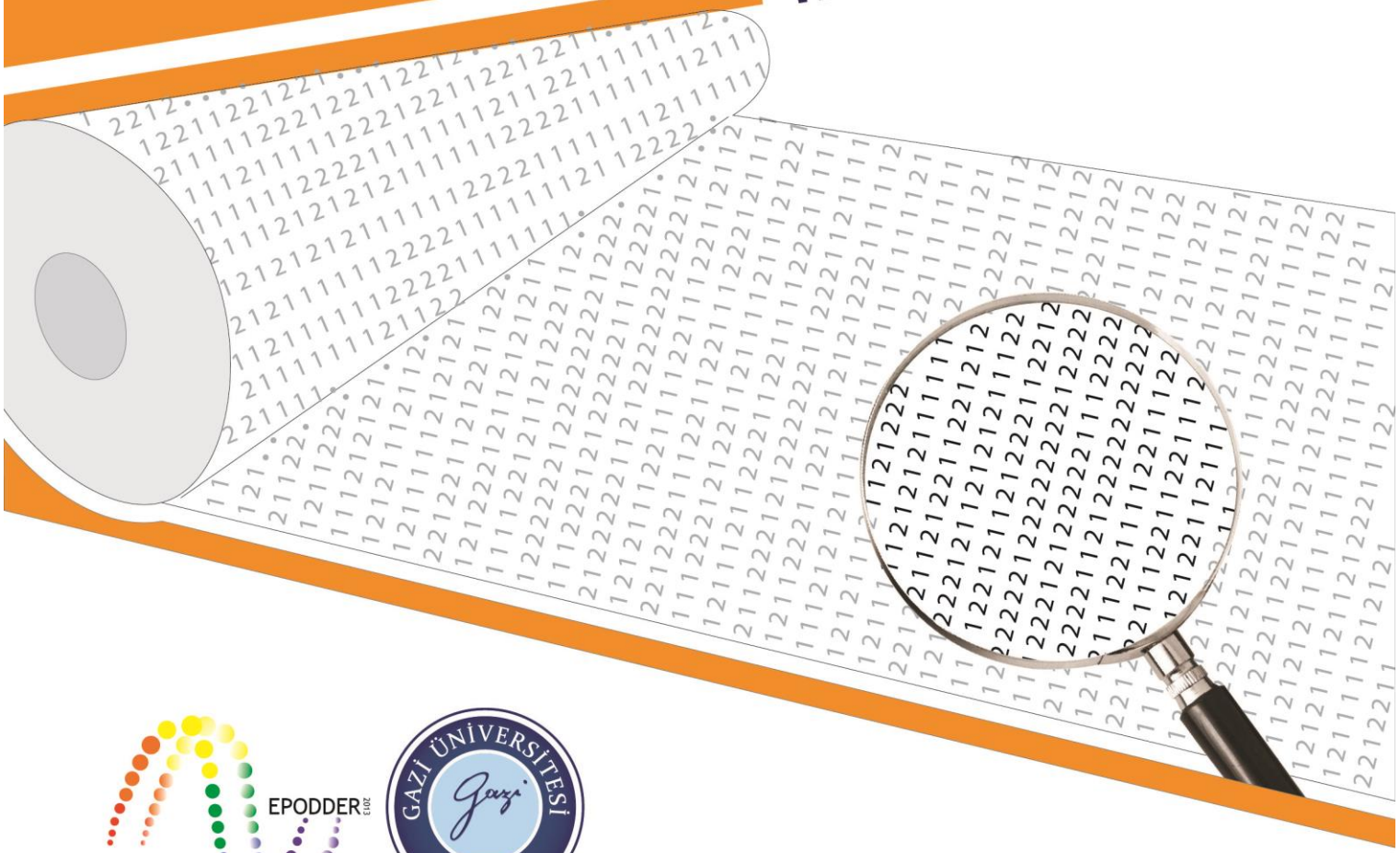
2020 CMEEP

1 - 4 Eylül 2021

7. Uluslararası Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi

Kongre Bildiri Kitapçığı



EPODDER

GAZİ ÜNİVERSİTESİ 1926

7. Uluslararası
Eğitimde ve Psikolojide
Ölçme ve Değerlendirme Kongresi

CMEEP-2020

1-4 Eylül 2021

VII. ULUSLARARASI
EĞİTİMDE VE PSİKOLOJİDE ÖLÇME ve DEĞERLENDİRME KONGRESİ

Kongre Düzenleme Kurulu Onursal Başkanı
Prof. Dr. Musa YILDIZ

Kongre Düzenleme Kurulu Eş Başkanları
Prof. Dr. Mehtap ÇAKAN
Prof. Dr. Nuri DOĞAN

Kongre Onur Kurulu
Prof. Dr. Ali BAYKAL
Prof. Dr. Durmuş Ali ÖZÇELİK
Prof. Dr. İlhan AKHUN
Prof. Dr. Mehmet Fuat TURGUT
Prof. Dr. Nizamettin KOÇ
Prof. Dr. Süleyman Çetin ÖZOĞLU
Prof. Dr. Yaşar BAYKUL
Doç. Dr. Halil TEKİN
Dr. Fethi TOKER

Kongre Sekreteryası
Dr. Öğr. Üyesi Mahmut Sami KOYUNCU
Dr. Öğr. Üyesi Mehmet ŞATA
Arş. Gör. Dr. Ayşenur ERDEMİR
Arş. Gör. Dr. Elif SEZER
Arş. Gör. Dr. Serpil ÇELİKTEN DEMİREL
Arş. Gör. Dr. Tuba GÜNDÜZ
Arş. Gör. Dr. Yıldız YILDIRIM
Arş. Gör. Dr. Zafer ERTÜRK
Arş. Gör. Aybüke DOĞAÇ
Arş. Gör. Ergün Cihat ÇORBACI
Arş. Gör. Esra OYAR
Arş. Gör. F. Gül İNCE ARACI
Arş. Gör. Merve YILDIRIM SEHERYELİ
Arş. Gör. Muharrem ŞENGÜL
Arş. Gör. Oya ERDİNÇ AKAN
Arş. Gör. Seyhan SARITAŞ AKYOL
Arş. Gör. Sinem ŞENFERAH
Uzm. Metehan GÜNGÖR

Kongre Düzenleme Kurulu Yöneticisi
Prof. Dr. Mahmut SELVİ

Kongre Düzenleme Kurulu

Prof. Dr. Hakan Yavuz ATAR
Prof. Dr. İsmail KARAKAYA
Prof. Dr. Mehtap ÇAKAN
Prof. Dr. Nilüfer KAHRAMAN
Prof. Dr. Nuri DOĞAN
Prof. Dr. Şeref TAN
Doç. Dr. Ayfer SAYIN
Doç. Dr. C. Deha DOĞAN
Doç. Dr. Dilara BAKAN KALAYCIOĞLU
Doç. Dr. Emine ÖNEN
Doç. Dr. Eren Can AYBEK
Doç. Dr. Melek Gülşah ŞAHİN
Doç. Dr. Murat Doğan ŞAHİN
Dr. Öğr. Üyesi Eren Halil ÖZBERK
Dr. Öğr. Üyesi Görkem CEYHAN
Dr. Öğr. Üyesi Vahit BADEMCİ
Arş. Gör. Ömer KAMIŞ
Arş. Gör. Sebahat GÖREN

Yayına Hazırlayan ve Kapak Tasarım
Arş. Gör. Muharrem ŞENGÜL

VII. ULUSLARARASI
EĞİTİMDE VE PSİKOLOJİDE ÖLÇME ve DEĞERLENDİRME KONGRESİ
Bilim Kurulu

Prof. Dr. Adnan ERKUŞ
Prof. Dr. Adnan KAN
Prof. Dr. Akihito KAMATA
Prof. Dr. Bayram BIÇAK
Prof. Dr. Bayram ÇETİN
Prof. Dr. Bruno D. ZUMBO
Prof. Dr. Cindy M. WALKER
Prof. Dr. Devrim ALICI
Prof. Dr. Duygu ANIL
Prof. Dr. Ezel TAVŞANCIL
Prof. Dr. Gülşah BAŞOL
Prof. Dr. Hakan ATILGAN
Prof. Dr. Hakan Yavuz ATAR
Prof. Dr. Halil Giray BERBEROĞLU
Prof. Dr. Hülya KELECİOĞLU
Prof. Dr. İbrahim Alper KÖSE
Prof. Dr. İsmail KARAKAYA
Prof. Dr. Jeroen VERMUNT
Prof. Dr. Jimmy de la TORRE
Prof. Dr. Kadriye ERCİKAN
Prof. Dr. Mehtap ÇAKAN
Prof. Dr. Neşe GÜLER
Prof. Dr. Nilüfer KAHRAMAN
Prof. Dr. Niyazi KARASAR
Prof. Dr. Nizamettin KOÇ
Prof. Dr. Nuri DOĞAN
Prof. Dr. Ömay ÇOKLUK BÖKEOĞLU
Prof. Dr. Rahime Nükhet ÇIKRIKÇI
Prof. Dr. Satılmış TEKİNDAL
Prof. Dr. Selahattin GELBAL
Prof. Dr. Şener BÜYÜKÖZTÜRK
Prof. Dr. Şeref TAN
Prof. Dr. Terry A. ACKERMAN
Prof. Dr. Tuncay ÖĞRETMEN
Prof. Dr. Yaşar BAYKUL
Prof. Dr. Zekeriya NARTGÜN

Doç. Dr. Ayfer SAYIN
Doç. Dr. Bilge GÖK
Doç. Dr. Burak AYDIN
Doç. Dr. Burcu ATAR
Doç. Dr. C. Deha DOĞAN
Doç. Dr. Dilara BAKAN KALAYCIOĞLU
Doç. Dr. Durmuş ÖZBAŞI
Doç. Dr. Emine ÖNEN
Doç. Dr. Eren Can AYBEK
Doç. Dr. Ergül DEMİR
Doç. Dr. Erkan Hasan ATALMIŞ
Doç. Dr. Gökhan AKSU
Doç. Dr. Güçlü ŞEKERCİOĞLU
Doç. Dr. H. Deniz GÜLLEROĞLU
Doç. Dr. Hakan KOĞAR
Doç. Dr. Halil İbrahim SARI
Doç. Dr. Hasan TABAK
Doç. Dr. Hatice KUMANDAŞ ÖZTÜRK
Doç. Dr. İrfan YURDABAKAN
Doç. Dr. K. Zülfikar DENİZ
Doç. Dr. Kübra ATALAY KABASAKAL
Doç. Dr. Melek Gülşah ŞAHİN
Doç. Dr. Meltem ACAR GÜVENDİR
Doç. Dr. Murat Doğan ŞAHİN
Doç. Dr. Mustafa Yüksel ERDOĞDU
Doç. Dr. Nezaket Bilge UZUN
Doç. Dr. Okan BULUT
Doç. Dr. Önder SÜNBÜL
Doç. Dr. Özen YILDIRIM
Doç. Dr. Özlem Yeşim ÖZBEK
Doç. Dr. Ragıp TERZİ
Doç. Dr. Safiye BİLİCAN DEMİR
Doç. Dr. Sedat ŞEN
Doç. Dr. Seher YALÇIN
Doç. Dr. Serkan ARIKAN
Doç. Dr. Sevilay KİLMEN
Doç. Dr. Tahsin Oğuz BAŞOKÇU
Doç. Dr. Tülin OTBİÇER ACAR
Doç. Dr. Ufuk AKBAŞ
Doç. Dr. Yeşim ÖZER ÖZKAN

Kongre Açılış Konuşması - I

Sayın Millî Eğitim Bakan Yardımcım Prof. Dr. Petek AŞKAR, Sayın Gazi Eğitim Fakültesi Dekanım (Prof. Dr. Mahmut Selvi), Sayın Kongre Eş Başkanları (Prof. Dr. Nuri Doğan ve Prof. Dr. Mehtap Çakan) ve Değerli Katılımcılar,

Bugün açılışını yaptığımız ve 1-4 Eylül 2021 tarihleri arasında gerçekleştirilecek olan 7. Uluslararası Eğitimde Ölçme ve Değerlendirme Kongresi'ne ben de hoş geldiniz diyorum. Öncelikle Kongreye Gazi Üniversitesi olarak ev sahipliği yapmaktan mutluluk duyduğumuzu belirtmek isterim.

Erken çocukluk döneminden başlayarak üniversite yaşantısına ve hatta sonrasına kadar devam eden öğrenme sürecinde, öğrencilerimizin neyi ne kadar öğrendiklerinin belirlenmesi için ölçme ve değerlendirme önem arz etmektedir. Sınıf içi ortamlardaki ölçme ve değerlendirme süreçleri kadar, Kongre'nin bir paneli de olan kademeler arası geçiş sürecindeki ölçme ve değerlendirmeler de ülkemizde öne çıkmaktadır. Ayrıca ülkemizin katıldığı TIMSS, PISA gibi uluslararası uygulama sonuçlarını doğru okuyabilmek için de ölçme ve değerlendirme bilgisi ön plana çıkmaktadır.

İçinde bulunduğumuz 21. yüzyılda bireyden beklenen beceriler, önceki yüzyıllara göre farklılık göstermektedir. Artık eğitim-öğretimle birlikte öğrenmeyi öğrenebilen, yaratıcı düşünebilen, muhakeme, problem çözme gibi düşünme becerileri gelişmiş bireyler yetişmesi beklenmektedir. Öğretim sürecine bağlı olarak ölçme ve değerlendirme sürecinin de 21.yyda dönüşmesi beklenmektedir ki Kongre kapsamında bu konunun da ele alındığını görüyorum.

Yaklaşık iki yıldır içinde bulunduğumuz salgın dönemi, bizlere eğitim-öğretimin her bir ögesinin ne kadar önemli olduğunu göstermiş oldu. Çevrim içi ortamda öğrencilere nasıl sınavlar uygulanacağı, bu sınavın nasıl değerlendirileceği salgının ilk başlarında fazlaca tartışıldı. Bu süreçlere ilişkin de bilimsel sonuçların bu Kongrede ele alınacağını düşünüyorum.

Dört gün sürecek Kongre kapsamında gerçekleştirilecek paneller, çalıştaylar, çağrılı konuşmacıların sunumları ve sözlü sunularla ölçme ve değerlendirmenin farklı boyutları ele alınacağını düşünüyorum. Çevrim içi gerçekleştirilecek Kongrede yüz yüze yapılacak çalıştaylara da ev sahipliği yapmaktan mutluluk duyduğumuzu belirtmek isterim.

Kongre'nin düzenlenmesinde başta Kongre Eş Başkanlarına ve Kongre Düzenleme Kurulu'na teşekkür ederim. Kongre'ye destek sağlayan Millî Eğitim Bakanlığına, TÜBİTAK'a, yayınevlerine ve diğer destekçilere teşekkür ederiz. Ayrıca Kongre'nin her aşamasına katkı sağlayan Gazi Eğitim Fakültesi Dekanlığına ve ev sahibi ana bilim dalımız Eğitimde Ölçme ve Değerlendirme Ana Bilim Dalı öğretim üyelerini tebrik eder ve kendilerine teşekkür ederim.

Kongrenin verimli ve hayırlı olması dileklerimle tekrar tüm katılımcılara hoş geldiniz diyorum.

Prof. Dr. Musa YILDIZ
Gazi Üniversitesi Rektörü

Kongre Açılış Konuşması - II

Sayın Millî Eğitim Bakan Yardımcım (Prof. Dr. Petek AŞKAR), Sayın Gazi Üniversitesi Rektörüm (Prof. Dr. Musa YILDIZ), Sayın Kongre Eş Başkanları (Prof. Dr. Nuri Doğan ve Prof. Dr. Mehtap Çakan) ve Değerli Katılımcılar,

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Derneği- EPODDER ile iş birliğiyle gerçekleştirdiğimiz 7. Uluslararası Eğitimde Ölçme ve Değerlendirme Kongresi'ne hoş geldiniz. Kongrenin eğitim camiasına hayırlı olmasını diliyorum.

Bilindiği gibi ölçme ve değerlendirme süreçleri, eğitim-öğretimimizin ayrılmaz bir parçası. Aslında birçok farklı amaçla gerçekleştirilse de biz ölçme ve değerlendirmeyi en çok puan ya da not verilirken, öğrenciler hakkında önemli kararlar alınırken görüyoruz. Örneğin bir öğrencinin hangi lisede ya da üniversite okuyacağının belirlenme sürecinde ölçme ve değerlendirme ön plana çıkıyor. Dün YKS yerleştirme sonuçları açıklandı. Üniversitemiz ve Fakültemizi kazanan öğrencilerimize de şimdiden hayırlı olsun diyorum. Dolayısıyla doğru bir şekilde gerçekleştirilen ölçme ve değerlendirme uygulamalarını, aslında bireysel farklılıkların öneminin tartışıldığı günümüzde adalet sağlayıcı bir unsur olarak görebiliriz.

Ülkemizde son yıllarda ölçme ve değerlendirme alanında yenilikçi yaklaşımların olması, birçok toplantının gündemini ölçme ve değerlendirmenin oluşturması da beklendik bir durum. İçinde bulunduğumuz küresel salgın döneminde de eğitim-öğretim sürecinin birçok ögesi gibi ölçme ve değerlendirme çalışmaları da öne çıktı. Biz de Dekanlığımız olarak çevrim içi toplantılar düzenleyerek ölçme ve değerlendirme alanında katkı sağlamaya çalıştık.

Bugün de yedincisi gerçekleştirilen Uluslararası Eğitimde Ölçme ve Değerlendirme Kongresi'ne ev sahipliği yapmaktayız. Desteklerini bizden esirgemeyen Üniversitemizin Rektörlüğüne teşekkür ederim. Kongre'nin düzenlenmesine katkı sağlayan Kongre eş başkanlarına, bu süreçte emek veren Eğitimde Ölçme ve Değerlendirme Ana Bilim Dalı öğretim üyelerine ve kurul üyelerine, katkı sağlayan kurum ve kuruluşlara teşekkür ederim. Kongre'nin eğitim camiasına hayırlı olmasını diliyorum.

Prof. Dr. Mahmut Selvi
Gazi Üniversitesi Gazi Eğitim Fakültesi Dekanı

Teşekkür Konuşması

Kongremize açılışına katılan Gazi Eğitim Fakültesi Dekanı Prof. Dr. Mahmut Selvi, Gazi Üniversitesi Rektörü Prof. Dr. Musa Yıldız ve Milli Eğitim Bakan Yardımcısı Prof. Dr. Petek Aşkar'a teşekkürlerimizi sunarız.

Kongre düzenleme kurulu eş başkanı Prof. Dr. Mehtap Çakan olmak üzere Gazi Üniversitesi Eğitimde Ölçme ve Değerlendirme öğretim üyelerine, düzenleme kurulu, bilim kurulu ve sekreteryada görevli tüm üyelere tüm süreçlerdeki destekleri ve gayretleri için teşekkür ederiz.

EPODDER yönetim kurulunda görev alan tüm hocalarımıza gayretleri için teşekkür ederiz. EPODDER'in yaklaşık üç ay önce yönetim kurulu değişti. Daha önce yönetim kurulu başkanı Prof. Dr. Hakan Yavuz Atar nezdinde tüm yönetim kuruluna katkılarından dolayı teşekkür ederiz.

Kongremiz kapsamında dört panel gerçekleştirildi. Prof. Dr. Şener BÜYÜKÖZTÜRK, Prof. Dr. Nuri DOĞAN, Doç. Dr. C. Deha DOĞAN ve Doç. Dr. Gülşen TAŞDELEN TEKER'e moderatörlükleri için onların nezdinde kongremize katkı sağlayan tüm panelistlere teşekkür ederiz.

Kongre davetimizi kabul ederek çağrılı konuşmacılarımız olan Prof. Dr. Akihito KAMATA, Prof. Dr. Bruno D. ZUMBO, Prof. Dr. Fatma BIKMAZ, Prof. Dr. İnayet AYDIN, Prof. Dr. Jimmy de la TORRE, Doç. Dr. Okan BULUT, Doç. Dr. Önder SÜNBÜL, Doç. Dr. Sedat ŞEN ve Doç. Dr. Serkan ARIKAN'a teşekkür ederiz.

Kongremizde yüz yüze gerçekleştirilen çalıştaylarda eğitim görevlisi olarak yer alan başta yurt dışından bizi kırmayan Prof. Dr. Akihito KAMATA'ya teşekkür ederiz. Yine çalıştay il dışından gelerek çalıştay düzenleyen Prof. Dr. Tuncay ÖĞRETMEN, Eğitimci Selim DAŞÇIOĞLU ve Eğitimci Zeynep UZUN'a teşekkür ederiz.

Kongremizdeki 24 farklı oturumda oturum başkanı olarak görev alan tüm hocalarımıza, bildiri sunumlarıyla katkı sağlayan tüm katılımcılarımıza ve dinleyicilerimize teşekkür ederiz.

Kongremizin başından sonuna kadar büyük bir özveriyle çalışan, açılış ve kapanış konuşmalarımızın moderatörlüğünü üstlenen Doç. Dr. Ayfer Sayın'a teşekkür ederiz. Yine her türlü organizasyonda görev alan Arş. Gör. Sebahat Gören'e, Arş. Gör. Ömer Kamış'a; bize teknik anlamda önemli destekler veren Özgür Güler'e teşekkür ederiz.

Kongrenin mutfağında yer alarak asıl yükü taşıyan sekreterya grubumuza teşekkür ederiz. Sekreterya grubundaki tüm arkadaşların tek tek isimlerini söylemek isterim:

Dr. Öğr. Üyesi Mahmut Sami KOYUNCU
Dr. Öğr. Üyesi Mehmet ŞATA
Arş. Gör. Dr. Ayşenur ERDEMİR
Arş. Gör. Dr. Elif SEZER
Arş. Gör. Dr. Serpil ÇELİKTEN DEMİREL
Arş. Gör. Dr. Tuba GÜNDÜZ
Arş. Gör. Dr. Yıldız YILDIRIM
Arş. Gör. Dr. Zafer ERTÜRK
Arş. Gör. Aybüke DOĞAÇ
Arş. Gör. Ergün Cihat ÇORBACI
Arş. Gör. Esra OYAR
Arş. Gör. F. Gül İNCE ARACI
Arş. Gör. Merve YILDIRIM SEHERYELİ
Arş. Gör. Muharrem ŞENGÜL
Arş. Gör. Oya ERDİNÇ AKAN
Arş. Gör. Seyhan SARITAŞ AKYOL
Arş. Gör. Sinem ŞENFERAH
Uzm. Metehan GÜNGÖR

Prof. Dr. Nuri Doğan
Kongre Eş Başkanı

SUNUMLAR

BİLDİRİ

Yapay sinir ağları ile küçük örneklemde bireyselleştirilmiş bilgisayarlı testler için kalibrasyon çalışması

Hüseyin Yıldız ve Eda Akdoğdu

Giriş

Bilgisayar teknolojisinin gelişimi, ölçme alanında Madde Tepki Kuramı (MTK) modellerinin ve onun önemli uygulamalarından olan Bireyselleştirilmiş Bilgisayarlı Testlerin (BBT) daha rahat kullanımına imkân vermiştir (Weiss ve Kingsbury, 1984). BBT geleneksel kağıt kalem testlerinin aksine bireylere yönlendirilen maddelerin ve test uzunluklarının farklılaşmasına imkan sağlamaktadır. BBT'lerin temel amacı, bireylere çok kolay ya da çok zor gelecek, dolayısıyla bireyin bulunduğu yetenek düzeyinde yeterli bilgi üretimi gerçekleştiremeyecek maddelerden kaçınarak test uzunluklarının kısaltılması amaçlanmaktadır (Weiss ve Kingsbury, 1984; Davey, 2011). Her bireye bulunduğu yetenek düzeyinde yüksek bilgi üreten maddelerin yönlendirilmesi daha düşük standart hataya sahip yetenek kestirimlerinin daha az sayıda madde kullanılarak yapılmasına imkan sağlamaktadır (Weiss, 1982). Belirtilen avantajların sağlanabilmesi için aşılması gereken bazı güçlüklere sahiptir. Kullanılan maddelerin kalitesi ve madde havuzunun genişliği BBT ile yapılan yetenek kestirimlerinin doğruluğunu doğrudan etkilediğinden (Stocking, 1994; Flaugher, 2000) uygulama öncesi geniş bir madde havuzunun oluşturulması ve oluşturulan havuzdaki maddelerin yeterli büyüklükte örnekleme uygulanarak havuzun kalibre edilmesi gerekmektedir. Ancak geniş madde havuzu ve havuzun kalibrasyonu için geniş örneklemlere uygulama yapılması gerekliliği ciddi zaman, emek ve ekonomik kaynak gerektirebilir. Bu çalışma BBT havuz kalibrasyonunu Yapay sinir Ağı (YSA) algoritması kullanarak çok daha küçük örneklemler ile gerçekleştirmeyi amaçlamaktadır. Elde edilen bulguların aşağıdaki sorunların çözümüne katkı sunacağı düşünülmektedir:

1.Sonuçları bakımından yüksek öneme sahip sınavlarda sınav öncesi madde havuzunun geniş örneklemlerle kalibrasyonunun oluşturduğu güvenlik sorunu.

2.Madde ifşalarının doğuracağı sonuçlardan kaçınmak için gerekli olan madde havuzu güncellemelerinde tekrar tekrar büyük örneklem uygulamalarına ihtiyaç duyulması.

Yöntem

Bu araştırmada yapay sinir ağları ile küçük örneklemde bireyselleştirilmiş bilgisayarlı testler için kalibrasyon çalışması gerçekleştirmek amacıyla veriler yapay olarak türetilmiştir. Bu yönüyle araştırma simülasyon çalışmasıdır. Yapılan bu çalışma uygulamada yaşanan bir sorunu çözmeye, iyileştirmeye dayalı bir araştırma olduğundan temel araştırma niteliği taşıdığı söylenebilir (Büyüköztürk ve diğ., 2020).

Araştırma veri üretimi, YSA eğitimleri ve BBT simülasyonu aşamalarından oluşmaktadır. Araştırmada kullanılan veriler üretilirken 2PL model altında, a ve b parametreleri sırasıyla log-normal [$a \sim lnN(0.0,\ 0.2)$] ve normal dağılımdan [$b \sim N(0,1)$] çekilmiştir. Ayrıca yetenek parametrelerinin üretimi için de yine normal dağılım [$\theta \sim N(0,\ 1)$] kullanılmıştır. Üretilen verilerin a ve b parametreleri "*irtGUI*" paketi (Yıldız, 2021) ile gerçekleştirilmiştir. YSA eğitim ve kestirimleri WEKA v.3.8.4. programı aracılığı ile yürütülmüştür. Son olarak gerçekleştirilen BBT simülasyonları için "*catR*" paketi (Magis ve Raiche, 2012) kullanılmıştır. Çalışmanın istatistiksel analizleri aşağıdaki basamaklarla yürütülmüştür:

1. 2000 kişi 250 madde veri seti üretildi. Veri setine ait a ve b parametreleri ile KTK'ya dayalı güçlük ve ayırıcılık parametreleri hesaplanıp YSA için öğrenme verisi olarak kullanıldı.

2. 2000 kişi 100 maddelik bir başka veri seti üretildi. 2000 kişi için MTK'ya dayalı a ve b parametreleri kestirilerek gerçek parametreler olarak kaydedildi. Veri setinden rastgele 100 birey seçildi.

3. Seçilen bireyler ile KTK'ya dayalı güçlük ve ayırıcılık parametreleri kestirildi. Önceden eğitilmiş YSA algoritması ile güçlük ve ayırıcılık indeksleri kullanılarak aynı maddeler için a ve b parametreleri kestirildi ve kestirilen parametreler olarak kaydedildi.

4. catR paketi "simulate respondents" fonksiyonu ile 100 simülatif birey için aynı yapay katılımcılar için gerçek ve kestirilen parametreler kullanılarak ayrı ayrı BBT simülasyonları yürütüldü. Her birey için yetenek kestirimleri gerçekleştirildi.

5. Gerçek ve kestirilen parametreler kullanılarak hesaplanan yetenek değerleri arasındaki korelasyon, RMSE ve BIAS değerleri raporlaştırıldı.

Sonuçlar

Bu araştırma büyük örneklemler kullanılarak MTK analiz araçları ile kalibre edilmiş madde havuzu kullanılarak yapılan yetenek kestirimleri ile küçük örneklemler kullanılarak YSA ile kalibre edilmiş madde havuzu kullanılarak yapılan yetenek kestirimleri arasıdaki ilişki düzeyini ortaya çıkarmıştır. İlişkiyi ortaya koyan korelasyon katsayısı, RMSE, BIAS, ortalama mutlak hata değerleri elde edilmiştir. Buna göre iki farklı yetenek parametresi seti arasındaki korelasyon katsayısı .928, RMSE .383, BIAS .031 ve ortalama mutlak hata .300 olarak hesaplanmıştır. Bu değerler incelendiğinde YSA ile küçük örneklemler kullanılarak yapılan kalibrasyon çalışması konvansiyonel yöntemlerle yapılan kalibrasyon çalışmasıyla oldukça benzer sonuçlar verdiğini göstermektedir. Bu çalışma farklı makine öğrenmesi

algoritmaları, farklı veri üretim koşulları, farklı öğrenme ve test verisi büyüklükleri ve farklı BBT simülasyon koşulları altında tekrarlanarak daha yüksek ilişkilerin elde edilebileceği düşünülmektedir.

Kaynaklar

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş. ve Demirel, F. (2020). *Eğitimde bilimsel araştırma yöntemleri.* Pegem Akademi.

Davey, T. (2011). *A guide to computer adaptive testing systems.* Council of Chief State School Officers.

Flaugher, R. (2000). Item pools. In H. Wainer (Eds.), *Computerized adaptive testing: A primer* (2nd ed., pp. 37–59). Lawrence Erlbaum Associates.

Frank, E., Hall, M. A., and Witten, I. H. (2016). *The WEKA workbench: Data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann. https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf

Magis, D., and Raiche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software*, *48*(8), 1-31. https://doi.org/10.18637/jss.v048.i08

Parshall, C. G., Spray, J. A., Kalohn, J. C. and Davey, T. (2002). *Practical considerations in computer-based testing.* Springer

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (ETS Research Report 94-5). Educational Testing Service.

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*(4), 473-492. https://doi.org/10.1177/014662168200600408

Yıldız, H. (2021). *irtGUI: Item response theory analysis with a graphic user interface* (version 0.2.) [Computer Software.] https://doi.org/10.6084/m9.figshare.14229953.v1

Okuduğunu anlama ve duygusal okuryazarlık becerilerinin ölçülmesinde görsel olarak zenginleştirilmiş yenilikçi madde formatları

Ömer Topraktepe ve Nilüfer Kahraman

Giriş

Çocukların okul ve okul dışı yaşantılarında karşılaşabilecekleri eğitsel metinleri okuma ve anlama becerilerinin iyi olmasının, gelişim ve öğrenme süreçlerini olumlu etkileyeceği düşünülmektedir (Baştuğ ve diğ., 2021). Erken çocuklukta okur-yazarlık eğitimi konusunu çalışan alan yazın incelendiğinde, okuma ve anlama kavramlarının iki temel boyut olarak tanımlandığı ancak, günümüze değin yapılan araştırmalarda, daha çok anlama boyutu üzerinde durulduğu görülmektedir (Özkan ve Başkan, 2020; Sadoski ve Pavio, 2007; Yılmaz, 2021). Daha güncel çalışmalarda ise araştırmacıların, okur-yazarlık becerisinin ön plana alınması ve sadece tanımlama, kavrama gibi bilişsel becerilerin değil; iletişim, liderlik gibi diğer özelliklerin de çalışılması gerektiğini sıkça ele aldığı görülmektedir (Cansoy, 2018; Harari, 2018; Lamb ve diğ., 2017).

Akyol'un (2011) "yazar ile okuyucu arasında aktif ve etkili iletişimi gerekli kılan, dinamik bir anlam kurma süreci" olarak tanımladığı okuma'nın, çağın getirdiği teknolojik imkânlar ile desteklenmiş görsel ve işitsel metinleri de içerecek şekilde, yani daha geniş bir açıdan ele alınması gerektiği açıktır (Duran ve Topbaşoğlu, 2018). Bu bağlamda, çocukların öğrenme yaşantıları sırasında etkileşimde bulundukları yazılı/görsel/işitsel metinleri okuma/anlama becerilerinin, karşılaşılan metinde geçen kişi-olay-durum etkileşimlerini bir bütün olarak algılamaları ve kendileri ile bağdaştırarak içselleştirmelerinin, yani duygusal okur-yazarlıklarının önemli olduğu gittikçe daha çok tartışılmaktadır (Kaya, 2018; Steiner, 2020).

Duyuşsal alan içerisinde yer aldığı düşünülen duygusal okur-yazarlık becerilerinin, gözlenip, ölçülmesinin, bilişsel alan içerisinde yer aldığı düşünülen okur-yazarlık becerilerine göre daha zor olduğu düşünülmektedir (Kaya, 2018; Turgut ve Baykul, 2019). Teknolojinin hızla gelişmesi ile birlikte, kâğıt kalem testleriyle kolayca ölçülemeyen bu özellikleri ölçmek için yenilikçi madde formatları çalışmaya başlanmıştır (Demirtaşlı, 2017). Yenilikçi madde formatı, geleneksel testlerle değerlendirmesi zor olan özelliklerin, multimedya araçlarını da kullanarak değerlendirilmesine imkân sağlayan madde formatlarıdır (Zenisky ve Sireci, 2002). Multimedya unsurları eklenerek zenginleştirilmiş yenilikçi

maddelerin, ölçülmek istenen hedefe yönelik verilen konunun kavranmasına katkıda bulunabileceği düşünülmektedir (Strain-Seymour ve diğ., 2009).

Bu kapsamda eldeki çalışmanın odak noktası, ilkokul 4. sınıf öğrencilerinin Türkçe dersinde okuduğunu anlama ve duygusal okuryazarlık becerilerinin gelişiminin izlenmesi ile ilgili yapılacak olan uygulama üzerinden, metin ile birlikte verilen resimler (iki boyutlu); metin ile birlikte verilen videolar (üç boyutlu) kullanılarak yenilikçi bir madde formatı tasarlamak ve geliştirilen madde formatının hem bilişsel hem de duyuşsal becerilere ilişkin ölçme sonuçlarına getireceği veri zenginliğini uygulamalı olarak araştırmaktır. Bu amaç doğrultusunda aşağıdaki sorulara yanıt aranacaktır.

1.Ölçmeye konu öğrenme ve gelişim durumlarını izlemede tekrarlı olarak kullanılacak ölçeklerdeki maddelerin iki boyutlu veya üç boyutlu olmalarının, tek boyutlu olmalarına göre, çocukların okuduğunu anlama ve duygusal okuryazarlık becerileri ile ilgili daha detaylı, geri bildirim değeri daha yüksek bilgiler sağlama açısından getireceği katkılar ne düzeydedir?

1.1.Tek boyutlu, iki boyutlu ve üç boyutlu madde formatı kullanılarak uygulanan testlerden elde edilen puanlara dayalı *test istatistikleri ve madde istatistikleri* hangi düzeydedir? Üç ayrı formattaki testten elde edilen istatistikler arasında anlamlı bir farklılık var mıdır?

1.2.Tekrarlı ölçümlerin uygulandığı üç ayrı zaman noktası dikkate alındığında, öğrencilerin *gelişim eğrileri* farklılık göstermekte midir?

1.3.Güvenirlik kanıtı olarak; tek boyutlu, iki boyutlu ve üç boyutlu madde formatı kullanılarak uygulanan testlere ilişkin elde edilen *test bilgi fonksiyonları* nasıl bir değişim göstermektedir? Üç ayrı formata ilişkin, test bilgi fonksiyonlarından ulaşılan bilgi düzeyleri arasında iki ve üç boyutlu ölçekler lehine bir farklılık var mıdır?

1.4.Geçerlik kanıtı olarak; tek boyutlu, iki boyutlu ve üç boyutlu madde formatı ile hazırlanan testlerden alınan puanların, sınıf öğretmenlerinin ilgili kazanımlara yönelik öğrencilere verdikleri puanlarla göstermiş oldukları ilişki ne düzeydedir?

Yöntem

Araştırmanın çalışma grubunu, 2020-2021 eğitim öğretim yılı içerisinde, İstanbul Sultangazi İlkokulu'nda 4. sınıfa kayıtlı olan 419 öğrenci arasından, gönüllülük esasıyla seçilecek öğrenciler oluşturacaktır. İlgili okulun seçilmesindeki sebep, zaman ve iş gücü kaybını önleyerek, en ulaşılabilir yanıtlayıcıların varlığından kaynaklanmaktadır. Uygulamalar yüz yüze olacaktır. Veri toplama araçları, araştırmacı tarafından uygulanacaktır. Araştırma, ilkokul 4. sınıfta öğrenim gören üç farklı gruptaki öğrencilere Türkçe dersi kapsamında, bir hafta aralıklarla aynı maddelerin farklı formatlarını içeren (multimedya unsurları ile zenginleştirme biçimiyle birbirinden ayrılan), üç ayrı okuduğunu anlama ve duygusal okuryazarlık bütünleşik ölçeğinin uygulanmasını kapsamaktadır. Araştırma, tekrarlı gözlemlerle öğrencilerin okuduğunu anlama ve duygusal okuryazarlık becerilerinin gelişiminin izlenmesi açısından boylamsal desen kategorisine girmektedir. Boylamsal desen çoğunlukla gelişim ile ilgili veri toplamak ve

belirli bir süreç içerisindeki değişim durumlarını incelemek amacıyla tekrar eden ölçme işlemlerinin uygulanmasını içermektedir (Fraenkel ve Wallen, 2009). Veri toplama aracı olarak, yalnızca sunuluş biçimiyle birbirinden ayrılan, madde kökünde ortak metinlerin ve bu metne bağlı maddelerin yer aldığı üç tür ölçek yer alacaktır. Oluşturulan ilk iki ölçek, yenilikçi madde formatında hazırlanacaktır. Yenilikçi madde formatı; geleneksel testlerle değerlendirilmesi zor olan bilgi, beceri ve yeteneklerin ses, grafik, video gibi multimedya araçlarını da kullanarak değerlendirme imkânı sağlayan madde formatlarıdır (Zenisky ve Sireci, 2002). Alan yazında yapılan araştırmalar, yenilikçi madde tipinde hazırlanan testlerin daha fazla bilgi sağlayarak daha geçerli ve güvenilir sonuçlar ürettiğini ortaya koymaktadır (Jodoin, 2003; Martinez, 1991; Wan ve Henly, 2012). Madde köküne eklenecek ek kaynakların ölçme ve değerlendirme sürecine getireceği katkılardan yola çıkılarak, hem okuduğunu anlama hem de duygusal okuryazarlık becerilerinin gelişiminin izlenmesinde ilk olarak, ortak metinlere eklenmiş olan videoların yer aldığı üç boyutlu ölçekler; ikinci olarak, ortak metinlere eklenmiş olan resimlerin yer aldığı iki boyutlu ölçekler oluşturularak, yenilikçi madde formatında testler kullanılacaktır. Üçüncü olarak ise diğer iki ölçekle ortak metinlerin ve bu metne bağlı maddelerin bulunduğu, yalnız yazılı metin içeren tek boyutlu ölçeklerden oluşan testler yer alacaktır.

Sonuçlar

Tek boyutlu, iki boyutlu ve üç boyutlu madde formatı kullanılarak uygulanan testlerden elde edilen puanlara dayalı hesaplanan *madde istatistikleri* arasında anlamlı bir farklılık çıkması beklenmektedir. Testin boyutu arttıkça, madde istatistiklerine dayalı olarak elde edilecek olan *madde güçlük indeksinin* 1.0 değerine daha yakın; *madde ayırt edicilik gücü indeksinin* ise daha yüksek değerler alması beklenmektedir. Üç ayrı madde formatı kullanılarak uygulanan testlerden elde edilen puanlara dayalı *test istatistikleri* arasında da anlamlı fark çıkması beklenmektedir. Öğrencilerin testin boyutu arttıkça, test içerisinde yer alan metinleri ve bu metinlere dayalı maddeleri daha iyi anlayacağı düşünülerek, toplam test puanlarında da bir artış olacağı böylece testin boyutu arttıkça aritmetik ortalamanın da yükseleceği beklenmektedir.

Testin boyutu arttıkça metni anlamaya katkı sağlayacak destekleyici unsurların da artması ve özellikle duygusal okuryazarlık ile ilgili oluşturulan maddelere ilişkin daha derinlemesine yanıtlar vermeye ortam oluşturması yönüyle testin boyutunun artmasının öğrencilerin okuduğunu anlama ve duygusal okuryazarlık becerilerinin gelişimine destek olacağı, böylelikle daha fazla gelişim sağlayacağı beklenmektedir.

Oluşturulan test formatının boyutu arttıkça tüm yetenek düzeylerinde daha fazla bilgi vermesi beklenmektedir. Testin boyutu arttıkça, öğrencilerin teste ilişkin aldığı puanlar ile öğretmenlerin ilgili kazanımlara yönelik öğrencilere verdikleri puanlar arasında ilişkinin düzeyinin daha yüksek olması beklenmektedir.

Kaynaklar

Akyol, H. (2011). *Türkçe öğretim yöntemleri* (4. Baskı). Pegem Akademi.

Baştuğ, M., Hiğde, A., Çam, E., Örs, E. ve Efe, P. (2021). *Okuduğunu anlama becerilerini geliştirme, stratejiler, teknikler, uygulamalar* (2. baskı). Pegem Akademi.

Cansoy, R. (2018). Uluslararası çerçevelere göre 21. yüzyıl becerileri ve eğitim sisteminde kazandırılması. *İnsan ve Toplum Bilimleri Araştırmaları Dergisi*, *7*(4), 3112-3134.

Demirtaşlı, N. (2017). Bilgisayar destekli testler için yenilikçi maddeler. F. Odabaşı, B. Akkoyunlu ve A. İşman (Eds.), *Eğitim teknolojileri okumaları* içinde (ss. 59-74). Vadi. http://www.tojet.net/ebook/eto_2017.pdf

Duran, E. & Topbaşoğlu, N. (2018). *Hikâye kitaplarında anlama.* Pegem Akademi.

Fraenkel, J. R., and Wallen, N. E. (2009). *How to design and evaluate research in education.* McGraw-Hill.

Harari, Y. N. (2018). *21. yüzyıl için 21 ders.* Kolektif Kitap.

Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computerbased testing. *Journal of Educational Measurement, 40*(1), 1-15. http://www.jstor.org/stable/1435051

Kaya, S. G. (2018). *Görsel olarak zenginleştirilmiş yenilikçi madde formatı geliştirilmesi ve uygulamalı olarak değerlendirilmesi* (Tez No: 528242) [Yüksek lisans tezi, Gazi Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Lamb, S., Maire, Q., and Doecke, E. (2017). *Key skills for the 21st century: An evidence-based review.* Report prepared for the State of New South Wales (Department of Education) Sydney. https://vuir.vu.edu.au/35865/1/Key-Skills-for-the-21st-Century-Analytical-Report.pdf

Martinez, M. E. (1991). A comparison of multiple-choice and constructed figural response items. *Journal of Educational Measurement, 28*(2), 131–145. https://doi.org/10.1111/j.1745-3984.1991.tb00349.x

Özkan, E. ve Başkan, A. (2020). Okuma ve söz varlığı ilişkisi. M. Kaya ve M. N. Kardaş (Eds.), *Okuma eğitimi* içinde (ss. 67-78). Pegem Akademi.

Sadoski, M., & Paivio, A. (2007). Toward a unified theory of reading. *Scientific Sutudies of Reading, 11*(4), 337-356. https://doi.org/10.1080/10888430701530714

Steiner, C. (2020). *Akıllı bir kalple duygusal okuryazarlık [Emotional Literacy: Intelligence with a Heart]* (M. Şahin ve F. Erden, Çev.; 4. baskı). Nobel Akademik Yayıncılık (2003).

Çoktan seçmeli maddelerin aynı grupta tekrar kullanımının maddelerin psikometrik özelliklerine etkisi

Levent Yakar

Anahtar kelimeler: Çoktan seçmeli madde, vize, final, madde ayırt edicilik, madde güçlüğü

Giriş

Çoktan seçmeli maddeler uygulama ve değerlendirme kolaylığı ve objektifliği sayesinde eğitimde en sık kullanılan ölçme araçlarındandır. Ancak hazırlama aşamasındaki zorluk göz önünde bulundurulduğunda ise çoktan seçmeli maddelerin her durumda kullanışlı olduğu söylenemez (Doğan, 2020). Özellikle işe yarar, mantıklı çeldiricilerin kolay bir şekilde yazılamaması çoktan seçmeli maddelerin kullanımını sınırlandıran etmenler arasında yer almaktadır (Haladyna ve Downing, 1993). Bu durum diğer pek çok ölçme aracına kıyasla çoktan seçmeli maddelerin hazırlanmasını daha maliyetli hale getirmektedir.

Çoktan seçmeli maddelerin kullanışlılığını artırmak için öncelikle hazırlama maliyetini düşürmek gerekmektedir. Test konu alanı ve hazırlama sürecinde uzmanlaşmak daha kaliteli ve hızlı maddeler üretilmesini bir noktaya kadar destekleyecek unsurlar olduğu söylenebilir. Bunun yanı sıra maddeleri birden fazla kullanımı da soru hazırlama maliyetini ciddi bir şekilde azaltan bir uygulamadır. Ancak maddelerin doğru bir şekilde tekrar kullanımında, güvenirlik hesaplama yöntemlerinden test-tekrar test yönteminde ifade edildiği gibi ilk cevabın hatırlanmaması gerekmektedir (Başokçu, 2020). Üstelik çoktan seçmeli maddeler değer biçmeye yönelik değerlendirme amacını taşıyan bir başarı testi ise öğrencilerin verdikleri cevabın doğruluğunu sonrada kontrol etmeleri de maddelerin tekrar kullanımına ilişkin bir sorun kaynağı olarak görülebilir.

Dönem sonu sınavları öncesinde "Hocam, vize konularından da soru gelecek mi?" şeklindeki soru hemen hemen tüm öğretim elemanlarının karşılaştığı üniversite öğrencilerinin klasik soruları arasında yer almaktadır. Cevabın "Evet" olması durumunda vize konuları için öncekilerden bağımsız yeni maddeler hazırlanması test hazırlama işlemini çok daha fazla emek gerektiren bir sürece dönüştürecektir. Aynı maddelerin kullanımında ise öğrencinin cevabı hatırlayabilmesi, maddenin psikometrik özelliklerini dolayısıyla da testin ve maddenin güvenirliğini tehdit eden bir unsur olabilir.

Bu araştırmada çoktan seçmeli maddelerin üniversitedeki ara ve dönem sonu sınavlarda tekrar kullanımı durumunda maddelerin psikometrik özelliklerine etkisi incelenmiştir. Bu amaçla,

a. Çoktan seçmeli maddelerin tekrar kullanımında madde ayırt edicilik indeksi farklılaşmakta mıdır?
b. Çoktan seçmeli maddelerin tekrar kullanımında madde güçlük indeksi farklılaşmakta mıdır?
c. Maddeyi bir kez ve iki kez alanlar için hesaplanan madde ayırt edicilik indeksi farklılaşmakta mıdır?
d. Maddeyi bir kez ve iki kez alanlar için hesaplanan madde güçlük indeksi farklılaşmakta mıdır?

araştırma soruları yanıtlanmaya çalışılmıştır.

Yöntem

Bu araştırmada maddelerin bir kez ve iki kez iletildiği gruplar bulunmaktadır. Uygulamanın gruplara göre araştırmacı tarafından farklılaştırıldığı düşünüldüğünde çalışmanın deneysel araştırma türünde olduğu söylenebilir. Grupların oluşturulması tesadüfi yöntem kullanılması nedeniyle çalışma gerçek deneysel araştırma desenine sahiptir (Akbay, 2019).

Araştırmanın çalışma grubunu araştırmacının 2020-2021 Bahar döneminde, 5 şubede verdiği Eğitimde Ölçme ve Değerlendirme dersini alan öğrenciler oluşturmaktadır. İlahiyat fakültesinde 4 şubede 509, Eğitim fakültesinde bir şubede ise 64 toplamda 573 öğrenci bulunmaktadır.

Ara dönem sınavı için 62 çoktan seçmeli madde hazırlanmış ve bunlardan 25'i öğrenme yönetim sistemi aracılığıyla öğrencilere atanmıştır. Dönem sonu sınavında ise ara dönem konularına ait sorulardan 11, son konulara ilişkin yeni hazırlanan 40 maddeden 14 tanesi yine öğrenme yönetim sistemi aracılığıyla öğrencilere atanmıştır. Atama işlemi konu ağırlıkları da göz önünde bulundurularak sistem tarafından rastgele gerçekleştirilmiştir. Sınavlar ayrı günlerde 30 dakikalık zaman dilimi içerisinde çevrimiçi ve eşzamanlı olarak gerçekleştirilmiştir. Sınavlarda gözetmen uygulaması kullanılmamıştır. Sınavın gerçekleştirildiği öğrenme yönetim sistemindeki her bir öğrencinin her bir soruya verdiği yanıtlar indirilmiş ve öğrencinin işaretlediği seçenek cevap anahtarı yardımıyla doğru (1) – yanlış (0) listesine dönüştürülmüştür. Ardından Excel yardımıyla her iki sınavda da aynı soru(ları) alan deney grubu tespit edilerek kontrol ve deney grubu için madde parametreleri elde edilmiştir.

Excel'de elde edilen veriler SPSS'e aktarılarak analizler gerçekleştirilmiştir. İlk araştırma sorusu için öncelikle, vize öncesi konulara ait 62 sorudan en az birini her iki sınavda da alan, her bir madde için ortalama 41'er öğrenciden oluşan deney grubu tespit edilmiştir. Bu grubun maddeler için ilk ve ikinci sınavdaki doğru yanıt ortalaması hesaplanarak madde güçlük parametreleri elde edilmiş ve ilişkili örneklemler t testi ile analiz edilmiştir. İkinci araştırma sorusu için madde ayırt edicilik indeksi madde puanı-toplam puan korelasyonu yöntemi ile hesaplanmıştır. Bu araştırma sorunda deney grubunun ilk ve ikinci sınavdaki ortak maddelere verdikleri yanıtlardan madde ayırt edicilik indeksleri elde edilmiş ve bunlar ilişkili örneklemler t testi ile analiz edilmiştir. Üçüncü araştırma sorusu için deney grubunun ikinci

sınavda tekrar aldığı sorulara verdiği yanıtlar için elde edilen madde güçlük parametreleri, kontrol grubunun ikinci sınavda verdiği yanıtlar için elde edilen madde güçlük parametreleri ile bağımsız örneklemler t testi ile analiz edilmiştir. Bu işlem son araştırma sorusu için ise madde ayırt edicilik parametreleri için gerçekleştirilmiştir.

Sonuçlar

Her iki sınavda da aynı soruları alan deney grubu için oluşan madde parametrelerinin vize-final sınavları karşılaştırılmasına ilişkin sonuçlar Tablo 1'de sunulmuştur.

Tablo 1

Deney Grubu Ön Test-Son Test Madde Parametreleri Karşılaştırması

Değişken	*N*	Ön Test Ort	Ss	Son Test Ort	Ss	*t*	*p*	*η*
Madde Güçlük	62	.51	.18	.54	.18	-2.82	.006	.37
Madde Ayırt Edicilik	62	.35	.5	.40	.15	-2.32	.024	.31

Tablo 1'de görülen eşleştirilmiş örneklem t testi sonuçları incelendiğinde, soruları tekrar alan öğrenci grubundan elde edilen sonuçlarda vize sınavına kıyasla final sınavında madde güçlük ve madde ayırt edicilik indekslerinde istatistiksel olarak anlamlı bir artış olduğu görülmektedir. Bu artışların etki büyüklüğü incelendiğinde ise farka ilişkin etkinin .2-.5 arasında olduğu ve düşük etkiye sahip olduğu görülmüştür (Cohen, Manion ve Morrison, 2007).

Tablo 2

Deney ve Kontrol Grubu Son Test Madde Parametreleri Karşılaştırması

Değişken	*N*	Ön Test Ort	Ss	Son Test Ort	Ss	*t*	*p*
Madde Güçlük	62	.54	.18	.50	.18	-1.225	.22
Madde Ayırt Edicilik	62	.40	.15	.35	.15	-1.62	.11

Deney ve kontrol grupları cevapları için elde edilen madde parametrelerinin karşılaştırılmasının bulunduğu Tablo 2'ye bakıldığında deney grubundan elde edilen madde güçlük ve madde ayırt edicilik parametrelerinin kontrol grubuna göre kısmen yüksek olduğu ancak aradaki farkın istatistiksel olarak anlamlı olmadığı görülmektedir.

Sonuçlar

Diğer ölçme araçlarına kıyasla çoktan seçmeli maddeler hazırlanması daha fazla emek gerektiren ölçme araçlarıdır. İhtiyaç hissedilmesi durumunda yeniden madde hazırlamanın pek mümkün olmadığı

durumlarda maddelerin tekrar kullanımı gerekebilir. Bu araştırmada, çoktan seçmeli maddelerin aynı grup üzerinde tekrar kullanımının maddelerin psikometrik özelliklerine etkisi incelenmiştir.

Araştırmada, maddelerin hem vize hem de final sınavlarında tekrar kullanıldığı durumda madde güçlük indekslerinin kısmen arttığı ve soruların daha kolay hale geldiği görülmüştür. Bunda öğrencilerin ilk sınavda yaptığı hataların farkına vararak aynı hatayı ikinci sınavda yapmamaları etkin olabilir. Ancak, iki sınav arasındaki güçlük farkının etki büyüklüğünün düşük düzeyde olması bu değişimin ciddi boyutlarda olmadığını göstermektedir. Benzer şekilde madde ayırt edicilik indekslerinin de bir miktar artarak maddelerin daha ayırt edici hale geldiği görülmüştür. Buradaki değişimin de etki büyüklüğü yine düşük düzeyde olduğu görülmüştür. Bu durum, maddedeki performansı artan öğrencilerin bu artışı genele de yansıtmış olabilmelerinden kaynaklanabilir.

Final sınavında maddeyi ilk kez alan öğrencilerden oluşan kontrol grubu ile ikinci kez alan deney grubu için oluşan madde parametrelerine bakıldığında ise maddelerin her iki grup için de benzer zorluğa ve ayırt ediciliğe sahip olduğu görülmüştür. Bu sonuç, ilk iki araştırma problemindeki gözlenen kısmi farkın sadece aynı soruları tekrar almaktan kaynaklanmadığını göstermektedir.

Bu sonuçlardan hareketle, uzaktan ölçme yöntemi ile gözetmensiz yapılan bir sınavda bile madde parametrelerinde ciddi farklılıklar görülmemesi, maddelerin aynı grupta tekrar kullanımının sorun teşkil etmeyeceğini göstermektedir. Ancak, bu sonuç değerlendirilirken, alternatifli çok sayıda soru hazırlanmasının ve uygulamalar arasında 7 haftalık süre olmasının, vizedeki sorularının final sınavında hatırlanmasını zorlaştırmış olabileceği de göz önünde tutulmalıdır.

Kaynaklar

Akbay, T. (2019). Deneysel Araştırmalar. S. Şen ve İ. Yıldırım (Eds.), *Eğitimde araştırma yöntemleri* içinde (ss.155-180). Nobel Akademik Yayıncılık.

Başokçu, T. O. (2020). Ölçme süreç ve sonuçlarının nitelikleri: Ölçme hatası, güvenirlik, geçerlik ve kullanışlılık. N. Doğan (Ed.), *Eğitimde ölçme ve değerlendirme* içinde (ss. 32-75). Pegem Akademi.

Cohen, L. M., Manion, L., and Morrison, K. (2007). *Research methods in education*. Routledge.

Doğan, N. (2020). Geleneksel Ölçme ve Değerlendirme Teknikleri I: Yanıtı Seçmeyi Gerektiren Ölçme Araçları. N. Doğan (Ed.), *Eğitimde ölçme ve değerlendirme* içinde (ss. 114-139). Pegem Akademi.

Haladyna, T. M., and Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, *53*(4), 999-1010. https://doi.org/10.1177/0013164493053004013

Performans değerlendirmede yeni bir yaklaşım

Nuri Doğan, Sümeyra Soysal ve Mine Demirbaş

Anahtar kelimeler: Göz izleme, puanlayıcılar arası uyum, puanlayıcılar arası güvenirlik

Giriş

Değerlendirme, ölçme sonuçlarının bir ölçütle karşılaştırılarak ölçülen nitelik hakkında karar verme sürecidir. Bu şekilde çok genel bir tanıma sahip olan değerlendirmenin en doğru şekilde yapılması açısından, ölçümlerin geçerli ve güvenilir olması önemli bir rol oynamaktadır. Farklı niteliklerin farklı ölçüm süreçleri ve yöntemleri bulunmaktadır. Bilişsel ve duyuşsal özelliklerin ölçülmesi belirli araçlara dayandırılarak yapılmaktadır. Bu tür özellikler doğrudan ölçülemediği için geliştirilen farklı ölçme araçları ve yaklaşımları ile ölçüm sonuçları elde edilmektedir. Bu nedenle dolaylı bir ölçme işlemi gerçekleştirilmektedir. Dolaylı ölçme sonuçlarının geçerlik ve güvenirlik açısından değerlendirilmesi için birçok aşamadan geçmesi gerekmektedir. Bu süreç oldukça zordur. Eğitim sürecinde uygulanan performas değerlendirmede de benzer güçlükler söz konusu olmaktadır.

Performans değerlendirilirken ürün ve sürecin değerlendirilmesinde sıklıkla rubriklerden (dereceli puanlama anahtarlarından) yararlanılır. Rubrik her ne kadar belirli kriterleri içeriyor olsa bile puanlayıcıların kanaatlerini yansıtması söz konusu olduğu için sübjektif puanlamadan tamamen kaçınmak mümkün değildir. Puanlayıcılıarın subjektiflik veya objektiflik düzeyini ortaya koymak için çeşitli istatistiksel tekniklerden yararlanılabilir. Son yıllarda geliştirilen teknolojik araçların da objektif ölçümlerin sağlanması için kullanılabildiği gözlenmektedir. Bu tekniklerden biri de göz izleme tekniğidir. Dolayısıyla bu araştırmada "puanlayıcı uyumunda kullanılan belirli istatistiksel tekniklerle elde edilen sonuçları göz izleme tekniğiyle desteklemek mümkün müdür?" sorusuna cevap aranmıştır.

Araştırma sonuçlarının yukarıda sözü edilen genel amaç yanında bazı alt amaçlara da hizmet edeceği düşünülmektedir. Bu alt amaçlar aşağıdaki şekilde sıralanabilir.

- Göz izleme yöntem ve araçlarının performans değerlendirmede kullanılabilirliğini belirlemek
- Puanlayıcıların performans değerlendirmede nerelere dikkat ettiği konusunda bir yargıya varabilmek
- Göz izleme yöntem ve araçlarının kullanışlılığı hakkında bilgi sahibi olmak

- Göz izleme yöntemiyle elde edilen farklı performans puanlarının güvenirlik açısından değerlendirmek
- Göz izleme tekniği ile elde edilen örüntüleri değerlendirmek

Öğrencilerin performansını değerlendirmede puanlayıcılar arası uyum sonuçları ile göz izleme tekniğinden elde edilen sonuçların ilişkisi nasıldır? Buna bağlı olarak:

1. Puanlayıcılar arası uyum nasıldır? (İkili ve üçlü karşılaştırma)
2. Performans değerlendirme sürecinin sonunda elde edilen puanların(klasik tekniklerle) tobii uygulaması sonucu elde edilen metrikslerle aralarında nasıl bir ilişki vardır?
3. Puanlayıcılar arası sınıf içi uyum ve göz izleme sonucunda elde edilen metriksler arasında uyum var mıdır?

Yöntem

Araştırmanın yöntemi betimsel araştırmalar kategorisine girmektedir. Farklı tekniklerle elde edilen bilgilerin ilişkilendirilmesi nedeniyle ilişkisel, farklı tekniklerle elde edilen bilgilerin değerlendirilmesi nedeniyle betimsel çalışmalar arasında sayılabilir. Araştırma betimsel ve karşılaştırmalı araştırma içerisinde değerlendirilebilir.

Çalışma grubu araştırmaya katılan alanında beş farklı uzman ve üç farklı performans görevini içermektedir. Sınıf içi uygulama çalışması eğitim felsefesi dersi kapsamında gerçekleştirilen grup çalışmasında yer alan sunum yapma performans görevleridir. Eğitimde ölçme ve değerlendirme, eğitim programları ve öğretimi, eğitim yönetimi, alanında farklı uzmanların görüşleri alınarak hazırlanan dereceli puanlama anahtarı, beş farklı Eğitimde Ölçme ve Değerlendirme alanında yeterliliğe sahip puanlayıcı tarafından doldurulmuştur. Sunumları yapan gruplardan birincisi yaratıcılık, ikincisi realizm ve üçüncüsü idealizm konularını sunum için seçmişlerdir.

Uygulama için lisans ve yüksek lisans da farklı branşlarda ve bölümlerde eğitime devam eden Hacettepe üniversitesi öğrencileri gönüllü olarak katılım sağlamışlardır. Uygulama için ilk aşama da grupla veya bireysel öğrenci sunumları videoya alınmıştır. Bunun yanında beş alan uzmanı hakem(gözlemci) sunum esnasında hazır bulunarak sınıf içinde değerlendirmelerini, uzman görüşleri alınarak hazırlanmış gözlem çizelgesine göre tamamlamışlardır. Daha sonra Tobii Pro X2-60 (Çeşitli araştırma ve senaryolar için özellikle göz izleme konusunda esneklik sağlayan güvenilir bir mobil cihazdır.) kullanılarak hakemler ekran üzerinden değerlendirmelerini tekrar yapmışlardır.

Veri toplama araçları performans görevi videoları ve puanlayıcıların performans değerlendirme için hazırlanan puanlama yönergesi ve göz izleme sonucunda elde edilen puanlayıcı görüntü örüntüleri olmak üzere üç boyutta ele alınabilir.

Analitik Gözlem anketi (ölçeği) oluşturulmadan önce literatür taraması yapılmıştır. Elde edilen sonuçlar maddeler haline getirilerek alan uzmanı kişilerin görüşüne sunulmuştur. Bunun sonucunda elde edilen maddeler kullanılarak gözlem anketinin son hali verilmiştir. Gözlem anketiyle pilot çalışma

yapılmıştır. Pilot çalışma sonucunda 19 maddenin yer aldığı gözlem anketinin son halini kullanılmaya karar verilmiştir.

Sunumları yapan gruplardan birincisi yaratıcılık, ikincisi realizm ve üçüncüsü idealizm konularını sunum için seçmişlerdir. 5 alan uzmanı puanlayıcı ve 3 farklı performans hem sınıf ortamında hem de göz izleme sistemiyle video üzerinden değerlendirilmiştir. Bunun sonucunda elde edilen veriler çeşitli istatistiksel işlemler yapılarak yorumlanmıştır. Öncelikle puanlayıcılar öğrenciler sınıfta gözlem yoluyla puanladılar. Bu sırada öğrenci performans kayıtları yapıldı. Puanlayıcılar daha sonra öğrenci performans videolarını izleyerek öğrencileri yeniden puanlardılar. Bu süreçte puanlayıcılara ilişkin göz izleme kaydı tutulmuştur. Örüntülerin analizinde betimsel istatistikler (odaklanma, sıçrama ve sıcaklık haritaları) ve içerik analizi kullanılacaktır. Bunun yanında puanlayıcılar arası uyum ikili karşılaştırmalarda Kappa, üç puanlayıcı aynı anda Kendall uyum katsayısı, genelleştirilmiş kappa ve bunun yanında diğer yöntemler kullanılarak değerlendirilmiştir.

Sonuçlar

Bulgular hazırlanmaya devam etmektedir.

Tablo 1

Birinci Video İçin Puanlayıcıların İlgi Alanı Üzerinde Toplam Sabitlenme Sürelerine Ait Betimsel İstatistikler

İlgi alanı üzerinde toplam sabitlenme süresi	Aoi ta	Aoi bi	Ort	Medyan	Toplam	İlgi alanı üzerindeki toplam süre	Toplam kayıt süresi
p1	217.9215	395.77	306.85	306.85	613.70	786.89	1456.75
p2	262.61	200.37	231.49	231.49	462.98	750.77	869.27
p3	313.00	246.85	279.92	279.92	559.84	748.20	955.14
ort	264.51	281.00	272.75	272.75	545.51	761.95	1573.42
Toplam zaman (%)	48.49	51.51					
Varyans	2262.62	10420.31	1458.20	1458.20	5832.8	468.00	1393886.43
Standart sapma	47.57	102.08	38.19	38.19	76.37	21.63	1180.63

AOI Total Fixation Duration: İlgi alanı üzerinde toplam sabitlenme süresi

Video incelenirken Tablo 1'de belirtilen iki ilgi alanı(AOI) üzerinden analiz yapılmaktadır. (Sunumu yapan birey ve sunum yapılan tahta) harcanan zaman göz önüne alındığında %51 sunum yapan kişi üzerinde %48 ise sunum yapılan tahta üzerinde gerçekleşmiştir. Ayrıca her bir etki alanının ortalama, varyans ve standart sapma değerleri tabloda yer almaktadır. Her bir puanlayıcının ilgi alanı üzerinde toplam sabitlenme süresi, ortalama, medyan, toplam, ilgi alanı üzerindeki toplam süre ve toplam kayıt süreleri yer almaktadır. 1.puanlayıcının ilgi alanları üzerindeki sabitlenme süresi incelendiğinde sunum yapan birey üzerinde, 2.ve 3. puanlayıcı sunum yapılan tahta üzerinde daha çok zaman geçirmişler. İlgi alanı üzerinde toplam sabitlenme süreleri görsel şekil 3 'te yer alan olarak ısı haritasında da incelenebilmektedir. Toplam kayıt süreleri farklı olsa bile puanlayıcıların ilgi alanı üzerinde geçirdikleri zamanlar birbirlerine oldukça yakın değerler almıştır.

Sonuçlar

Puanlayıcılar arası güvenirliği araştıran birçok çalışma yapılmıştır.(Şata,2019) Alanyazın incelendiğinde puanlayıcılar arası güvenirlik en çok dereceli puanlama anahtarları kullanılarak araştırılmıştır. Dereceli puanlama anahtarlarının güvenirliğini araştıran birçok çalışma yapılmıştır. Birçok etken karşılaştırılmıştır. Puanlayıcı sayısı, süresi, puanlanan sayısı değiştirilerek birçok karşılaştırma çalışması bulunmaktadır. Puanlayıcılar arası uyum analizleri için birçok farklı istatistiksel yöntem kullanılmıştır. Bu istatistikler genel olarak verinin parametrik veya nanparametrik olmasına göre, puanlayıcı sayısına göre farklılık gösterse bile yapılan çalışmalarda benzer sonuçlar verdikleri gözlenmektedir. Bu istatistiklere alternatif bir uyum veya güvenirlik kanıtı popüler sorunların başında gelmektedir.

Puanlayıcılar arası güvenirliği etkileyen birçok faktör bulunmaktadır. Bazılarının kaynağı belli olsa bile kaynağı belli olmayan bir etki de söz konusudur. Bu etkileri belirleyerek hata payını en aza indirerek puanlayıcılar arası yüksek bir uyum sağlamak, yapılan çalışmalarda en çok belirlenen hedeflerin başında gelmektedir. Kaynağı belli olmayan etkiler göz ardı edilemeyecek derecede önemli bir paydaya sahiptir. Puanlayıcıların tutumları, önyargıları, dikkatleri birçok etki düşünülebilir. Bu etkileri ölçmek ve bu ölçümlerden bir istatistiksel veri elde etmek günümüze kadar pek yaygın değildir. Psikolojik ölçümlerde bireyin dikkati, odaklanması, göz hareketleri, bakış süreleri yeni yapılan çalışmalarda birçok faktörün açıklanmasında kullanılan yeni veriler olarak karşımıza çıkmaktadır. Yapılan bu çalışmada ise bu yeni nesil verilerin puanlayıcılar arası uyum ve güvenirlik için bize farklı bakış açıları ve yorum yapma imkânı verdiği gözlemlenmiştir.

Göz izleme çalışması sonucunda, her bir sunum videosunda Puanlayıcıların odaklandığı ilgi alanları çalışmada yer alan dereceli puanlama anahtarına göre iki alan olarak belirlenmiştir. Bunlar sunum yapan birey ve sunum yapılan tahtadır. Dereceli puanlama anahtarı sunumun içeriği değil sunum yapan bireyin sunum yeteneğini ölçen bir özellik göstermektedir. Bu nedenle çalışmada yer alan puanlayıcıların sunum yapan birey ilgi alanına daha fazla odaklanması ve bu alan üzerinde daha fazla zaman geçirmesi beklenmektedir. Göz izleme sonucu elde edilen metriksler incelendiğinde genel olarak sonuçlar bu beklentiyi karşılar niteliktedir. Bu sonuçlar bize ölçeğin ölçmek istediği şeye odaklanıldığı sonucuna ulaşabiliriz. Bu sonuç bize ölçeğimizin geçerli olduğuna ve güvenilir sonuçlar verdiğine kanıt oluşturmaktadır. Her bir puanlayıcının her bir görev için benzer verilerinin olması da çalışmamızda odaklandığımız puanlayıcı güvenirliği için önemli bir bulgu niteliğindedir.

Kaynaklar

Şata, M. (2019). *Performans değerlendirme sürecinde puanlayıcı eğitiminin puanlayıcı davranışları üzerindeki etkisinin incelenmesi* (Tez No: 626117) [Doktora tezi, Gazi Üniversitesi]. Yükseköğretim Kurulu Ulusal Tez Merkezi.

Test kullanımına ilişkin bir inceleme: Sınav kaygısı envanteri

Betül Alatlı

Anahtar kelimeler: Test kullanımı, geçerlilik, güvenilirlik, raporlama, sınav kaygısı envanteri

Giriş

Doğrudan gözlenemeyen ve sosyal bilimlerde sıklıkla inceleme konusu olan zekâ, tutum, kişilik, inanç, algı gibi özelliklerin ölçülebilmesi için ilgili özelliği uyaran maddelerin yer aldığı testler kullanılmaktadır. Eğitimde ve psikolojide kullanılan ölçme araçları için genel bir isim olarak "test" kavramı kullanılmaktadır. Testler, büyük gruplarda, hızlı, bilimsel anlamda geçerli ve güvenilir veriler toplamayı sağlaması bakımından oldukça önemli görülmektedir (Cronbach, 1990). Türkiye'de, Türkiye Ölçme Araçları Dizini'nde (TOAD) 9016 test geliştirme ve uyarlama çalışmasının yer aldığı düşünülürse test kullanımının yaygınlığı ve testlerin oldukça önemli bir çalışma alanı olduğu görülmektedir. Literatür incelendiğinde test geliştirme ve uyarlama çalışmalarının ilgili adımlar ve birçok değişken bakımından incelendiği çalışmalar mevcuttur (Çüm ve Koç, 2013; Acar-Güvendir ve Özer-Özkan, 2015; Ergene, 2020; Erkuş, 2016; Şahin ve Boztunç Öztürk, 2018; Hambleton ve diğ., 2005; Murphy ve Davidshofer, 2005). Test geliştirme ve uyarlama çalışmaları kadar testlerin kullanımı da oldukça önemli bir konudur. Test kullanıcılarının dikkat etmesi gereken noktalar ve raporlaması gereken durumlar American Educational Research Association (AERA), American Psychological Association (APA) ve National Council on Measurement in Education (NCME) (2014) tarafından ayrıntılı bir şekilde ele alınmıştır. Literatürde test kullanımına ilişkin yapılan araştırmalar da yer almaktadır (Vacha-Haase ve diğ., 1999; Vacha-Haase, 1998; Simmelink ve Vacha-Haase, 1999; Barnes, Harp ve Jung, 2002). Ancak araştırmalarda belli dergiler veya belli bir testi kullanan araştırmalar incelenmiş ve genellikle güvenilirliğin raporlanması bakımından ele alınmıştır. Bu araştırmada ise Türk kültürüne uyarlanan literatürde de sıklıkla kullanılan Öner (1990) tarafından Türk kültürüne uyarlanan "Sınav Kaygısı Envanteri"nin kullanıldığı makalelerin ilgili kültüre uyarlanan bir ölçme aracının kullanımı bakımından dikkat edilmesi gerekenler doğrultusunda incelenmesi amaçlanmıştır.

Yöntem

Türk kültürüne uyarlanan Sınav Kaygısı Envanteri'nin kullanımı bakımından ilgili makalelerdeki eğilimin ortaya konulmasının amaçlandığı bu araştırma tarama modelinde betimsel bir araştırmadır

(Karasar, 2009). Araştırmanın çalışma grubunu Sınav Kaygısı Envanterinin Türk kültürüne uyarlanması ile elde edilen Türkçe formunun kullanıldığı, 2000-2020 yılları arasında yayınlanmış, tam metnine ulaşılabilen 45 makale oluşturmaktadır. Araştırma kapsamında incelenen makaleler Social Sciences Citation Index (SSCI), Alan indeksi, Ulakbim ve diğer indekslerde taranmaktadır. Araştırmaların yayın dili Türkçe ve İngilizce'dir. Araştırma verilerinin elde edilmesinde belgesel tarama veri toplama tekniği kullanılmıştır. Belgesel tarama tekniği ile var olan kayıt veya belgelerin incelenmesi ile veriler elde edilebilmektedir (Karasar, 2009). Çalışmada araştırmacı tarafından geliştirilen "Hedef kültüre uyarlanan ölçeklerin kullanımına ilişkin kontrol formu" kullanılmıştır. Formun geliştirilmesinde doktora eğitimini tamamlamış dokuz ölçme ve değerlendirme uzmanının görüşü alınmıştır. Verilerin analizinde içerik analizi kullanılmıştır. Verilerin analizi için araştırmacı tarafından geliştirilen kontrol formunda yer alan maddeler dikkate alınmıştır. Daha sonra yapılan kodlamalara göre ilgili maddeler bakımından frekans ve yüzde değerleri elde edilmiştir. Elde edilen verilerin güvenilirliği için 45 makale arasından tesadüfen seçilen 10 makale farklı bir uzman tarafından incelenmiştir. Miles ve Huberman (1994) tarafından geliştirilen güvenilirlik katsayısı formülü "Güvenirlik Katsayısı = Uzlaşma Sayısı / (Uzlaşma Sayısı + Uzlaşmama Sayısı)" kullanılacaktır.

Sonuçlar

Araştırma bulguları araştırmacı tarafından geliştirilen kontrol formundan elde edilen veriler doğrultusunda elde edilecektir. Buna göre araştırma kapsamında yer alan makalelerin ilgili testin adının, geliştirilme ve uyarlama çalışmalarının kaynaklarının, testin yanıtlarının ve puanlanmasına ilişkin bilgilerin, testin geçerlilik ve güvenilirliğine ilişkin bilgilerin doğru bir şekilde raporlanması bakımından incelenecektir. Bu doğrultuda elde edilen bulgulara göre testlerin kullanımına ilişkin dikkat edilmesi gereken adımlar bakımından ilgili makalelerdeki benzerlikler ve farklılıklar doğrultusunda var olan durum ortaya konulacaktır. Testlerden elde edilen puanlar doğrultusunda elde edilen bulgulara göre önemli sonuçlar elde edilmektedir. Bu nedenle testlerin kullanımına ilişkin araştırmacılara ya da test kullanıcılarına önemli öneriler bu araştırma sonuçlarına göre ortaya konulacaktır.

Kaynaklar

Acar-Güvendir, M. ve Özer-Özkan, Y. (2015). Türkiye'deki eğitim alanında yayımlanan bilimsel dergilerde ölçek geliştirme ve uyarlama konulu makalelerin incelenmesi. *Elektronik Sosyal Bilimler Dergisi, 14*(52), 23-33. https://doi.org/10.17755/esosder.54872

AERA, APA, and NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Barnes, L. B. L., Harp, D., and Jung, W.S. (2002). Reliability generalization of scores on the spielberger state–trait anxiety inventory. *Educational and Psychological Measurement, 62*(4), 603–618. https://doi.org/10.1177/0013164402062004005

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). Harper Collins Publishers.

Çüm, S. ve Koç, N. (2013). Türkiye'de psikoloji ve eğitim bilimleri dergilerinde yayımlanan ölçek geliştirme ve uyarlama çalışmalarının incelenmesi. *Eğitim Bilimleri ve Uygulama*, *12*(24), 115-135.

Ergene, Ö. (2020). Matematik eğitimi alanında ölçek geliştirme ve ölçek uyarlama makaleleri: Betimsel içerik analizi. *Yaşadıkça Eğitim, 34*(2), 360-38. https://doi.org/10.33308/26674874.2020342207

Erkuş, A. (2016). Ölçek geliştirme ve uyarlama çalışmalarındaki sorunlar ile yazım ve değerlendirilmesi. Pegem Atıf İndeksi, 1221-1234. https://doi.org/10.14527/9786053183563.075

Şahin, M. ve Boztunç-Öztürk, N. (2018) Eğitim alanında ölçek geliştirme süreci: Bir içerik analizi çalışması. *Kastamonu Eğitim Dergisi, 26*(1), 191-199. https://doi.org/10.24106/kefdergi.375863.

Hambleton, R. K., Merenda, P. F., and Spielberger, C. D. (2005). *Adapting educational and psychological tests for cross-cultural assessment.* Lawrence Erlbaum.

Murphy, K. R., and Davidshofer, C. O. (1998) *Psychological testing: Principles and applications* (4th ed.) Prentice Hall.

Öner, N. (1990). *Sınav kaygısı envanteri el kitabı.* İstanbul: Yükseköğrenimde Rehberliği Yayma Vakfı.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*(1), 6-20. https://doi.org/10.1177/0013164498058001002

Vacha-Haase, T., Ness, C., Nilsson, J., and Reetz, D. (1999). Practices regarding reporting of reliability coefficients: A review of three journals. *The Journal of Experimental Education, 67*(4), 335-341.

Küçük örneklemlerde bilişsel tanılama: Yapay sinir ağı, parametrik olmayan bilişsel tanılama ve DINA modelinin sınıflandırma performanslarının karşılaştırılması

Emine Yavuz ve Hakan Yavuz Atar

Anahtar kelimeler: Küçük örneklem, parametrik olmayan bilişsel tanılama, yapay sinir ağı, DINA

Giriş

Geleneksel test puanlarının karşılaştırılmaları, grupların akademik başarıları arasındaki farkı ortaya koyarken, grupların akademik alandaki bilişsel niteliklere göre nasıl farklılaştığı hakkında bilgi vermemektedir. Bu detaylı bilginin elde edilmesi için bilişsel tanı modelleri (BTM'ler) ortaya çıkmıştır. BTM'ler öğrencileri bilişsel nitelik sınıflarına yerleştirirken aynı zamanda maddelerle ilgili parametreleri kestirebilmektediler. Bazen sınıf gibi küçük örneklemlerde öğretme ve öğrenme sürecini izlemek için BTM'lerin madde parametrelerinin kestirimlerinde kullanılan algoritmalarının güvenilir sonuçlar vermediği gözlenmektedir (bkz. Akbay, 2016). Sınıf gibi küçük örneklemlerde odak genellikle öğrencilerin bilgiyi hangi düzeyde edindikleri veya becerilerinin hangi düzeyde geliştiği olduğundan (Chiu ve Köhn, 2019) sınıf düzeyi bilişsel tanı değerlendirmelerinde parametrik olmayan tekniklerin güvenilir ve kullanışlı olabilecekleri düşünülmektedir. Parametrik olmayan teknikler, madde parametre değerlerinden çok öğrencileri nitelik sınıflarına yerleştirmeye odaklanmaktadırlar. Sınıf gibi küçük örneklemlerdeki uygulamalarda bazen elde edilen veride kayıp verinin çok olduğu veya bu uygulamalarda madde sayısının az veya ölçülen niteliğin çok olduğu görülmektedir. Bu tür koşullarda da BTM'lerin öğrencileri nitelik sınıflarına yerleştirirken yanlı sonuçlar verebildikleri görülmüştür (Shuying, 2016; Shu ve diğ., 2013). Bu nedenle küçük örneklem, az sayıda madde, çok sayıda nitelik veya kayıp verinin bulunduğu koşullarda kullanılabilecek parametrik olmayan alternatif yöntemlerin geliştirilmesi ve bu yöntemlerin en iyi sınıflandırma performansı gösterdikleri koşulların belirlenmesi ihtiyacı ortaya çıkarmıştır. Bu bağlamda bu araştırma, bilişsel tanılamada kullanılan, BTM'lerle bazı benzerlik ve farklılıklara sahip alternatif yöntemlerden yapay sinir ağına (YSA'ya) ve parametrik olmayan bilişsel tanılamaya (POBT'ye) odaklanmaktadır. Bu araştırmada, YSA'ya ve POBT'ye ait nitelik (NSO) ve örüntü (ÖSO) düzeyi sınıflandırma oranlarının örneklem büyüklüğü, kayıp veri oranı, madde ve nitelik sayısı koşullarında Deterministic-Input, Noisy-And Gate (DINA) modeli temelli simülasyon veri setlerinde öncelikle birbirleriyle, daha sonra DINA modeli ile karşılaştırılmaları amaçlanmaktadır.

Yöntem

Simülasyon araştırması şeklinde modellenen bu çalışmada veri setlerinin üretimine Q matris yapısına karar verilerek başlanmıştır. Araştırmada bir maddede birden fazla niteliğin ölçüldüğü karmaşık Q matris yapısı temel alınmıştır. Koşullara ait Q matrislerinde niteliklerin benzer sayıda madde ile ölçülmesine dikkat edilmiştir. Ek olarak, niteliklerin üretilmesinde gerçek durumu betimleyebilmek için nitelikler arası ilişkilerin bulunduğu ve nitelik örüntülerinin eşit dağılmadığı çok değişkenli normal dağılım tercih edilmiştir. Simülasyon veri yapısına DINA modelin temel alınmasına karar verildikten sonra kaydırma (s) ve tahmin (g) madde parametre değerleri belirlenmiştir. S ve g parametreleri maddelerin ayırt ediciliğini belirlemektedir ve bu değerlerin 0,5'i aşması halinde güvenilir sınıflandırma oranlarına ulaşılamamaktadır (Chiu ve Douglas, 2013). Bu nedenle mevcut araştırmada tüm koşullarda madde ayırt edicilikleri karışık (orta) düzeyde, yani *s* ve *g* parametre değerleri U[0; 0,3] olarak belirlenmiştir. Simülasyon faktörleri alan yazın göz önüne alınarak madde (10, 15 ve 20) ve nitelik sayısı (3 ve 5), örneklem büyüklüğü (30, 60 ve 90) ve kayıp veri oranı (0; 0,05 ve 0,10) olarak belirlenmiştir. Bu doğrultuda araştırma kapsamında 3x2x3x3x3= 162 karşılaştırma durumu oluşturulmuştur. Araştırmada verilerin düzenlenmesinde Microsoft Excel 2019, verilerin üretilmesinde ve analizinde r ve faktöriyel ANOVA'nın yapılmasında IBM SPSS 26.0 programlarından faydalanılmıştır.

Sonuçlar

Araştırmada öncelikle YSA, POBT ve DINA modelinin her birine ait sınıflandırma oranlarının koşullar arasında nasıl değiştiği, ardından bu üç yöntemin koşullar arasında birlikte nasıl değiştikleri incelenmiştir. Analizler sonucunda yöntemlere ait sınıflandırma oranlarının koşullar arası değişimlerinde bazı benzerlik ve farklılıklar tespit edilmiştir. Örneğin, örneklem büyüklüğünün artmasıyla YSA ve POBT'ye ait NSO ve ÖSO'ların doğrusal değişimleri (artış veya azalış) hakkında bir sistematiklik belirlenemezken, DINA modeline ait NSO ve ÖSO'ların azaldıkları belirlenmiştir. Kayıp veri oranının artmasıyla YSA'ya ait NSO ve ÖSO'larının doğrusal değişimleri hakkında bir sistematiklik belirlenemezken, POBT ve DINA modeline ait NSO ve ÖSO'ların azaldıkları belirlenmiştir. Nitelik sayısının artmasıyla YSA'ya ait NSO'ların artarken, POBT ve DINA modeline ait NSO'ların azaldığı görülmektedir. Buna ek olarak nitelik sayısı arttıkça tüm yöntemlerin ÖSO'ları azalmaktadır. Son olarak, madde sayısının artmasıyla YSA'ya ait NSO ve ÖSO'lar azalırken POBT ve DINA modeline ait NSO ve ÖSO'lar artmaktadır. YSA, POBT ve DINA modelinin her birine ait sınıflandırma oranlarının koşullar arasında birlikte nasıl değiştikleri incelendiğinde, tüm koşullarda POBT'nin YSA'dan daha yüksek sınıflandırma oranlarına sahip olduğu belirlenmiştir. Ayrıca POBT, DINA modelinden biraz düşük fakat karşılaştırılabilir sınıflandırma oranlarına sahiptir. YSA'nın ve POBT'nin bilişsel tanılama alanında yeni kullanılmaya başlandığı göz önüne alındığında genellenebilir sonuçlara ulaşmak için daha çok çalışmanın yapılması gerektiği düşünülmektedir.

Kaynaklar

Akbay, L. (2016). Relative efficiency of the nonparametric approach on attribute classification for small sample cases. *Journal of European Education 6*(1), 43-59. https://doi.org/10.18656/jee.20599

Chiu, C.-Y., and Köhn, H.-F. (2019). Consistency theory for the general nonparametric classification method. *Psychometrika, 84*(3), 830–845. https://doi.org/10.1007/s11336-019-09660-x

Shu, Z., Henson, R., and Willse, J. (2013). Using neural network analysis to define methods of DINA model estimation for small sample sizes. *Journal of Classification*, *30*(2), 173-194. https://doi.org/10.1007/s00357-013-9134-7

Shuying, S. (2016). *Nonparametric diagnostic classification analysis for testlet based tests* (Publication No. 10154646) [Doctoral dissertation, The University of North Carolina]. ProQuest Dissertations & Theses Global.

Parametrik olmayan bilişsel tanılama, yapay sinir ağı ve DINO modelinin sınıflandırma performanslarının karşılaştırılması

Emine Yavuz ve Hakan Yavuz Atar

Anahtar kelimeler: Tanılayıcı değerlendirme, biçimlendirici değerlendirme, parametrik olmayan bilişsel tanılama, yapay sinir ağı, DINO, PISA

Giriş

Son yıllarda 21. yüzyıl becerilerinin değerlendirilmeleri ve tanılayıcı ve biçimlendirici değerlendirmelerin artması bilişsel tanı modellerinin (BTM'lerin) gelişimini hızlandırmıştır. Yapılan bazı çalışmalarda BTM'lerin çok sayıda niteliğin, küçük örneklemin ve az sayıda maddenin bulunduğu koşullarda öğrencilerin nitelik sınıflarına yerleştirilmesinde yanlı sonuçlar verdiği görülmüştür (Shuying, 2016; Shu, Henson, & Willse, 2013). Ölçme ve değerlendirme etkinliklerinde, özellikle öğrenci yeterliklerinin belirlendiği çalışmalarda araştırma amacına en uygun yöntemin kullanılması önemlidir. Bu nedenle BTM'lerin yanlı sonuçlar verebilecek koşullarda kullanılması için alternatif yöntemlerin ve bu yöntemlerin hangi koşullar altında daha iyi performans gösterdiklerinin belirlenmesine ihtiyaç vardır. Bu düşünceden hareketle mevcut araştırma, parametrik olmayan bilişsel tanılama (POBT) ve yapay sinir ağlarına (YSA) odaklanmaktadır. BTM'lerle analiz için Q matrisinin ve veri yapısının (telafisel veya telafisel olmayan) bilinmesi hususunda benzerlik gösteren bu yöntemler, öğrencileri nitelik profillerine sınıflandırmada farklı bakış açıları kullanmaları yönüyle birbirinden farklıdır. YSA ideal cevap örüntüleriyle eğitildikten sonra öğrencileri sınıflandırırken, POBT, öğrencilerin ideal cevap örüntüleri ile gözlenen cevap örüntülerini çeşitli uzaklık ölçütleri kullanarak karşılaştırıp öğrencileri nitelik sınıflarına yerleştirmektedir. Bu araştırmada, POBT ve YSA'ya ait nitelik (NSO) ve örüntü (ÖSO) düzeyi sınıflandırma oranlarının çeşitli koşullarda Deterministic-Input, Noisy-Or Gate (DINO) temelli simülasyon veri setlerinde öncelikle birbirleriyle, sonra DINO model ile karşılaştırılmaları, ardından POBT, YSA ve DINO modelinin nitelik (NSOB) ve örüntü (ÖSOB) düzeyi sınıflandırma oranlarının benzerliklerinin PISA 2015 İşbirlikli problem çözme veri seti üzerinde incelenmesi amaçlanmaktadır.

Yöntem

Betimsel ve simülasyon araştırma modelinin kullanıldığı bu araştırmada öncelikle gerçek veri setlerinin yapısı belirlenmiş, ardından alan yazın taraması da dikkate alınarak simülasyon veri setleri

üretilmiştir. Bu nedenle simülasyon veri setleriyle gerçek veri setleri bazı benzer özelliklere sahiptir. Örneğin her iki veri setinde nitelikler arası ilişki bulunmaktadır ve madde ayırt edicilikleri orta (karmaşık) düzeydedir. Gerçek veri setinin yapısı telafisel modellerden DINO modele uygun iken, simülasyon veri setleri de DINO model temel alınarak üretilmişlerdir. Bazı koşullarda gerçek veri setlerine ait Q matrisleri karmaşık yapılıdır. Simülasyon veri setleri üretilirken Q matris yapılarının da karmaşık olmasına dikkat edilmiştir. Ek olarak, simülasyon veri setlerinin üretilmesinde her niteliğin benzer sayıda madde ile ölçülmesine ve niteliklerin üretilmesinde çok değişkenli normal dağılımın kullanılmasına dikkat edilmiştir. Araştırmada simülasyon ve gerçek veri setleri için karşılaştırma koşulları, alan yazın ve gerçek veri yapısı dikkate alınarak oluşturulmuştur. Bu bağlamda simülasyon veri setleri için karşılaştırma koşulları: Madde (15, 30 ve 45) ve nitelik sayısı (3, 5 ve 7), örneklem büyüklüğü (30, 100 ve 500) ve kayıp veri oranıdır (0; 0,05 ve 0,10). Gerçek veri setleri (PISA 2015 İPÇ uygulaması) için karşılaştırma koşulları nitelik sayısı (3, 7 ve 11) ve örneklem büyüklüğüdür (30, 100 ve 500). Araştırma problemlerin cevaplanması için faktöriyel ANOVA yapılmıştır.

Sonuçlar

Araştırmada ilk önce simülasyon veri setleri üzerinde her bir yöntemin sınıflandırma oranlarının koşullar arasında nasıl değiştiği, ardından yöntemlere ait sınıflandırma oranlarının koşullar arasında birlikte nasıl değiştiği ve son olarak, gerçek veri setleri üzerinde yöntemlerin sınıflandırma oranlarının koşullar arası birlikte nasıl değiştiği incelenmiştir. Sonuç olarak, hem simülasyon hem de gerçek veri setlerinde POBT'ye ait NSO ve ÖSO'larının YSA'ya ait NSO ve ÖSO'lardan yüksek, POBT ve DINO modeline ait NSO ve ÖSO'ların ise genel olarak benzer olduğu belirlenmiştir. Buna ek olarak, örneklem büyüklüğü arttıkça POBT ve YSA'ya ait NSO ve ÖSO'ların doğrusal değişimleri (artış veya azalış) hakkında herhangi bir sistematikliğe rastlanmamıştır. Bilişsel tanı alan yazınında POBT ve YSA yeni çalışılmaya başlandığından mevcut araştırma bulgularını destekleyecek veya çürütecek daha fazla çalışmanın yapılması önerilmektedir.

Kaynaklar

Shu, Z., Henson, R., and Willse, J. (2013). Using neural network analysis to define methods of DINA model estimation for small sample sizes. *Journal of Classification*, *30*(2), 173-194. https://doi.org/10.1007/s00357-013-9134-7

Shuying, S. (2016). *Nonparametric diagnostic classification analysis for testlet based tests* (Publication No: 10154646) [Doctoral dissertation, The University of North Carolina]. ProQuest Dissertations & Theses Global.

Test geliştirme sürecinde madde ve test istatistiklerinin uzman görüşlerinden elde edilen sonuçlar ile karşılaştırılması

Ayfer Sayın ve Sebahat Gören

Anahtar kelimeler: Test geliştirme, madde güçlüğü, madde ayırt ediciliği, uzman görüşü

Giriş

Ülkemizde seçme, yerleştirme, sınıflama, değerlendirme gibi farklı amaçlar doğrultusunda hem MEB, ÖSYM gibi kurumlar tarafından gerçekleştirilen hem de öğretmenler tarafından sınıf içi ölçme uygulamalarında yapılan ve çoktan seçmeli olarak uygulanan birçok başarı testi uygulanmaktadır. Diğer test türlerinde olduğu gibi başarı testi geliştirme süreci de birçok adımı kapsayan sistematik bir işlemdir. Asıl istenen bu işlemin tüm aşamalarının titizlikle yapılıp güvenirlik ve geçerliğin istenilen kullanım için gerekli düzeydeki puanları sağlayacak minimum uzunlukta bir test elde etmektir (Crocker ve Algina, 1986).

Özçelik (2009) test geliştirme aşamalarını sırasıyla amacın saptanması, kapsamın belirlenmesi, soru türlerine ve sayısına karar verilmesi, yazılan soruların gözden geçirilmesi, deneme formunun uygulanması, ön uygulamanın yapılması, ön uygulamadan elde edilen sonuçlara göre madde analizlerinin yapılması, nitelikli maddelerin seçilmesi, nihai testin oluşturulması, nihai testin test istatistiklerinin yapılması şeklinde tanımlamıştır. Gerek ülkemizde seçme ve yerleştirme amacıyla hazırlanan sınavlarda gerekse öğretmenlerin sınıf içi ölçmelerde kullandıkları çoğu başarı testlerinde ön uygulamanın yapılması ve daha sonraki aşamaların uygulanmasında bazı eksiklikler gözlenmektedir.

Alan yazında daha çok psikolojik test ve ölçeklerin test geliştirme süreçlerine uygunluğu ile ilgili çalışmalara rastlanmakta olup (Çüm ve Koç, 2013; Şahin ve Boztunç Öztürk, 2018; Karadağ, 2011; Mor Dirlik, 2014; Öztürk, Eroğlu ve Kelecioğlu, 2015) başarı testlerinin bu hususta incelenmesine yönelik çalışmaların sayısı oldukça azdır (Boyraz, 2018; Mutluer ve Yandı, 2012; Karadağ, 2011). Araştırmacıların kendi geliştirdikleri başarı testlerinde bile test geliştirme aşamalarına yeterince özen gösterilmediği görülmüştür. Mutluer ve Yandı (2012)' nın 2010-2012 yılları arasında yapılan lisansüstü tezlerin incelenmesine yönelik çalışmasında 50 adet tezin yalnızca 16'sında test geliştirme aşamalarının tamamının gerçekleştirildiği görülmüştür. Boyraz (2018) 62 doktora tezinde ölçme aracı olarak kullanılan başarı testlerini incelemiş ve bu inceleme sonucunda başarı testi geliştirmenin her aşamasında yetersizlikler olduğu sonucuna ulaşmıştır.

Sonuç olarak bu çalışmada başarı testi geliştirme sürecinde ön uygulama sonrasında hesaplanan madde ve test istatistiklerinin uzmanlardan elde edilen sonuçlarla ne düzeyde tutarlı olduğu belirlenmeye çalışılacaktır. Uzmanların test geliştirme sürecindeki etkilerinin belirlenmesi, ön uygulamanın gizlilik, katılımcı sayısının az olduğu vb. durumlarda uzman görüşlerine dayalı test geliştirme sürecine de kaynaklık edeceği düşünülmektedir.

Yöntem

Bu araştırmada test geliştirme sürecinde ön uygulama sonrasında hesaplanan madde ve test istatistikleri ile uzman görüşlerinden elde edilen sonuçların detaylı incelenmesi ve karşılaştırılması yapıldığından çalışma bu yapısı itibariyle betimsel bir araştırmadır (Büyüköztürk ve diğ., 2020).

Çalışma grubunu 2018-2019 eğitim-öğretim yılında bir devlet üniversitenin eğitim fakültesinde üçüncü sınıfta öğrenim gören ve ölçme ve değerlendirme dersine 14 hafta devam eden toplam 165 öğretmen adayı oluşturmaktadır. Çalışma grubunun %30'unu (n=50) okul öncesi öğretmenliği, %17'sini (n=27) kimya öğretmenliği, %53'ünü de (n=88) İngilizce öğretmenliği bölümlerinde okuyan öğretmen adayları oluşturmaktadır. Uzman görüşü olarak araştırmada ölçme ve değerlendirme alanında doktora derecesi almış 11 öğretim elemanının görüşleri alınmıştır.

Araştırmacılar tarafından geleneksel ölçme araçları, alternatif ölçme araçları, madde istatistikleri, test istatistikleri ve test puanlarının yorumlanması olmak üzere belirlenen beş konu alanını kapsayan 12 çoktan seçmeli maddeden oluşan bir başarı testi geliştirilmiştir. Katılımcıların testte yer alan maddelere vermiş oldukları cevaplar doğrultusunda testin yapı geçerliğini belirlemek amacıyla öncelikle tetrakorik korelasyon matrisine dayalı açımlayıcı faktör analizi hesaplanmıştır. Hesaplama öncesinde KMO testinin 0,91 ve Barlet's testinin anlamlı bulunması neticesinde veri setinin faktör analizine uygun olduğu tespit edilmiştir.

Açımlayıcı faktör analizi sonucunda öz değeri 1'den büyük iki faktör olduğu belirlenmiştir. Öz değeri 1'den büyük olan iki faktör yer alsa da testte yer alan maddelerin tek bir faktör altında toplandığı görülmüştür. Ayrıca testte yer alan maddelerin faktör yük değerlerinin 0,616 ile 0,844 arasında değişiklik gösterdiği, maddelerin açıklayıcılıklarının yüksek düzeyde olduğu bulunmuştur. 12 maddeden oluşan testin toplam varyansının %57'sine açıklık getirdiği belirlenmiştir.

Maddelerin niteliği ile ilgili en çok kullanılan göstergeler madde güçlük indeksi ve madde ayırt edicilik indeksidir (Özçelik, 2014). Araştırmada, uzmanlara ön uygulamanın yapılacağı örneklem grubu tanıtılarak başarı testindeki her bir madde için hem ayırt edicilik (1=negatif, 2=düşük, 3=orta, 4=iyi, 5=çok iyi) hem de güçlük (1=çok zor, 2=zor, 3=orta, 4=kolay, 5=çok kolay) düzeylerinin tahmin edilmesi istenmiştir. 165 öğretmen adayının katıldığı ön uygulamadan elde edilen sonuçlar 1-5 olarak kategorilendirilmiştir. Daha sonra ön uygulama sonucu elde edilen madde ve test istatistikleri ile 11 uzmandan elde edilen ortalama tahminler arasındaki ilişki incelenecektir. Ayrıca her bir uzmanın görüşü ile ön uygulamadan elde edilen sonuçlar ayrı ayrı değerlendirilecektir.

Sonuçlar

Ön uygulama sonrasında hesaplanan madde ve test istatistiklerinin uzman görüşlerinden elde edilen ortalama tahminler ile tutarlı olması, teorik olarak istenen bir sonuçtur. Fakat uzmanların genel olarak maddeleri olduğundan daha kolay ya da zor ve daha düşük ya da yüksek ayırt edicilikte değerlendirdikleri ile ilgili yorumlara ulaşılabilir. Uzmanların niteliği, madde yazma konusundaki deneyimleri, maddelerin içeriği, alt testlerin özellikleri gibi değişkenler tahminler üzerinde oldukça etkili olabileceğinden bu etkilerin incelendiği çalışmaların yapılması önerilebilir. Elde edilen sonuçlar, madde ve test istatistiklerini doğru tahmin etmek için kaç uzmandan görüş alınması gerektiği hakkında bilgi verebilir ve puanlayıcılar arası uyum katsayısı da hesaplanabilir. Ayrıca testin kapsamını oluşturan geleneksel ölçme araçları, alternatif ölçme araçları, madde istatistikleri, test istatistikleri ve test puanlarının yorumlanması olmak üzere belirlenen beş konu alanına göre de yorumlar yapılabilir.

Kaynaklar

Boyraz, C. (2018). Investigation of achievement tests used in doctoral dissertations department of primary education (2012-2017). *Inonu University Journal of the Faculty of Education*, *19*(3), 14-28. https://doi.org/10.17679/inuefd.327321

Büyüköztürk, Ş., Çakmak, E. K., Akgün, O. E., Karadeniz, Ş. ve Demirel, F. (2020). *Bilimsel araştırma yöntemleri.* Pegem Akademi Yayıncılık.

Crocker, L., and Algina, J. (1986). *Introduction to classical and modern test theory.* Cengage Learning.

Çüm, S. ve Koç, N. (2013). Türkiye'de psikoloji ve eğitim bilimleri dergilerinde yayımlanan ölçek geliştirme ve uyarlama çalışmalarının incelenmesi. *Eğitim Bilimleri ve Uygulama, 12*(24), 115-135.

Karadağ, E. (2011). Eğitim bilimleri doktora tezlerinde kullanılan ölçme araçları: Nitelik düzeyleri ve analitik hata tipleri. *Kuram ve Uygulamada Eğitim Bilimleri, 11*(1), 311-334.

Mor-Dirlik, E. (2014). Ölçek geliştirme konulu doktora tezlerinin test ve ölçek geliştirme standartlarına uygunluğunun incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 5*(2), 62-78. https://doi.org/10.21031/epod.63138

Mutluer, C. ve Yandı, A. (2012, Eylül). *Türkiye'deki üniversitelerde 2010-2012 yıllarınarasında yayımlanan tezlerdeki başarı testlerinin incelenmesi* [Sözlü bildiri]. III Ulusal Eğitim ve Psikolojide Ölçme ve Değerlendirme Kongresi, Abant İzzet Baysal Üniversitesi, Bolu, Türkiye.

Özçelik, D. A. (2009). *Test hazırlama kılavuzu* (4. baskı). Pegem Akademi Yayıncılık.

Özçelik, D. A. (2014). *Öğrenme öğretim ve değerlendirme ile ilgili bir sınıflama.* Pegem Akademi Yayıncılık.

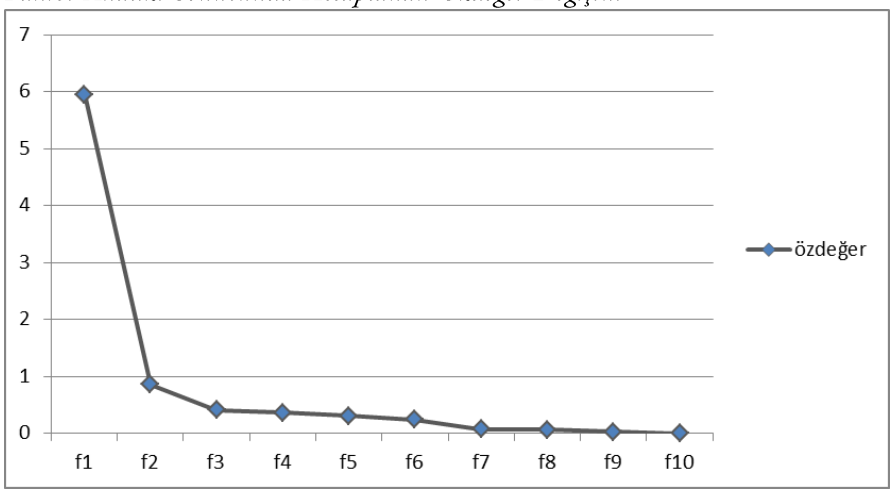
Öztürk, N. G., Eroğlu, M. G. ve Kelecioğlu, H. (2015). *Eğitim alanında yapılan ölçek uyarlama makalelerinin incelenmesi. Eğitim ve Bilim, 40*(178), 123-137. http://dx.doi.org/10.15390/EB.2015.4091

Şahin, M. ve Boztunç-Öztürk, N. (2018). Eğitim alanında ölçek geliştirme süreci: Bir içerik analizi çalışması. *Kastamonu Eğitim Dergisi, 2*(1), 191-199. https://doi.org/10.24106/kefdergi.375863

Ek-1

Şekil 1

Faktör Analizi Sonucunda Hesaplanan Özdeğer Değişim



Çok boyutlu madde tepki kuramı modellerinde yetenek parametresi değişmezliği

Gökhan Kumlu, Çiğdem Reyhanlıoğlu ve Nuri Doğan

Anahtar kelimeler: Çok boyutlu madde tepki kuramı, parametre değişmezliği, simülatif araştırma, yetenek dağılımı, yetenek parametresi

Giriş

Eğitimde ve psikolojide gerçekleştirilen ölçme işlemleri karmaşık süreçleri içinde barındırmaktadır. Bu karmaşık süreçleri standartlaştırmak amacıyla farklı ölçme kuramları geliştirilmiştir. Bu kuramların başlıcaları Klasik Test Kuramı (KTK), Madde Tepki Kuramı (MTK) ve Sato Test Kuramı'dır (STK). Bunlardan MTK ve STK modern test teorileri arasında yer almaktadır.

Klasik kuram ile modern test teorilerini birbirinden ayıran en önemli özelliklerinden biri madde parametrelerinin gruba bağlı olarak değişme/değişmeme durumudur. Klasik test teorisinde madde ve test istatistikleri gruba bağlı olarak hesaplanır (Hambleton ve diğ., 1991). Bu klasik kuramın önemli bir sınırlılığıdır. Bir gruba bir testin uygulanmasının ardından testin sonuçları doğrultusunda öğrencilerin testte zorlandığı görülmüş olsun. Bu durumda test zor olduğu için mi öğrenciler zorlanmıştır, yoksa öğrenciler düşük yetenek düzeyine sahip oldukları için mi zorlanmıştır? Bu durumu Hambleton ve diğ. (1991) madde özellikleri ile grup özelliklerinin birbirinden ayrılamaması şeklinde ifade etmiştir. Klasik kuramın bu sınırlılığına çözüm getirmiş olması modern test teorilerinin en önemli savlarından biridir (DeMars, 2010). Modern test teorileri arasında en yaygın kullanılan kuram MTK'dır. MTK temelde bireylerin maddelere verdikleri tepkiler ile o maddeler ile ilintili olan yetenekleri arasında matematiksel bir ilişki kuran olasılıksal bir teoridir (Faulkner-Bond ve Wells, 2016). Bu teoriyi esas alarak kullanılan modeller MTK modelleri olarak bilinir.

Geliştirilmiş olan ilk MTK modelleri tek boyutlu MTK modelleridir (Embretson ve Reise, 2000). Bu modeller tek boyutluluk özelliği gösteren maddelerden elde edilen verilere uygulanabilmektedir. Maddelerin tek boyutlu olması, gözlenen değişkenlerin tek bir örtük özelliğin bir fonksiyonu olmasını ifade etmektedir (de Ayala, 2009). Eğitim ve psikolojide ölçülmesi hedeflenen özelliklerin karmaşıklığı göz önünde bulundurulduğunda gözlenen değişkenleri tek bir örtük özellikle açıklamak yeterli olmamaktadır. Bu nedenle gözlemleri birden fazla örtük özellik ile aynı anda açıklayan modellerin geliştirilmesi ihtiyacı doğmuştur. Bu ihtiyaç doğrultusunda çok boyutlu MTK modelleri geliştirilmiştir

(Reckase, 2009). Çok boyutlu MTK modelleri iki kategorili veriler için faktör analitik yöntemleri uygulama sürecinde ortaya çıkan bazı problemlere çözüm getirmek amacıyla gerçekleştirilen çalışmalar sonucunda elde edilmiştir (Reckase, 2009).

MTK'da parametre değişmezliği (DeMars, 2010) madde parametrelerinin yetenek parametrelerinden ve benzer şekilde yetenek parametrelerinin de madde parametrelerinden bağımsız bir şekilde kestirilmesi anlamına gelir. MTK varsayımlarının karşılanması durumunda parametre değişmezliğinin sağlaması beklenir (Doğan ve Kılıç, 2017). Ancak alan yazında varsayımlar karşılanmış olsa da parametre değişmezliğinin kesin olarak sağlandığı ifade edilmemektedir. Bu durumun nedeni parametre değişmezliğinin farklı faktörlerden etkilenmesidir. Çok boyutlu MTK için alan yazın çalışmaları incelendiğinde parametre değişmezliğinin etkilendiği faktörlerden birinin örneklem büyüklüğü olduğu görülmektedir (Bolt ve Lall, 2003; Knoll ve Berger, 1991; Köse, 2010; Reyhanlıoğlu ve Doğan, 2020; Sünbül, 2011). Daha açık bir ifade ile parametre değişmezliği uygun örneklem büyüklükleri için sağlanabilir. Çok boyutlu MTK için parametre değişmezliğinin etkilendiği bir başka faktör de test uzunluğudur (de la Tore ve Patz, 2005; Koğar, 2014; Köse, 2010; Zhang, 2008). Örneklem büyüklüğünde olduğu gibi parametre değişmezliğinin sağlanması için de uygun test uzunluklarının kullanılması gerekir. Tek boyutlu MTK için parametre değişmezliğini etkileyen faktörlerden birinin de yetenek dağılımı olduğu görülmektedir (Doğan, 2002). Çok boyutlu MTK'da, yetenek dağılımını parametre kestirimleri açısından odağına alan bir çalışmaya rastlanmıştır (Kirisci ve diğ., 2001). Bu çalışmada normal, sağa çarpık ve basık yetenek dağılımlarından elde edilen parametreler arasında anlamlı bir farklılığın olmadığı belirlenmiştir. Ancak çok boyutlu MTK'da parametre değişmezliğinin boyutlara göre farklılık gösteren yetenek dağılımından etkilendiğini açıkça ortaya koyan bir çalışmaya rastlanmamıştır.

Bu alana yönelik çalışmaların alan yazında sınırlı olmasından dolayı, bu çalışmanın alan yazına sağlayacağı katkı açısından önemli olduğu düşülmektedir. Bu doğrultuda araştırmanın problem cümlesi "Çok boyutlu Madde Tepki Kuramı (ÇBMTK) modelleri birey yetenek parametrelerinin boyutlarda farklı dağılıma sahip olduğunda parametre değişmezliği nasıl etkilenmektedir?" şeklinde yapılandırılmıştır.

Yöntem

Çok boyutlu MTK modellerinde bireyin yetenek parametresinin boyutlarda farklı dağılım göstermesi durumunda parametre değişmezliğine etkisini belirlemek amacıyla araştırma simülatif bir çalışma olarak yürütülmüştür. İki boyutlu veri seti ile yürütülen simülasyon çalışmasına ilişkin koşullar Tablo 1'de verilmiştir.

Tablo 1

Simülasyon Koşulları

Yetenek Dağılımı	Örneklem Büyüklüğü	Test Uzunluğu	Alt Boyutlardaki Madde Sayısı Oranı	Boyutlar Arası Korelasyon	MTK Modeli
Normal Dağılım (N (0, 1); N (0, 1))	1000	40	%50-%50	0.25	1PLM
Normal-Çarpık Dağılım (N (0, 1); PÇ (-1.5, 1))	3000	80	% 75-%25	0.50	2PLM
Zıt Yönde Çarpık Dağılım (PÇ (-1.5, 1); NÇ (1.5, 1))					3PLM
Aynı Yönde Çarpık Dağılım (NÇ (1.5, 1); NÇ (1.5, 1))					

Bu çalışmada veriler iki boyutlu olarak üretilmiştir. Boyutlardaki yetenek dağılımları (i) birinci boyutta normal dağılım, ikinci boyutta normal dağılım olmak üzere *Normal Dağılım*; (ii) birinci boyutta normal dağılım, ikinci boyutta pozitif çarpık dağılım olmak üzere *Normal-Çarpık Dağılım*; (iii) birinci boyutta pozitif çarpık dağılım, ikinci boyutta negatif çarpık dağılım olmak üzere *Zıt Yönde Çarpık Dağılım*; (iv) birinci boyutta negatif çarpık dağılım, ikinci boyutta negatif çarpık dağılım olmak üzere *Aynı Yönde Çarpık Dağılım* olarak veriler üretilmiştir. Verilerin üretimi aşamasında örneklem büyüklüğü ve test uzunluğu için iki farklı koşul ele alınmıştır (1000 ve 3000 kişi; 40 ve 80 madde). Test uzunluğuna ilişkin madde sayısı alt boyutlarda %50-%50 (20-20 madde ve 40-40 madde) ve %75-%25 (30-10 madde ve 60-20 madde) olarak iki farklı oranda belirlenmiştir. Veri setinde yer alan iki boyut arasındaki ilişki düzeyi 0.25 ve 0.50 olarak saptanmıştır.

Çok boyutlu MTK modellerine göre üretilen verilerde, 1PL model için madde ayırt edicilik düzeyini gösteren *a* parametresi alt testlerde sabit tutularak değerleri 0.5 ile 2 arasında değişen, ortalaması 1.25 ve standart sapması 0.5 olan tek biçimli (uniform) dağılımdan üretilmiştir. 2PL model için *a* parametresi aynı kalmak koşulu ile güçlük düzeyini gösteren *b* parametresi -3 ile +3 arasında değişen, ortalaması 0 ve standart sapması 1 olan normal dağılımdan üretilmiştir. Çok boyutlu MTK modeline göre veri yapısının iki boyutlu olmasından dolayı tek boyutlu modellerde kullanılan *b* parametresi yerine çok boyutlu veri üretim aşamasında kesişim (intercept) parametresi olan *d* parametresi kullanılmıştır. *d* parametresine ilişkin değerler üretilen *a* ve *b* parametrelerinden yararlanılarak aşağıda verilen formül yardımıyla elde edilmiştir (Reckase, 2009).

$$MDIFF = \frac{-d_i}{MDISC}$$

3PL model için *a* ve *b* parametreleri aynı kalmak koşulu ile şans başarısı düzeyini gösteren *c* parametresi 0.05 ile 0.25 arasında değişen, ortalaması 0.15 ve standart sapması 0.05 olan tek biçimli (uniform) dağılımdan üretilmiştir.

Yetenek dağılımları 4, örneklem büyüklüğü 2, test uzunluğu 2, alt boyutlardaki madde sayısı oranı 2, boyutlar arası korelasyon 2, MTK modeli 3 farklı koşul olmak üzere toplamda 192 koşul ele alınmıştır.

Simülatif çalışmalara ilişkin sonuçların tutarlı olması açısından her veri seti için alan yazında önerilen (Kim & Kolen, 2006) 100 tekrar ile analizler yürütülmüştür. Dolayısıyla araştırma 19200 veri seti ile gerçekleştirilmiştir.

Verilerin üretilmesinde R yazılımından, analizlerinde ise IRTPRO 4.2 programından yararlanılmıştır. Verilerin analizinde kontrol altına alınan koşullar çerçevesinde yetenek parametresinin değişmezliğinin sağlanıp sağlanmadığı kontrol edilmiştir.

Sonuçlar

İki boyutlu veri seti üzerinden gerçekleştirilen Çok Boyutlu Madde Tepki Kuramı modellerine göre parametre kestirimi sonucunda yetenek parametresi değişmezliğinin 1PLM'den 2PLM ve 3PLM'ye doğru gidildikçe parametre sayısının artmasıyla modelin daha fazla karmaşıklaşmasından dolayı değişmezliğin sağlanmasının zorlaşması beklenmektedir. Ayrıca boyutlardaki yetenek dağılımlarının çarpıklaşmasının yetenek parametresi değişmezliğini sağlamayı zorlaştırması beklenmektedir. Diğer yandan bu çalışma kapsamında alan yazındaki sonuçlardan yararlanarak dikkate alınan koşullardan test uzunluğunun artması, örneklem büyüklüğünün artması ve boyutlar arasındaki korelasyon değerlerinin 0.25'ten 0.50'ye doğru yükselmesi boyutlara ilişkin yetenek dağılımındaki çarpıklığın etkisini minimize ederek yetenek parametresi değişmezliğini olumlu yönde etkilemesi beklenmektedir. Genel sonuç olarak da hangi koşulların kombinasyonu altında yetenek parametre kestirimlerinin standart hatalarının minimize edildiğine ilişkin bilgilere ulaşılması beklenmektedir.

Kaynaklar

Bolt, D. M., and Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using markov chain monte carlo. *Applied Psychological Measurement, 27*(6), 395–514. https://doi.org/10.1177/0146621603258350

de Ayala, R. J. (2009). *The theory and practice of item response theory.* The Guilford Press.

de la Tore, J., and Patz, R.J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics, 30*(3), 295-311. https://doi.org/10.3102/10769986030003295

DeMars, C. (2010). *Item response theory.* Oxford University Press.

Doğan, N. (2002). *Klasik test kuramı ve örtük özellikler kuramının örneklemler bağlamında karşılaştırılması.* (Tez No: 82042). [Doktora tezi, Hacettepe Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Doğan, N. ve Kılıç, A. F. (2017). Madde tepki kuramı yetenek ve madde parametreleri kestirimlerinin değişmezliğinin incelenmesi. Ö. Demirel ve S. Dinçer (Eds.), *Küreselleşen dünyada eğitim* içinde (ss. 297-314). Pegem Akademi.

Embretson, S. E., and Reise, S. P. (2000). *Item response theory for psychologists.* Lawrence Erlbaum Associates, Inc.

Faulkner-Bond, M., and Wells, C. S. (2016). A brief history of and introduction to item response theory. In C. S. Wells, and M. Faulkner-Bond (Eds.), *Educational measurement: From foundations to future* (pp. 107-125). The Guilford Press.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of item response theory.* Sage Publications.

Kim, S., and Kolen, M.J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education, 19*(4), 357-381.

Kirisci, L., Hsu, T., and Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25*, 146-162. https://doi.org/10.1177/01466210122031975

Knoll, D. L., and Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate behavioral research, 26*(3), 457-477. https://doi.org/10.1207/s15327906mbr2603_5

Koğar, H. (2014). *Madde tepki kuramının farklı uygulamalarından elde edilen parametrelerin ve model uyumlarının örneklem büyüklüğü ve test uzunluğu açısından karşılaştırılması* (Tez No. 378546) [Doktora tezi, Hacettepe Üniversitesi]. Yükseköğretim Kurumu Tez Merkezi.

Köse, A. (2010). *Madde tepki kuramına dayalı tek boyutlu ve çok boyutlu modellerin test uzunluğu ve örneklem büyüklüğü açısından karşılaştırılması* (Tez No: 285760) [Doktora tezi, Ankara Üniversitesi). Yükseköğretim Kurulu Ulusal Tez Merkezi.

Reckase, M.D. (2009). *Multidimensional item response theory.* Springer Sience & Business Media.

Reyhanlıoğlu, Ç. ve Doğan, N. (2020). An analysis of parameter ınvariance according to different sample sizes and dimensions in parametric and nonparametric ıtem response theory. *Journal of Measurement and Evaluation in Education and Psychology, 11*(2), 98-112. https://doi.org/10.21031/epod.584977

Sünbül, Ö. (2011). *Çeşitli boyutluluk özelliklerine sahip yapılarda, madde parametrelerinin değişmezliğinin klasik test teorisi, tek boyutlu madde tepki kuramı ve çok boyutlu madde tepki kuramı çerçevesinde incelenmesi* (Tez No: 274531) [Doktora tezi, Mersin Üniversitesi] Yükseköğretim Kurumu Tez Merkezi.

Zhang, B. (2008). Application of unidimensional item response models to tests with items sensitive to secondary dimensions. *The Journal of Experimental Education, 77*(2), 147-166. https://doi.org/10.3200/JEXE.77.2.147-166

Değişen madde fonksiyonu belirleme yöntemleri performanslarının karşılaştırılması: Mantel-haenszel, lojistik regresyon ve Lord ki-kare

Münevver Başman

Anahtar kelimeler: Değişen madde fonksiyonu, lojistik regresyon, Lord ki-kare, Mantel-Haenszel

Giriş

Testler; beceriler, yetenekler ve diğer psikometrik özellikler gibi gizil özellikleri değerlendirmek için kullanılan sistematik süreçleri içeren araçlardır (Linn ve Gronlund, 2000). Testlerden elde edilen sonuçlar ile farklı özelliklere sahip gruplar birbirleriyle karşılaştırılabilmekte ve karşılaştırma sonuçlarına göre çeşitli kararlar alınabilmektedir. Ancak test maddeleri bir grup lehine yanlıysa ve adil değilse testin geçerliği etkilenir (Kane, 2006; Messick, 1989). Bu nedenle bireylerin hayatında önemli bir yere sahip olan testlerin güvenirlik ve geçerlik çalışmaları yapılması gerekmektedir.

Testin geçerliğini sağlamanın bir yolu, tüm maddelerin farklı birey grupları arasında benzer şekilde çalışmasıdır. Ancak, eşit yetenek düzeyine sahip farklı gruplardan bireyler aynı test maddesi üzerinde farklı performans gösterdiğinde, değişen madde fonksiyonu (DMF) oluşur. Diğer bir deyişle DMF, yeteneği aynı olan alt grupların maddeyi doğru cevaplama olasılığının farklılaşmasıdır (Gao, 2019; Hambleton ve diğ., 1991).

Bir maddenin DMF'ye sahip olup olmadığını belirlemek için bir takım istatistiksel yöntemler geliştirilmiştir. DMF belirleme yöntemleri temel olarak, gözlenen puan grubunu dikkate alan Klasik Test Kuramı (KTK) ve gizil değişken grubunu dikkate alan Madde Tepki Kuramı'na (MTK) göre sınıflandırılır. Bu yöntemler; örneklem büyüklüğü gereksinimleri, grup özellik dağılımlarındaki farklılıklardaki duyarlılık, I. Tip hata oranları ve istatistiksel güç değerleri, farklı DMF türlerindeki duyarlılık gibi bazı koşullar altında diğer yöntemlerden daha avantajlı olabilmektedir. Bu nedenle DMF analizlerinin kullanıldığı çalışmalarda birden fazla yöntem kullanılması önerilmektedir (Camilli ve Shepard, 1994; Osterlind ve Everson, 2009).

Yapılan çalışmalar incelendiğinde DMF belirleme yöntemlerinin bazı değişkenler göz önünde bulundurularak performanslarının incelendiği görülmüştür. Bu araştırmanın amacı üç parametreli (3PL) modeline dayalı tek biçimli DMF varlığında örneklem büyüklüğü, test uzunluğu ve DMF oranı

değiştiğinde I.Tip hata oranı ve istatistiksel güç oranları kullanılarak Mantel-Haenszel (MH), Lojistik regresyon (LR) ve Lord'un ki-kare (LORD) yöntemlerinin karşılaştırılmasıdır.

Yöntem

Referans ve odak grupları için örneklem büyüklükleri, testin uzunluğu ve DMF gösteren maddelerin oranı gibi bağımsız değişkenlerin manipüle edildiği üç DMF yönteminin güç ve I. Tip hata oranlarını incelemek için bir Monte Carlo simülasyonu yapılmıştır.

Veriler türetilirken gerçek verilerin (kağıt-kalem uygulaması yapılan TIMSS Matematik testi) güçlük ve ayırt edicilik parametre dağılımları kullanılmıştır. Şans parametresi .20 olarak sabit tutulmuştur.

Simülasyon kapsamında örneklem büyüklüğü (500, 2000), test uzunluğu (10, 20, 30), DMF içeren madde yüzdesi (%10, %20) manipüle edilen koşullar olarak ele alınırken, DMF biçimi tek biçimli DMF ve 3PL model sabit koşullar olarak ele alınmıştır. 12 koşul kapsamında gerçekleştirilen araştırma için 100 tekrar gerçekleştirilmiştir. Böylece toplam 1200 veri üretilmiştir. Her bir veri seti için üç DMF yöntemi ile DMF analizleri yapılmıştır. Veriler WinGen3 programıyla oluşturulmuş ve R 4.0.2 istatistik yazılımında bulunan difR paketi ile analiz edilmiştir.

Yöntemlerin performansını karşılaştırmak için I.Tip hata ve güç oranları kullanılmıştır. Bradley'e (1978; aktaran Hidalgo ve diğ., 2016) göre I. Tip hata oranı .025 ile .075 arasında olmalıdır. Yöntemlerin gücünün yeterli olabilmesi için en az .80 olması gerekir ve bu ölçütler alanyazında yaygın olarak kullanılmaktadır (Atar, 2007).

Belirlenen üç yöntemin I.Tip hata ve güç oranları arasında anlamlı farklılık olup olmadığı tek yönlü ANOVA ile incelenmiştir. Ele alınan faktörlerin yöntemler üzerindeki hem ana etkileri hem de etkileşim etkilerini belirlemek için de faktöriyel ANOVA yapılmıştır. İlgili analizlerin ve post hoc karşılaştırmaların istatistiksel anlamlılık bulguları incelenmiştir.

Sonuçlar

Yöntemlerin ortalama I. Tip hata oranlarının anlamlı olarak farklılık göstermediği bulunmuştur ($F(2, 3597)= 3.036$, $p> .05$). Güç oranları incelendiğinde ise yöntemler arasında anlamlı farklılık olduğu görülmektedir ($F(2, 3597)= 31.721$, $p< .05$). Post hoc analizi sonucunda LORD yönteminin güç oranının MH ve LR yöntemlerinin güç oranlarından anlamlı şekilde daha az olduğu bulunmuştur.

Örneklem büyüklüğüne göre tüm yöntemler karşılaştırıldığında, büyük örneklemde I. Tip hata oranlarının küçük örneklemden daha yüksek olduğu görülmüştür. Küçük örneklem için tüm yöntemlerin güç oranları her koşulda .80'in altında ve büyük örneklem için MH ve LR yöntemlerinin güç oranları her koşulda .80'in üzerindedir. Örneklem büyüklüğüne göre tüm yöntemler karşılaştırıldığında, büyük örneklemde güç oranlarının daha yüksek olduğu görülmüştür. Yöntemlerin

I.Tip hata ve güç oranlarında ilgilenilen faktörlerin ana etkileri ve etkileşim etkilerinin de olduğu görülmektedir.

Bu çalışmada 3PL, tek biçimli DMF, çeşitli koşullar ve üç farklı yöntem ele alınmıştır. Benzer koşullarda farklı DMF yöntemleri incelenebilir. Farklı koşullar ortaya konulup (farklı yetenek dağılımı, eşit olmayan örneklem büyüklüğü gibi) bu koşullarda yöntemlerin performansları karşılaştırılabilir.

Kaynaklar

Atar, B. (2007). *Differential item functioning analyses for mixed response data using IRT likelihood-ratio test, logistic regression, and GLLAMM procedures* (Tez No. 3263842) [Doctoral dissertation, University of Florida State] ProQuest Dissertations & Theses Global.

Camilli, G. (2006). Test fairness. In Brennan, R. L. (Ed). *Educational Measurement* (pp. 221–257). American Council on Education.

Camilli, G., and Shepard, L. A. (1994*). Methods for identifying biased test items.* Sage Publications.

De Ayala, R. J. (2009). *The theory and practice of item response theory.* Guilford Press.

DeMars, C. E. (2009). Modification of the Mantel-Haenszel and logistic regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics, 34*(2), 149-170. https://doi.org/10.3102/1076998607313923

Dorans, N. J., & Holland, P. W. (1992). *DIF detection and description: Mantel-Haenszel and standardization 1, 2.* ETS Research Report Series.

Fidalgo, A. M., Mellenbergh, G. J., and Muñiz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online*, *5*(3), 43-53.

Gao, X. (2019). *A comparison of six DIF detection methods.* [Master's Theses, Shandong Universit]. https://opencommons.uconn.edu/gs_theses/1411

Gierl, M. J., Jodoin, M. G., and Ackerman, T. A. (2000, 24-27 April). *Performance of Mantel-Haenszel, simultaneous item bias test, and logistic regression when the proportion of DIF items is large.* [Paper presentation]. The Annual Meeting of the American Educational Research Association (AERA) New Orleans, Louisiana, USA.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Sage.

Hidalgo, M. D., López-Martínez, M. D., Gómez-Benito, J., and Guilera, G. (2016). A comparison of discriminant logistic regression and ıtem response theory likelihood-ratio tests for differential ıtem functioning (IRTLRDIF) in polytomous short tests. *Psicothema*, *28*(1), 83-88. https://doi.org/10.7334/psicothema2015.142

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17– 64). American Council on Education/Praeger

Keklik, D. E. (2014). Comparison of Mantel-haenszel and logistic regression techniques in detecting differential item functioning. *Journal of Measurement and Evaluation in Education and Psychology*, *5*(2), 12-25. https://doi.org/10.21031/epod.71099

Linn, R. L., and Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.). Upper Saddle River.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). MacMillan.

Osterlind, S. J., and Everson, H. T. (2009). *Differential item functioning.* Sage Publications, Inc.

Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement, 18*(1), 15–25. https://doi.org/10.1177/014662169401800102

Bireyselleştirilmiş bilgisayarlı sınıflama testlerinde yetenek kestirim yöntemlerinin farklı kesme noktalarına göre karşılaştırılması

Sümeyra Soysal ve Ceylan Gündeğer

Giriş

Bireyselleştirilmiş bilgisayarlı test uygulamalarının (BBT) temel amacı, akıllı bir madde seçim algoritması kullanarak maddeleri sınava giren kişinin yeteneğiyle eşleştirmektir. BBT, testi sınava giren kişiye göre ayarlayarak, geleneksel kağıt ve kalem testlerine göre daha verimli testler sağlar (Bennett, 2001, Boo ve Vispoel, 1998; Schmit ve Ryan, 1993 Wang ve diğ., 2008). Bazen testler kişilerin genel yetenek düzeyini belirlemekten ziyade onları geçti-kaldı, başarılı-başarısız veya daha fazla kategoride sınıflama amacı taşıyabilir. Bu durumda BBT'nin özel bir uygulaması olan Bireyselleştirilmiş Bilgisayarlı Sınıflama Testlerinin (BBST) kullanımı tercih edilebilir. Weiss ve Kinsbury (1984) "tepki modeli, madde havuzu, başlangıç kuralı, madde seçim yöntemi, yetenek kestirim yöntemi ve testi sonlandırma kuralı" olmak üzere BBT'nin altı bileşenden oluştuğunu belirtmiştir. BBST'de ise bu ilk beş bileşen sabit kalırken sonlandırma kuralı sınıflandırma kriterleri ile sağlanır. İki kategorili sınıflandırma durumları için BBST uygulamalarında, testi sonlandırmak için en fazla tercih edilen iki yaklaşımdan biri Ardışık olasılık oran testi (AOOT), diğeri ise güven aralığı (GA) yöntemidir. Reckase (1983), AOOT'yi sınava giren kişinin yetenek parametresinin kesme puanının üzerinde belirli bir noktaya veya kesme puanının altında belirli bir noktaya eşit olduğuna dair bir hipotez testi olarak tanımlar. GA, yönteminde ise sınava giren kişinin yetenek parametresi için kestirilen güven aralığı kesme puanının tamamen üzerinde veya altında olduğunda test sona erdirilir (Thompson, 2007). BBST'de madde seçim yöntemleri de iki kategoride sınıflanmaktadır: (i) bireyin geçici olarak kestirilen yetenek düzeyinde en yüksek maddenin seçimi (KY) veya (ii) seçilen kesme noktasında en yüksek bilgiyi veren maddenin seçimi (KN) (Thompson, 2007). Alanyazında AOOT yöntemine göre GA yönteminin sınıflama doğruluğu yüksek daha kısa testlerle sınıflama yapılabildiğine dair sonuçlar mevcuttur (Ayan, 2018; Eggen ve Straetmans, 2000; Gündeğer ve Doğan, 2018; Gündeğer ve Soysal, 2021; Nydick ve diğ., 2012) Ayrıca, Thompson (2009) AOOT yönteminin KN madde seçiminde, GA yönteminin ise KY madde seçiminde daha doğru ve yeterli sonuçlar verdiğini belirtmektedir. Alanyazında kesme noktasının BBST üzerindeki etkisine dikkat çeken çalışmalara pek rastlanmamıştır. Önceki araştırmaların bu sonuçlarından yola çıkılarak bu çalışmada, BBST'de yetenek kestirimi (Maksimum Olabilirlik, Ağırlıklandırılmış Olabilirlik, Beklenen Sonsal Dağılım, Maksimum Sonsal Dağılım) ve farklı kesme noktalarının, kesme noktası temelline dayalı madde

seçimi ile GA sınıflama yöntemi performansının test uzunluğu, sınıflama doğruluğu, 1. ve 2. tip hata, RMSE üzerindeki etkisinin incelenmesi amaçlanmaktadır.

Yöntem

Bu çalışma, farklı yetenek kestirim yöntemleri ile kesme noktaları arasındaki ilişkiyi üretilmiş veri seti üzerinden incelemesi bakımından bir simülasyon çalışmasıdır. Çalışmada yetenek parametreleri N (0,1) dağılımından 1000 kişi üzerinden üretilecektir. Madde havuzu, faktör yükleri 0.30 ve üzerinde olacak şekilde sabitlenerek b parametresi N(0,1); a parametresi Lognormal [1.5, 0.5] ve c parametresi Beta (6,16) olacak şekilde, gerçek uygulamalara daha yakın bir değer olan 200 madde üzerinden 3 parametreli lojistik model temel alınarak ve 0.0 kesme noktasında en yüksek bilgiyi verecek şekilde üretilecektir.

Çalışmanın manipüle edilen bağımsız değişkenlerini, üç farklı kesme noktası ve dört farklı yetenek kestirim yöntemi olmak üzere toplam 12 koşul oluşturmaktadır. Kesme noktaları, Spray ve Reckase'in (1994) çalışmasında olduğu gibi -0.5, 0.0 ve 1.0 şeklinde belirlenirken; yetenek kestirim yöntemlerinden Maksimum Olabilirlik Kestirimi (Birnbaum, 1968), Ağırlıklandırılmış Olabilirlik Kestirimi (Warm, 1989), Beklenen Sonsal Dağılım (Bock ve Aitkin, 1981) ve Maksimum Sonsal Dağılım (Samejima, 1969) yöntemleri ele alınacaktır.

Araştırmada madde seçme yöntemi kesme noktası temelli Maksimum Fisher Bilgisi; sınıflama kriteri ise Güven Aralığı yöntemi olarak sabitlenecektir. BBST uygulaması R ortamında (R Core Team, 2013) maksimum test uzunluğu 50 olacak şekilde ve 25 replikasyonla gerçekleştirilecektir. Çalışmanın bağımlı değişkenleri olan test uzunluğu, sınıflama doğruluğu, 1. ve 2. tip hata, RMSE ve gerçek yetenek ile kestirilen yetenek arasındaki korelasyon (r) şeklinde belirlenmiştir. Bulgular bu değerlerin 25 replikasyondan elde edilen ortalamaları alınarak özetlenecektir.

Sonuçlar

Spray ve Reckase'in (1994) çalışmasında AOOT sınıflama yöntemi kullanılmış olup kesme noktasına göre sonuçlar arasında büyük bir farklılık olmadığı; bir başka ifadeyle sonuçlar arasında bir uyum olduğu görülmüştür. Bu çalışmada ise GA sınıflama yönteminde kesme noktasına göre önemli bir farklılık çıkmayacağı düşünülmektedir. Ancak bu çalışma, kesme noktası bakımından hiç çalışılmamış bir sınıflama yöntemini (GA) ele alması sebebiyle sonuçlar analizler tamamlanınca görülecektir. Bununla birlikte yetenek kestirim yöntemleri arasında sınıflama doğruluğu bakımından olmasa da, test uzunluğu, RMSEA, r ve özellikle 1. ve 2. tip hatalar bakımından bir farklılık olacağı düşünülmektedir. Test uzunluğu bakımından BSD yönteminin daha etkili olacağı düşünülmektedir. Sınıflama doğruluğu yetenek kestirim yöntemleri fark etmeksizin genellikle 0.90 ve üzerinde hesaplanmaktadır. Ancak 1. ve 2. tip hata BBST çalışmalarında pek üzerinde durulan bir konu olmadığı için sonuçlar bu açıdan önem arz etmektedir. Alanyazındaki çalışmalar incelendiğinde, gerçekte başarısız sınıfında olan öğrencilerin başarılı veya gerçekte başarılı sınıfında olan öğrencilerin başarısız sayılma durumlarına ilişkin hataların

ele alındığı bir BBST çalışması bulunmadığı görülmüştür. Bununla birlikte özellikle yüksek riskli testler için (kesme noktasının 1.0 olduğu koşullar) uygun yetenek kestirimi yönteminin ne olacağı sorusu analizler sonucunda cevaplanmış olacaktır. Çalışmanın bu bakımdan alanyazına katkı sağlayacağı düşünülmektedir.

Kaynaklar

Ayan, C. (2018). *Bilişsel tanı modelinde geleneksel ve bilgisayarlı sınıflamalı test uygulamalarının psikometrik özelliklerinin karşılaştırılması* (Tez No: 531700) [Doktora Tezi, Ankara Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Bennett, R. E. (2001). How the Internet will help large-scale assessment reinvent itself. *Education Policy Analysis Archive, 9*(5), 1-23. https://epaa.asu.edu/ojs/article/view/334/460

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.). *Statistical theories of mental test scores* (pp. 397-472). Addison-Wesley.

Bock, R. D., and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm. *Psychometrika, 46, 443-459.* https://doi.org/10.1007/BF02293801

Eggen, T. J. H. M. & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60*(5), 713-734. https://doi.org/10.1177/00131640021970862

Gündeğer, C. & Doğan, N. (2018). The effects of item pool characteristics on test length and classification accuracy in computerized adaptive classification testings. *Hacettepe University Journal of Education, 33*(4), 888-896. https://dergipark.org.tr/tr/pub/hunefd/issue/39869/472906

Gündeğer, C, & Soysal, S. (2021, June). *Güçlü ve zayıf tek boyutlu madde havuzlarının bireyselleştirilmiş bilgisayarlı sınıflama testi kriterleri üzerindeki etkisi.* The 4th Annual Meeting of International Conference on Data Science and Applications. Vancouver, British Columbia, Canada.

Nydick, S. W., Nozawa, Y. & Zhu, R. (2012, April). *Accuracy and efficiency in classifying examinees using computerized adaptive tests: An application to a large scale test.* The Annual Meeting of the National Council on Measurement in Education. Vancouver, British Columbia, Canada.

R Core Team (2013). *R: A language and environment for statistical computing* (version 3.0.1) [Computer Software]. http://www.R-project.org/

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237-254). Academic Press.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores.* Educational Testing Service.

Schmit, M. J., and Ryan, A. M. (1993). Test-taking disposition: A missing link? *Journal of Applied Psychology, 77*, 624-637. https://doi.org/10.1037/0021-9010.77.5.629

Spray, J. A., and Reckase, M. D. (1994, April 5-7). *The selection of test ıtems for decision making with a computer adaptive test* [Paper Presentation]. The Annual Meeting of the National Council on Measurement in Education. NewOrleans, USA.

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment, Research, and Evaluation, 12*(1), 1-13. https://doi.org/10.7275/fq3r-zz60

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69*(5), 778-793. https://doi.org/10.1177/0013164408324460

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427-450. https://doi.org/10.1007/BF02294627

Wang, S., Jiao, H., Young, M. J., Brooks, T., and Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*(1), 5-24. https://doi.org/10.1177/0013164407305592

Weiss, D. J., and Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*(4), 361-375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Dezavantajlı ilkokul öğrencilerinin okuduğunu anlama becerilerinin değerlendirilmesi sürecinde puanlayıcı yanlılığının incelenmesi

Yusuf Kızıltaş, Mehmet Şata ve Fuat Elkonca

Anahtar kelimeler: Dezavantajlı ilkokul öğrencileri, okuduğunu anlama, puanlayıcı yanlılığı, çok yüzeyli Rasch.

Giriş

Okuma, ön bilgilerin kullanıldığı, okuyucu-yazar arasında dinamik bir etkileşimi özünde barındıran, düzenli bir ortamda gerçekleştirilen bir anlam kurma süreci olarak değerlendirmektedir. Anlam kurmadan kasıt, okunan metin ve okuyucunun ön bilgileri arasında bir bağ oluşturmaktır (Akyol, 2019). Bu bağ, okuduğunu anlama sürecini başlatan ve hızlandıran önemli bir etkiye sahiptir. Okunan metni anlamayı, kavramayı ve metinden anlam çıkarmayı kapsayan okuduğunu anlama becerisi, okumanın temel amacını oluşturmaktır (Grabe ve Stoller, 2002; Hans ve Hans, 2015). Bu bakımdan okuduğunu anlama sürecinin yeterince tartışılması önem arz etmektedir. Okuduğunu anlamayı değerlendirmeyi bu tartışmada ayrı bir noktaya konumlandırmak gerekir.

Okunan metinden çıkarılan anlamın değerlendirilmesi önemlidir. Okuduğunu anlama sürecinin kalitesini, başarısını kısacası performansını ortaya çıkarmak ve dönüt vermek açısından gereklidir (Klingner ve diğ., 2007). Ancak değerlendirme sürecinin şeffaf ve adil olması kadar değerlendirmeyi sağlayacak araçların güvenilir ve geçerli olması da gerekmektedir. Öte yandan hedef kitlenin ve özelliklerinin de değerlendirme sürecini etkileyen bir faktör olduğu (Bergh, 2010) bilinmektedir. Dovidio ve diğ. (2002) da bu durumun öğrencilerin başarı düzeyini olumsuz etkilediğini belirtmektedirler. Dolayısıyla sınıf içerisinde okuduğunu anlamayı değerlendirmede öğretmenlerin ifade edilen dinamikleri ve hassasiyetleri göz önünde bulundurmaları etkili iletişim becerilerine (Taşdelen ve diğ., 2014) sahip olmaları gerekmektedir. Ancak öğretmenlerin zaman zaman başarılı/başarısız, tanılanmış/tanılanmamış, risk altındaki, dezavantajlı vb. öğrencilere yani hedef kitleye karşı değerlendirme sürecinde yanlı davrandıkları da bilinmektedir. Özellikle de sınıf içerisindeki dezavantajlı (İYEP programında yer alan, RAM tarafından özel öğrenme güçlüğü teşhisi konulan, Türkçeyi yabancı dil olarak öğrenen mülteci çocuklar, mevsimlik tarım işçisi çocuklar, geçici koruma altındaki çocuklar vb.) bireylere karşı değerlendirmelerde yanlılık çoğu zaman görülebilmektedir. Dezavantajlı öğrencilerin okuduklarını anlamalarını değerlendirme süreci de bu yanlılıktan etkilenebilmektedir. Sınıf içerisinde

puanlayıcı konumunda bulunan öğretmenlerin dezavantajlı öğrencilerin okuduklarını anlamalarını değerlendirmelerdeki yanlılıklarını tartışmak ve ortaya koymak bu açıdan önem arz etmektedir.

Alanyazın incelendiğinde doğrudan dezavantajlı ilkokul öğrencilerin okuduklarını anlamalarını değerlendirmede sınıf öğretmenlerinin yanlılığını inceleyen çalışmaların eksikliği açık bir şekilde görülmektedir. Milanowski (2017) de yaptığı araştırmasında dezavantajlı öğrencilerin düşük performanslarının/puanlarının, öğretmenlerin bir yanlılığının mı yoksa tamamen gerçeğin bir yansıması mı olduğunu sorgulamıştır. Ancak dezavantajlı öğrenciler üzerinde yapılan puanlayıcı yanlılığı ile ilgili farklı araştırmalara rastlanmıştır (Choi ve Young-Min, 2011; Farrokhi ve diğ., 2012; Wesolowski ve diğ., 2015). Özellikle de Türkiye'de bu anlamda herhangi bir çalışmaya rastlanmamış olması yapılan bu araştırmayı daha da anlamlı kılmaktadır. Toptaş (2020) ortaokul öğrencilerinin yazma performanslarında puanlayıcı yanlılığını ve güvenirliğini çalışmıştır. Ancak bu araştırma kapsamında yer alan ortaokul öğrencilerinin tanılanmamış oldukları, dezavantajlı gruplarda da yer almadıkları anlaşılmaktadır.

Bu araştırmada sınıf öğretmenlerinin çeşitli dezavantajlara sahip dördüncü sınıf öğrencilerine karşı okuduklarını anlamalarını değerlendirme sürecindeki yanlılıklarını tartışmak amaçlanmıştır. Dezavantajlı öğrencilere de destek eğitimi sunan İYEP (ilkokullarda yetiştirme programı) kapsamında yer almış/alan öğrencilerin dördüncü sınıflarda yer alması, bu sınıf seviyesine kadar RAM'a yönlendirilme durumlarının olup olmamasının netlik kazanmış olması, bu sınıf seviyesinin seçilmesini gerektirmiştir.

Bu amaçlar kapsamında aşağıdaki sorulara yanıt aranmıştır:

1. Dezavantajlı olan ve dezavantajlı olmayan öğrencilerin okuduğunu anlama düzeyleri farklılık göstermekte midir?
2. Puanlayıcılar dezavantajlı olan ve dezavantajlı olmayan öğrencilerin okuduğunu anlama becerilerini değerlendirmede puanlayıcı yanlılığı göstermekte midir?
3. Dezavantajlı ve dezavantajlı olmayan öğrenciler için okuduğunu anlama rubriğinin kriterleri farklılık göstermekte midir?

Yöntem

Bu araştırmada nicel araştırma yöntemlerinden ilişkisel tarama deseni kullanılmıştır. İlişkisel tarama modelinde amaç, iki veya daha fazla değişken arasındaki ilişkinin varlığının ve derecesinin (Karasar, 2009) herhangi bir müdahalede bulunulmadan (Büyüköztürk, Kılıç-Çakmak, Akgün, Karadeniz ve Demirel, 2018) incelemektir. Araştırmada dezavantajlı (13 kişi) ve dezavantajlı olmayan (35 kişi) 48 4.sınıf öğrencisinin okuduğunu anlama becerileri değerlendirilmiştir. Öğrencilerin performanslarını sınıf öğretmeni olan üç puanlayıcı değerlendirmiştir. Öğrencilerin okuduğunu anlama becerileri için MEB tarafından hazırlanan 'Oyun ve Arkadaşlık' isimli metin verilmiştir. Öğrencilerin okuduğunu anlama becerilerinin ölçülmesinde ise Baştuğ ve diğ. (2019) tarafından geliştirilen okuduğunu anlama değerlendirme dereceli puanlama anahtarı kullanılmıştır. İlgili ölçme aracı yedi kriter ve dörtlü derecelendirmeye sahiptir.

Veri analizinde çalışmanın doğasına uygun olan yöntemlerden biri olan çok yüzeyi Rasch ölçme modeli kullanılmıştır (Linacre, 2012). Araştırmada puanlayıcılar, öğrenciler, öğrenci türü (dezavantajlı ve dezavantajlı olmayan), kriterler olmak üzere dört yüzey bulunmaktadır. Araştırmada tüm yüzeylerin birbiri ile etkileşimde olduğu tamamen çaprazlanmış desen kullanılmıştır.

Veri analizinde kullanılan çok yüzeyli Rasch ölçme modelinden elde edilen ölçümlerin tutarlı ve geçerliği için karşılanması gereken varsayımlar test edilmelidir. İlk olarak, tek boyutluluk için faktör analizi yapılmıştır. Yapılan analiz sonucunda yapının tek faktörlü bir yapıda olduğu ve toplam varyansın yaklaşık %45'ini açıkladığı bulunmuştur (ilgili veri için kriterlerin faktör yükleri sırasıyla; .583; .746; .758; .653; .672; .588 ve .677). Yapının tek faktörlü olması yerel bağımsızlığın sağlandığına işaret ettiğinden yerel bağımsızlığın da sağlandığı varsayılmıştır. Son olarak model veri uyumu için standartlaştırılmış artık değerler incelenmiştir. Model veri uyumunun sağlanması için ±2 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısı toplam gözlem sayısının %5'inden fazla olmaması ve ±3 aralığının dışında kalan standartlaştırılmış artık değerlerin de toplam veri sayısının %1'inden fazla olmaması gerektiği belirtilmiştir (Linacre, 2017). Standartlaştırılmış artık değerler incelendiğinde, ±2 aralığında 44 (%4.37) ve ±3 aralığında ise 3 (%0.30) değerin olduğu bulunmuş ve model veri uyumunun kabul edilebilir düzeyde olduğu sonucuna ulaşılmıştır (toplam gözlem sayısı 3x7x48 = 1 008).

Sonuçlar

Araştırmada dezavantajlı gruptaki öğrencilerin okuduğunu anlama becerilerilerinin değerlendirilmesi sürecinde puanlayıcı yanlılığının incelenmesi amacıyla gerçekleştirilmiştir. Bu amaç doğrultusunda yanıt aranan araştırma sorularına ilişkin bulgular sırasıyla sunulmuştur. İlk olarak dezavantajlı ve dezavantajlı olmayan öğrencilerin okuduğunu anlama becerilerinin yeterlik düzeylerinin farklılık durumu incelenmiş ve elde edilen bulgulara göre istatistiksel olarak farklılığın olduğu bulunmuştur (χ^2 (sd) = 13.8 (1); p< .05). Logit değerleri incelendiğinde, dezavantajlı öğrenciler -0.21 iken dezavantajlı olmayan öğrenciler 0.21 olduğu bulunmuştur. Ayrıca ayırma oranının, ayırma indeksinin ve ayırma indeksi güvenirliğinin yüksek olduğu ve sonuç olarak her iki grubun farklı yeterlik düzeyine sahip olduğu tespit edilmiştir.

Araştırmanın ikinci sorusu olarak puanlayıcı olan öğretmenler dezavantajlı ve dezavantajlı olmayan öğrencilerin okuduğunu anlama becerilerini değerlendirme sürecinde istatistiksel olarak farklı puanlama davranışlarını gösterme durumları incelenmiştir. Bu amaç doğrultusunda öğrenci türü x puanlayıcı etkileşimleri incelenmiştir. Yapılan analizler neticesinde puanlayıcıların puanlama sürecinde yanlı puanlamalar yaptığı belirlenmiştir (χ^2 (sd) = 20.0 (6); p< .05). Puanlayıcılar dezavantajlı öğrencilere cömert davranır iken dezavantajlı olmayan öğrencilere katı davranış göstermiştir.

Son olarak her iki öğrenci grubunun dereceli puanlama anahtarındaki kriterlere göre farklılık gösterme durumları incelenmiştir. Bu amaç doğrultusunda öğrenci türü x kriter etkileşimleri incelenmiştir. Yapılan analizler neticesinde dereceli puanlama anahtarındaki ölçütlerin her iki öğrenci grubu için de benzer bir şekilde çalıştığı ve istatistiksel olarak anlamlı olmadığı bulunmuştur (χ^2 (sd) =

7.3 (14); $p > .05$). Başka bir ifade ile dereceli puanlama anahtarındaki ölçütlerde farklılaşan madde fonksiyonu bulunmamaktadır.

Sonuç olarak, dezavantajlı öğrencilerin okuduğunu anlama becerilerinin değerlendirildiği bu araştırmada, puanlayıcıların dezavantajlı öğrencilere daha fazla cömert davrandıkları fakat yine de bu grubun yeterliğinin daha düşük olduğu tespit edilmiştir. Ayrıca dereceli puanlama anahtarındaki ölçütlerin tüm öğrenciler için aynı şekilde ölçüm yaptığı diğer bir ifade ile yanlılık oluşturmadığı sonucuna ulaşılmıştır. Puanlayıcıların dezavantajlı öğrencilere daha cömert puan verdiklerinin nedenleri ve bu davranışın avantaj ve dezavantajları nitel araştırmalarla derinlemesine bir şekilde ortaya çıkartılabilir.

Kaynaklar

Akyol, H. (2019). *Türkçe ilk okuma yazma öğretimi.* Pegem Akademi.

Baştuğ, M., Hiğde, A., Çam, E., Örs, E. ve Efe, P. (2019). *Okuduğunu anlama becerilerini geliştirme stratejiler, teknikler, uygulamalar.* Pegem Akademi.

Bergh, V. L., Denessen, E., Hornstra, L., Voeten, M., and Holland, R. (2010). The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic achievement gap. *American Educational Research Journal*, *47*(2), 497-527. https://doi.org/10.3102/0002831209353594.

Büyüköztürk, Ş., Kılıç-Çakmak, E., Akgün, Ö., Karadeniz, Ş., ve Demirel, F. (2008). *Bilimsel araştırma yöntemleri.* Pegem Akademi.

Choi, S. K., and Young-Min, P. (2011). The rater characteristics and raters bias in korean language teacher's persuasive writing assessment. *Journal of Curriculum Evaluation, 14*(1), 201-228. https://doi.org/10.29221/jce.2011.14.1.201

Dovidio, J. F., Kawakami, K., and Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82*(1), 62-68. https://doi.org/10.1037//0022-3514.82.1.62

Farrokhi, F., Esfandiari, R., and Vaez Dalili, M. (2011). Applying the many-facet Rasch model to detect centrality in self-assessment, peer-assessment and teacher assessment. *World Applied Sciences Journal, 15*(11), 70-77. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.390.2450&rep=rep1&type=pdf

Grabe, W., and Stoller, F. (2002). *Teaching and researching reading.* Pearson Education Limited.

Hans, A., and Hans, E. (2015). Different comprehension strategies to improve student's reading comprehension. *International Journal of English Language Teaching, 3*(6), 61-69. https://www.eajournals.org/wp-content/uploads/Different-Comprehension-Strategies-to-Improve-Student---s-Reading-Comprehension.pdf

Karasar, N. (2012). *Bilimsel araştırma yöntemi.* Nobel Akademi Yayıncılık.

Klingner, J. K., Vaughn, S., and Boardman, A. (2007). *Teaching reading comprehension to students with learning difficulties.* The Guilford Press.

Linacre, J. M. (2012). *FACETS* (version 3.70.1) [Computer Software]. MESA Press.

Linacre, J. M. (2017). *FACETS* (version 3.80.0) [Computer Software]. MESA Press.

Milanowski, A. (2017). Lower performance evaluation practice ratings for teachers of disadvantaged students: Bias or reflection of reality? *AERA Open, 3*(1), 1-16. https://doi.org/10.1177/2332858416685550.

Taşdelen, H., Turhan, M., Erikci, M. ve Özkan, S. (2014). *Okullardaki dezavantajlı ve risk altındaki çocuklar*. MEB.

Toptaş, C. (2020). *Performansın değerlendirilmesinde farklılaşan puanlayıcı davranışlarının çok yüzeyli Rasch ölçme modeli ile incelenmesi* (Tez No: 629923) [Yüksek lisans tezi, Gazi Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Wesolowski, B. C., Wind, S. A., and Engelhard Jr, G. (2015). Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning. *Musicae Scientiae, 19*(2), 147-170. https://doi.org/10.1177/1029864915589014

Doğrulayıcı faktör analizi sonuçlarının farklı istatistik programlarına göre karşılaştırılması

Çiğdem Reyhanlıoğlu, Mehmet Taha Eser ve Gökhan Aksu

Anahtar kelimeler: Doğrulayıcı faktör analizi, Mplus, LISREL, JAMOVI, JASP

Giriş

Doğrulayıcı faktör analizi (DFA) eğitim bilimleri, sağlık bilimleri, doğa bilimleri ve beşeri bilimler gibi birçok alanda özellikle teorik model ile teorik olarak değişkenler arasında var olduğu kanıtlanmış ya da öngörülen doğrusal ve doğrusal olmayan ilişkilerin betimlendiği modellerin sınanmasında kullanılan çok değişkenli bir istatistiksel tekniktir (Aksu ve diğ., 2017; Leech ve diğ., 2005). Bir başka ifade ile DFA araştırmacı tarafından teorik olarak ortaya konulan değişkenler arasında belirlenmiş olan ilişki örüntülerinin gözlenen değişkenler tarafından doğrulanma/ doğrulanmama durumunu ortaya koymak amacıyla sıklıkla kullanılan bir tekniktir.

DFA çoklu regresyon analizini, faktör analizini ve varyans analizini bir arada kullanan karmaşık bir tekniktir. DFA'nın bu karmaşık yapısından dolayı DFA uygulamalarının gerçekleştirilmesi için bilgisayar teknolojisinin kullanılması zorunlu olmuştur (Raykov ve Marcoulides, 2006). DFA için kullanılan istatistik programlarının tarihsel süreçte gelişimine bakılacak olursa DFA uygulaması yapabilen ilk programın 1970'li yıllarda Jöreskog tarafından geliştirilen LISREL (Lineer Structural Relations) bilgisayar programı olduğu görülmektedir (Wikipedia, 2021). Sonraki yıllarda teknolojinin hızla gelişmesi ile birlikte farklı istatistik programlar da geliştirilmiştir. Günümüzde DFA uygulamaları için LISREL, Mplus, AMOS, EQS, JAMOVI, JASP gibi istatistik programlarından yararlanılmaktadır.

DFA uygulamasını gerçekleştiren istatistik programları genel olarak benzer özelliklere sahip olsalar da ürettikleri sonuçlar açısından farklılaşmaktadır (Davey ve Savla, 2010; Reisinger ve Turner, 1999).

Programlar arasındaki bu farklılıkların, programların birbirine göre üstün ve zayıf yönlerinin belirlenmesi ve sonuçlarının ortaya konması birçok alternatif arasından hangisinin tercih edileceği konusunda araştırmacılara yol göstermesi açısından önemlidir. Bu bağlamda DFA uygulamalarında alanyazın ve saha çalışmalarında sıklıkla kullanılan istatistik programlarından LISREL, Mplus, JAMOVI ve JASP programlarına ait DFA uygulama çıktılarının birbirine göre nasıl farklılaştığını temel alan bu çalışmanın önemli olduğu değerlendirilmektedir. Bu nedenle bu çalışmada LISREL, Mplus, JAMOVI

ve JASP programlarından elde edilen sonuçların birbirine göre üstün ve zayıf yönlerinin belirlenmesi amaçlanmıştır. Bu amaç doğrultusunda araştırma kapsamında cevap aranan sorular şu şekilde ifade edilebilir:

1. İstatistiksel programlar kestirilen model uyum indeksi türlerine göre farklılık göstermekte midir?
2. İstatistiksel programlara göre elde edilen faktör yükleri birbirinden farklı mıdır?
3. İstatistiksel programlara göre elde edilen modifikasyon indeksleri nasıldır?

Yöntem

Bu araştırmada farklı istatistiksel programlardan elde edilen DFA sonuçlarının karşılaştırılması ve karşılaştırma sonuçlarına göre programların güçlü ve zayıf yönlerinin ortaya koyulması amaçlanmıştır. Bu nedenle bu çalışma betimsel bir çalışmadır.

Araştırma sonuçları PISA 2018 uygulamasına katılan ve Türkiye örneklemini temsil eden 6890 öğrenciye ait cevaplar üzerinden elde edilmiştir. PISA 2018 Türkiye örnekleminde öğrencilerin küresel düşüncelilik duygusunu (GLOBMIND) değerlendiren ve ST219 kodlu boyutta yer alan 6 maddeye verilen öğrenci cevaplarından yararlanılmıştır. Bu maddelerin ölçtüğü yapılar "Dünya vatandaşlığı duygusu" (madde Q01), "dünyada başkalarına karşı sorumlu olma" (madde Q02, Q04 ve Q06), "Birbiriyle bağlantılı olma duygusu" (madde Q03) ve "Küresel öz yeterlilik" (madde Q05) şeklinde ifade edilebilir. Madde formatı "Kesinlikle katılmıyorum", "Katılıyorum", "Katılıyorum", "Kesinlikle katılıyorum" yanıt kategorilerinden oluşan dörtlü likert tipindedir (OECD, 2020).

Çalışmanın amacı doğrultusunda DFA analizleri için LISREL 8.80, Mplus 7, JAMOVI 1.6.23 ve JASP 0.14.1.0 istatistik programlarından yararlanılmıştır. Analize başlamadan önce DFA uygulamasına ilişkin varsayımların test edilmesine ilişkin sonuçlar şu şekilde özetlenebilir:

Örneklem büyüklüğü: DFA için uygun örneklem büyüklüğüne dair tam bir görüş birliğine varılmasa da DFA uygulamasında 1000 kişinin üzerindeki örneklemler "mükemmel" olarak ifade değerlendirilmektedir (Comrey ve Lee 1992). Çalışmanın bulguları 6890 kişiye ait veriler üzerinden elde edildiği için ilgili varsayım sağlanmıştır.

Çok Değişkenli Normal Dağılım: DFA için bu varsayımın test edilmesinde Kalaycı'nın (2014) önerisi doğrultusunda standartlaştırılmış hata değerlerine ilişkin grafiklerle birlikte çoklu çarpıklık ve basıklık istatistiklerine dayalı olan analitik testlerden yararlanılmıştır. Elde edilen sonuçlara göre ilgili varsayımın karşılandığı görülmüştür.

Varyans-kovaryans matrislerinin homojenliği: Bu varsayımın test edilmesinde kullanılan ve Box (1949) tarafından önerilen Box-M istatistiğine göre bu varsayımın karşılandığı görülmüştür.

Hataların bağımsızlığı ve otokorelasyon: Bu varsayım Durbin Watson (D-W) istatistiğinin kullanılmasıyla test edilmektedir (Nerlove ve Wallis, 1966). D-W istatistiğinin sonucuna göre herhangi bir değişkene

ait hata ile diğer değişkenlerin hataları arasında anlamlı bir ilişkinin olmadığı ve dolayısıyla ilgili varsayımın karşılandığı tespit edilmiştir.

Çoklu doğrusal bağlantı (Multicolinearity): Değişkenler arasında, işaretine bakılmaksızın 0.80'dan büyük ilişkinin olması durumunda çoklu bağlantı problemi ortaya çıkmaktadır (Aksu ve diğ., 2017). Yapılan korelasyon analizi sonucunda değişkenler arasındaki korelasyonların düşük olduğu görülmüştür.

Sonuçlar

Birinci araştırma probleminin çözümüne ilişkin bulgular Tablo 1'de görülmektedir.

Tablo 1

LISREL, Mplus, JAMOVI ve JASP'tan Elde Edilen Uyum İndeksi Çeşitleri

	LISREL	MPLUS	JAMOVI	JASP
χ^2	√	√	√	√
RMSEA	√	√	√	√
ECVI	√			√
AIC	√	√		
BIC		√	√	√
SSABIC		√		√
CAIC	√			
NFI	√			√
NNFI	√			√
PNFI	√			√
CFI	√	√	√	√
IFI	√			√
RFI	√			√
RMR	√			
GFI	√			√
PGFI	√			
SRMR		√	√	√
TLI		√	√	√
RNI				√
MFI				√

DFA'nın uygulandığı dört farklı istatistiksel program arasında en fazla uyum indeksi türünün JASP'ta hesaplanabildiği görülmektedir. JASP'ın ardından LISREL ve LISREL'in ardından Mplus gelmektedir. En az uyum indeksi çeşidinin ise JAMOVI'de hesaplanabildiği tabloda görülmektedir. Tek bir uyum indeksi sonucuna göre teorik model ile kurulan modelin birbiri ile uyum gösterdiği kararına varmak doğru değildir (Tabachnick ve Fidell, 2001; Kline, 2005). Bu bağlamda kullanılan paket programın kestirdiği uyum indeksi çeşitliliğinin fazla olması verilen kararın geçerliliğini arttıracaktır. Elde edilen sonuçlara göre uyum indeksi çeşitliliği bakımından en güçlü programın LISREL, en zayıf programın JAMOVI olduğu sonucuna ulaşılmıştır. Ayrıca programlarda ortak olarak kestirilen uyum indeksleri büyüklük olarak karşılaştırıldığında programlar arasında bazı farklılıkların olduğu görülmüştür.

İkinci araştırma probleminin çözümüne ilişkin bulgular incelendiğinde JAMOVI, JASP ve LISREL programlarının birbirine çok yakın büyüklükte standart hatalarla kestirilen faktör yüklerinin hemen hemen aynı olduğu görülmüştür. Bununla birlikte Mplus programı JASP ve JAMOVI'den farklı büyüklükte faktör yükleri üretmiştir.

Son olarak üçüncü araştırma probleminin çözümüne ilişkin bulgular incelendiğinde JASP, Mplus ve LISREL'den üretilen modifikasyon indeksleri analiz çıktısında düz bir tablo içerisinde verilmektedir. Bu durum sonuçların okunmasını zorlaştırmaktadır. JAMOVI'de ise modifikasyon indekslerinin matris içinde verilmesi sonuçları daha kolay anlaşılır hale getirmektedir. Üç programdan elde edilen sonuçlar incelendiğinde kestirilen modifikasyon indekslerinin birbirine çok yakın değerler aldığı görülmüştür.

Kaynaklar

Aksu, G., Eser, M. T. ve Güzeller, C. O. (2017). *Açımlayıcı ve doğrulayıcı faktör analizi ile yapısal eşitlik uygulamaları.* Detay Yayıncılık.

Bentler, P.M., and Yuan, K.H. (1999). Structural equation modeling with small samples: *Test statistics. Multivariate Behavioral Research, 34*(2), 181-197. https://doi.org/10.1207/S15327906Mb340203

Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika,* 36, 317–346.

Comrey, A. L., and Lee, H. B. (1992). *A first Course in factor analysis* (2nd ed.). Lawrence Erlbaum Associates, Inc

Davey, A., and Savla, J. (2010). *Statistical power analysis with missing data: A structural equation modeling approach.* Routledge.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). The Guilford Press.

Leech, N.L., Barrett, K.C., and Morgan, G.A. (2005). *SPSS for Intermediate Statistics; use and interpretation.* Lawrence Erlbaum Associates.

LISREL (2021, August 13). In *wikipedia the free ensiclopedia*. https://en.wikipedia.org/wiki/LISREL

Nerlove, M., and Wallis, K. (1966). Use of the Durbin-Watson statistic in inappropriate situations. *Econometrica, 34*(1), 235-238. https://doi.org/10.2307/1909870

OECD (2020). *PISA 2018 Technical Report.* OECD Publishing. https://www.oecd.org/pisa/data/pisa2018technicalreport/PISA2018_Technical-Report-Chapter-16-Background-Questionnaires.pdf

Raykov, T., and Marcoulides, G.A. (2006). *A first course in structural equation modeling.* Lawrence Erlbaum.

Reisinger, Y., and Mavondo, F. (2007). Structural equation modeling. *Journal of Travel & Tourism Marketing*, *21*(4), 41-71. https://doi.org/10.1300/J073v21n04_05

Tabachnick, B. G., and Fidell, L. S. (2001). *Using Multivariate Statistics* (4th ed.). Allyn and Bacon.

Likert ölçeklerde kategoriler analiz aşamasında birleştirilebilir mi? Geçerlik ve güvenirlik üzerine etkileri

Abdullah Faruk Kılıç ve İbrahim Uysal

Anahtar kelimeler: Kategori birleştirme, geçerlik, güvenirlik

Giriş

Eğitim bilimlerindeki araştırmalarda kullanılan veri toplama araçlarının büyük bir bölümü Likert tipindedir. Bu tür ölçeklerde bir takım durum katılımcılara sunulur ve bunlara belirli kategorileri (katılmıyorum [1], kararsızım [2], katılıyorum [3] gibi) işaretleyerek cevap vermeleri istenir. Likert tipi ölçeklerde 5 kategori anlaşılır, karşılaştırılabilir ve kullanışlı olması nedeniyle tercih edilirken, daha az ya da çok sayıda kategori kullanılabilir (Willits ve diğ., 2016). Likert tipi ölçeklerde çoğunlukla aynı sayıda kategoriye sahip maddeler kullanılmakla beraber bazı durumlarda farklı kategorilerde maddeler yer alabilmektedir. Chakrabartty (2020) davranış bilimlerinde cevap kategori sayısı farklılaşan Likert tipi ölçeklerin de sıklıkla kullanıldığını belirtmektedir. Araştırmacılar bu tür durumlarda faktör analizi uygulamak üzere veri analizi sürecinde kategorileri birleştirme yoluna gidebilmektedir. Bundan daha önemlisi farklı kategorilere sahip maddelerle veri toplayan araştırmacılar karşılaştırılabilir sonuçlar elde etmek için maddeleri tek bir kategori sayısına indirgemek isteyebilmektedir (Chakrabartty, 2020). Bazı ölçeklerin farklı kültürlere uyarlanma aşamasında kategori sayısının değiştirilebildiği dikkate alınırsa kültürlerarası karşılaştırma amacıyla yapılan çalışmalarda araştırmacılar veri analizi aşamasında kategori birleştirebilmektedir. Özellikle çok kategorili ölçeklerde bazı kategoriler işaretlenmediğinde araştırmacılar benzer şekilde kategori birleştirme eğilimi gösterebilmektedir.

Likert tipi ölçeklerde kategori sayısı yükseldikçe varyans ve güvenirlik artma eğiliminde olabilmekte dolayısıyla dağılımın şekli değişebilmektedir (Cook ve diğ., 2001). Kategori sayısının bu tür etkileri yanında, kategori birleştirmesi bilgi kaybı yaratabilmekte, duyarlılık (sensitivity) azalabilmekte ancak yukarıda sayılan nedenlerle bu işlemin yapılması gerekebilmektedir. Tüm bu adımların atılabilmesi için verilerin geçerlik ve güvenirliğinin düşmemesi gerekmektedir. Bu nedenledir ki veri analizi sürecinde kategori birleştirmenin güvenirlik katsayıları ve yapı geçerliği üzerindeki etkisi dikkatle incelenmelidir.

Chakrabartty (2020) 3, 4, 5 ve 7 kategorili her bir kategoride 5 maddenin yer aldığı ölçekte kategorileri üç farklı teknikle dönüştürerek birleştirmiştir. Araştırma sonucunda ise sosyal bilimler

alanındaki araştırmacılara kategori birleştirmesini önermiştir. Ancak bu araştırmada kategori birleştirmesinin faktör analizi sonuçları üzerindeki etkisi değerlendirilmediği gibi koşullar üzerinde değişiklik yapılmamıştır. Literatürde konuyla ilgili diğer araştırmalarda (örn. Schwarz ve diğ., 1991) ise veri analizi sürecinde kategori birleştirilmesine odaklanılmamıştır. Buradan hareketle gerçekleştirilen mevcut araştırmada veri analizi sürecinde kategori birleştirmesinin örneklem büyüklüğü, veri dağılımı, madde sayısı ve ortalama faktör yükü koşulları altında ortaya çıkardığı etki değerlendirilmiştir. Araştırmada cevabı aranan sorular aşağıda yer almaktadır.

1. Örneklem büyüklüğü, madde sayısı, ortalama faktör yükü ve veri dağılımı koşulları altında kategori birleştirmesinin Cronbach alfa ve McDonald omega güvenirlik katsayıları üzerindeki etkisi nedir?
2. Örneklem büyüklüğü, madde sayısı, ortalama faktör yükü ve veri dağılımı koşulları altında kategori birleştirmesinin ortalama faktör yüklerinin göreli yanlılığı ve doğru tahmin yüzdesi üzerindeki etkisi nedir?

Yöntem

Monte Carlo simülasyonu olarak tasarlanan bu çalışmada Likert tipi ölçeklerde kategori birleştirilmesinin geçerlik ve güvenirlik üzerindeki etkisinin belirlenmesi amaçlanmıştır. Araştırmanın simülasyon koşulları; kategori dönüşümü (5 kategoriden 4 kategoriye, 5 kategoriden 3 kategoriye ve 4 kategoriden 3 kategoriye), örneklem büyüklüğü (300, 500 ve 1000), madde sayısı (10 ve 15 madde), dağılım şekli (normal, sağa ve sola çarpık) ve ortalama faktör yüküdür (.40 ve .70) olarak belirlenmiştir. Araştırmada sabit simülasyon koşulu ölçme modelidir (tek boyutlu). Örneklem büyüklüğü 3 koşul, dağılım şekli 3 koşul, madde sayısı 2 koşul, ortalama faktör yükü 2 koşul ve kategori dönüşümü 3 koşul olmak üzere (3x3x2x2x3= 108) tamamen çaprazlanmış desen altında 108 simülasyon koşulunda çalışılmış olup her bir koşul için 1000 replikasyon yapılmıştır. Çalışmada sadece tek boyutlu yapılar incelenmiştir.

Veri üretimi için R yazılımında (R Core Team, 2020) bulunan lavaan (Rosseel, 2012) paketi kullanılmıştır. Veri setleri öncelikle sürekli ve çok değişkenli normal dağılım gösterecek şekilde üretilmiş, kategorik hale getirilirken belirlenen kesme noktalarından (threshold) yararlanılmıştır. Açımlayıcı faktör analizi (AFA) psych (Revelle, 2020) paketiyle gerçekleştirilmiştir. Açımlayıcı faktör analizinde faktör çıkarma yöntemi olarak temel eksenler (principal axis) yöntemi kullanılmıştır. AFA, polikorik korelasyon matrisi ile gerçekleştirilmiştir. Simülasyon sonuçları değerlendirilirken güvenirlik katsayılarından ve faktör analizi sonucunda elde edilen ortalama faktör yüklerinin göreli yanlılık değerlerinden yararlanılmıştır. Göreli yanlılık; ([ortalama faktör yükü]-[gerçek faktör yükü])/(gerçek faktör yükü) şeklinde hesaplanmıştır. Göreli yanlılık değerinin |GY|<.10 olması kabul edilebilir olarak değerlendirilmiştir (Flora ve Curran, 2004; Moshagen ve Musch, 2014). Ayrıca ortalama faktör yüklerinin ±%5'lik aralıkları incelenerek replikasyonların yüzde kaçında bu aralıkta ortalama faktör yükü

elde edildiği incelenmiş ve buna doğru tahmin yüzdesi denilmiştir. Güvenirlik katsayılarından McDonald'ın omega (McDonald, 1999) ve Cronbach'ın alfa katsayıları incelenmiştir.

Sonuçlar

Göreli yanlılık açısından incelendiğinde normal dağılım gösteren veri setlerinde dönüşüm yapılmış ve dönüşüm yapılmamış veri setleri arasında fark olmadığı gözlenmiştir. Sağa çarpık dağılan veri setlerinde dönüştürme yapıldığında da tüm koşullarda kabul edilebilir aralıkta yanlı sonuçlar elde edilmiştir. Ancak sola çarpık veri setlerinde örneklem büyüklüğünün 300, ortalama faktör yükünün .40 olduğu koşulda 4 kategoriden 3 kategoriye dönüşüm yapılması negatif yanlı sonuçlar ortaya çıkarmıştır. Aynı koşullar için örneklem büyüklüğü 500'e çıktığında yanlılık bir miktar azalmış ancak yine de kabul edilebilir aralıkta yer almamıştır.

Doğru tahmin yüzdesi (DTY) açısından incelendiğinde örneklem ve ortalama faktör yükünün artması DTY'yi yükseltmiştir. Dönüşüm yapılması, normal dağılım gösteren veri setlerinde DTY değerini önemli düzeyde değiştirmezken sola çarpık dağılım gösteren veri setlerinde önemli değişikliğe neden olmuştur. Özellikle bu durum 4 kategoriden 3 kategoriye yapılan dönüşümde gözlenmiştir.

Güvenirlik katsayıları açısından incelendiğinde ise McDonald omega katsayısının kategori dönüşümünden çok etkilenmediği ancak Cronbach alfanın özellikle 4 kategoriden 3 kategoriye gerçekleştirilen dönüşüm sonucunda düştüğü gözlenmiştir.

Kaynaklar

Chakrabartty, S. N. (2020). *Combining Likert items with different number of response categories. Proceedings on Engineering Sciences, 2*(3), 311-322. https://doi.org/10.24874/PES02.03.010

Cook, C., Heath, F., Thompson, R. L., and Thompson, B. (2001). Score reliability in webor internet-based surveys: Unnumbered graphic rating scales versus likert-type scales. *Educational and Psychological Measurement, 61*(4), 697–706. https://doi.org/10.1177/00131640121971356

Flora, D. B., and Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466–491. https://doi.org/10.1037/1082-989X.9.4.466

McDonald, R. P. (1999). *Test theory: A unified treatment.* Erlbaum.

Moshagen, M., and Musch, J. (2014). Sample size requirements of the robust weighted least squares estimator. *Methodology, 10*(2), 60–70. https://doi.org/10.1027/1614-2241/a000068

R Core Team (2020). *R: A language and environment for statistical computing* (version 3.0.1) [Computer Software]. http://www.R-project.org/

Revelle, W. (2020). *psych: Procedures for psychological, psychometric, and personality research* (version 2.0.12) [Computer software]. https://cran.r-project.org/package=psych

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Schwarz, N., Knäuper, B., Hans-J. Hippler, Noelle-Neumann, E., and Clark, L. (1991). Rating scales: Numeric values may change the meaning of scale labels. *The Public Opinion Quarterly, 55*(4), 570-582. http://www.jstor.org/stable/2749407

Willits, F. K., Theodori, G. L., and Luloff, A. E. (2016). Another look at Likert scales. *Journal of Rural Social Sciences, 31*(3), 126-139. https://egrove.olemiss.edu/jrss/vol31/iss3/6

Eğitim bilimleri alanında yapılan meta-analiz çalışmalarında heterojenlik yaklaşımlarının incelenmesi

Mahmut Sami Koyuncu, Ayşenur Erdemir ve Esra Oyar

Anahtar kelimeler: Doküman analizi, eğitim bilimleri, heterojenlik, meta-analiz

Giriş

Meta-analiz, farklı çalışmalardan elde edilen birleştirilmiş verileri analiz etmek için istatistiksel bir süreci ifade etmektedir. Bu nedenle de kısa ve öz güncel bilgilerin ana kaynağı olmaktadır. Bununla birlikte, bir meta-analizin genel sonuçları, büyük ölçüde meta-analitik sürecin kalitesine bağlıdır ve meta-analizin (meta-değerlendirme) kalitesinin uygun bir şekilde değerlendirilmesi zor olabilmektedir (Nakagawa ve diğ., 2017). Örneğin, meta analiz kapsamında incelenen çalışmalar için heterojenlik varsayımının test edilmesi gerekmektedir (Şen, 2019). Meta analiz sonucunda rapor edilen genel etki, etki büyüklükleri arasındaki heterojenlik veya tutarsızlık analizi yapılmadan uygun şekilde yorumlanamaz. Bu doğrultuda, yapılan çalışmalarda meta analizin heterojenliğin test edilip edilmediğinin incelenmesi gerekliliği oluşmaktadır (Nakagawa, ve diğ. 2017).

Alan yazın incelendiğinde yapılan meta-analiz çalışmalarında, heterojenlik testlerinin ihmal edilebildiği, heteorjenlik testi yapılsa bile raporlanmadığı, doğrudan heterojenlik testi sonucuna göre model seçilebildiği, rastgele veya sabit etki modelinin seçim mantığının genellikle belli olmadığı, heterojenlik yaklaşımlarının açıklanmadığı belirtilmektedir (Petitti, 2001; Xu ve diğ., 2008). Xu ve diğ. (2008) Ocak 2005-Nisan 2007 tarihleri arasında doğum sağlığı konularına odaklanan meta-analizlerde heterojenliğin araştırılmasındaki yaklaşımları değerlendiren bir araştırma yapmıştır. Petitti (2001) 1999-2001 yılları arasında yayımlanmış 16 tıp dergisinde üreme sağlığı konusundaki 34 meta-analiz çalışmasındaki heterojenlik yaklaşımları değerlendiren bir araştırma yapmıştır. Eğitim alanlarında yapılan meta analiz çalışmalarında heterojenlik yaklaşımlarının incelendiği bir çalışmaya rastlanamamıştır.

Türkiye DergiPark sisteminde Eğitim bilimleri konu alanında meta-analiz çalışması olarak yayınlanmış makalelerde heterojenlik yaklaşımlarının nasıl ele alındığının incelendiği bu çalışmada aşağıdaki sorulara yanıt aranmıştır: İncelenen makalelerde

1. İstatistiksel heterojenlik testlerinin raporlanma durumu nasıldır?

2. İstatistiksel heterojenlik testlerini değerlendirmek için kullanılan anlamlılık düzeyi kriteri nedir?
3. İstatistiksel heterojenlik testleri yapan meta-analizlerdeki rasgele/sabit etkiler modelleri kullanım durumu nasıldır?
4. Heterojenliğin olası sebeplerini inceleme durumu nasıldır?

Bu çalışma ile ele alınan makalelerde, istatistiksel heterojenlik testlerinin yapılıp yapılmadığının, sonuçların raporlanıp raporlanmadığının, istatistiksel heterojenlik testi için anlamlı bir bulgunun nasıl ele alındığının, sonuçların nasıl yorumlandığının incelenmesi amaçlanmıştır.

Bu çalışma yapılacak olan meta-analiz çalışmalarında heterojenlik testinin raporlanması, sonuçlarının değerlendirilmesi konusunda araştırmacılara yol gösterecektir. Böylece yapılacak olan meta-analiz çalışmalarında genellikle yapılan hatalara vurgu yapılarak, bu hataların en aza indirgenmesine katkı sunmasının önemli olduğu düşünülmektedir. Çalışma sonucunda meta-analizde istatistiksel heterojenliği değerlendirmek, rapor etmek ve araştırmak için sunulan öneriler sayesinde yapılacak meta-analiz çalışmalarının kalitesinin artmasına katkı sunmak hedeflenmektedir.

Yöntem

Bu araştırma, Türkiye DergiPark sisteminde Eğitim bilimleri konu alanında meta-analiz çalışma olarak yayınlanmış makalelerde heterojenlik yaklaşımlarının nasıl ele alındığının incelenmesi amaçlandığı için, nitel bir araştırma olup, nitel araştırma türlerinden durum çalışmasıdır. Veri toplama yöntemi olarak doküman analizi kullanılmıştır.

Araştırmanın evrenini DergiPark sisteminde yer alan tüm dergilerde 2013-2021 yılları arasında Eğitim, Eğitim Araştırmaları ve Eğitim, Bilimsel Disiplinler konu alanında yayımlanan 63 makale oluşturmaktadır. Makale seçiminde DergiPark sisteminde yer alan tüm dergilerde "meta-analiz", "meta analiz", "meta-analysis" ve "meta analysis" anahtar kelimesi ile arama yapılmıştır. Ancak çalışmaya sadece Eğitim, Eğitim Araştırmaları ve Eğitim, Bilimsel Disiplinler konu alanındaki çalışmalar dâhil edilmiştir. Bu çalışmalar içinden, meta-analiz çalışması olmayanlar araştırmaya dâhil edilmemiştir.

Araştırma kapsamındaki verilere DergiPark sistemi üzerinden ulaşılmıştır. Veri toplama aracı olarak, araştırmacılar tarafından geliştirilen ve araştırma kapsamındaki makalelerin heterojenlik yaklaşımlarını özetleyen standart bir *"meta-analiz heterojenlik yaklaşımı inceleme formu"* oluşturulmuştur. Bu form iki kısımdan oluşmaktadır. Birinci kısımda makalenin künyesi, ikinci kısımda ise makalenin heterojenlik yaklaşımı ile ilgili bilgileri yer almaktadır. Veri toplama aracının hazırlanmasında ilk olarak ilgili alan yazında daha önce başka araştırmacılar tarafından hazırlanılan formlar incelenmiştir. Daha sonra araştırmacılarla iletişime geçilip dönüt alınmış ve uzman görüşlerine göre heterojenlik yaklaşımı inceleme formunun son hali oluşturulmuştur.

Verilerin Analizinde betimsel istatistikler kullanılmıştır. Araştırma kapsamında incelenen makalelerden, rasgele seçilen 9 makale üzerinden meta-analiz heterojenlik yaklaşımı inceleme formu ile

kodlayıcılar arası puanlama güvenirliği belirlenmiştir. Her bir çalışma tüm araştırmacılar tarafından bağımsız olarak kodlanmış ve kodlayıcılar arası güvenirlik değeri .933 olarak yüksek bir değer elde edilmiştir.

Sonuçlar

Araştırma kapsamında 63 çalışma incelenmiştir. Ancak bu çalışmalardan 7 tanesi meta-analiz çalışması olduğunu belirtmesine rağmen, inceleme sonucunda meta-analiz çalışması olmadığı anlaşılmıştır. Özellikle bu çalışmaların meta-sentez ya da reviev çalışması olduğu görülmüştür. Dolayısıyla araştırma bulguları meta-analiz çalışması olmayan çalışmalar çıkarıldıktan sonra 56 çalışma üzerinden sunulmuştur.

İncelenen 56 meta-analiz çalışmasının %96 (n=54)'sında heterojenlik testinin yapıldığı, sadece %4 (n=2)'ünde heterojenlik testinin yapılmadığı görülmüştür. Hetorejenlik testi yapıldığı belirtilen 54 çalışmadan ise sadece 1 tanesinde heterojenlik testi sonuçların raporlanmadığı belirlenmiştir. Çalışmaların %87'sinde hetorejenliğin belirlenmesinde klasik yöntemlerden X^2, Q istatistiği ve I^2 indeksinin kullanıldığı belirlenmiştir.

İstatistiksel heterojenlik testlerini değerlendirmek için kullanılan anlamlılık düzeyi kriterinin çalışmaların %66 (n=36)'sında 0,05; çalışmaların %5,6 (n=3)'sında ise 0,01 olarak seçildiği görülmüştür. Ancak çalışmaların %27 (n=15)'sinde ise istatistiksel heterojenlik testlerini değerlendirmek için kullanılan anlamlılık düzeyi kriterinin belirtilmediği tespit edilmiştir.

Çalışmaların heterojenligin olası sebeplerini belirleme durumları incelendiğinde, heterojenlikle başa çıkma yaklaşımı olarak çalışmaların %50 (n=27)'sinde doğrudan *heterojenlikten ötürü random etikler modeli kullanıldığı*, %14.8 (n=8)'inde sadece *alt grup analizlerinin* yapıldığı, %11.1 (n=6)'inde ise *homojen bulunduğu için sabit ekiler modeli kullanıldığı* belirlenmiştir. Sadece incelenen çalışmaların %1,9 (n=1)'unda *homojen sonuç bulunup, çalışmalar bağımsız olduğu için randum etkiler modelinin kullanımının* tercih edildiği belirlenmiştir. Ayrıca çalışmalarda *heterojen bulunup ve hiçbir şey yapılmayan* %5.6 (n=3) çalışma olduğu belirlenmiştir.

Kaynaklar

Nakagawa, S., Noble, D. W., Senior, A. M., and Lagisz, M. (2017). Meta-evaluation of meta-analysis: Ten appraisal questions for biologists. *BMC biology*, *15*(1), 1-14. https://doi.org/10.1186/s12915-017-0357-7

Petitti, D. B. (2001). Approaches to heterogeneity in meta-analysis. *Statistics in Medicine, 20*(23), 3625-3633. https://doi.org/10.1002/sim.1091

Şen, S. (2019). SPSS ile meta-analiz nasıl yapılır? *Harran Maarif Dergisi, 4*(1), 21-49. 4(1), 21-49 http://dx.doi.org/10.22596/2019.0401.21.49

Xu, H., Platt, R. W., Luo, Z. C., Wei, S., and Fraser, W. D. (2008). Exploring heterogeneity in meta-analyses: Needs, resources and challenges. *Paediatric and Perinatal Epidemiology, 22*, 18-28. https://doi.org/ 10.1111/j.1365-3016.2007.00908.x

Çevrimiçi sınav uygulaması ve güvenlik tartışması

Selma Tosun ve Mehmet Tosun

Anahtar kelimeler: Çevrimiçi sınav, e-sınav, uzaktan ölçme, sınav güvenliği, bağıl değerlendirme

Giriş

Bilginin dijitalleşmesi kaçınılmaz olarak eğitimin yönünü de dijital alana kaydırmış beraberinde de tüm dünyada ve Türkiye'de uzaktan eğitim ve çevrimiçi ders uygulamaları artmıştır. Özellikle son yıllarda Türkiye'deki üniversitelerin neredeyse tamamında uzaktan eğitim uygulama merkezleri açılmış; ortak dersler başta olmak üzere bazı dersler bu yöntemler ile yürütülmeye başlanmıştır. Nitekim Covid19 pandemi tedbirleri kapsamında 127 kamu üniversitesinin tamamı uzaktan eğitim ile bu süreci yürütecekleri kararını açıklamışlardır. Bununla birlikte literatürde nüanslarla birbirinden ayrılan pek çok ilişkili kavram olsa da bilgisayarlı eğitim/uzaktan öğrenme ve geniş kitlelerin eğitimi için düşünülen açık öğretimle birlikte bu tür bir eğitimin nasıl ölçülüp değerlendirileceği de uzun süredir tartışılmaktadır. En klasik uygulama olan "uzaktan eğitim-yakından klasik sınav" metodu başta olmak üzere bu alandaki ölçme değerlendirme yöntemleri ve buna ek alternatif arayışları devam etmektedir. Bununla birlikte iyi planlanmış bir uzaktan öğrenme programı ile pandemi sürecinde zorunluluk ile ortaya çıkan bazı ölçme değerlendirme uygulamaları farklı ele alınmalıdır. Ülkemizde yükseköğretim uzaktan eğitim programlarında 2010 yılında ara sınavlar için uzaktan çevrimiçi sınav uygulaması yapılmış ve 2016 yılına kadar devam eden bu durum değişen mevzuat ve beraberinde yapılan güvenlik tartışmaları ile birlikte askıya alınmıştır. Gözetimli ya da gözetimsiz çevrimiçi sınavlar pandemi koşullarında tavsiyesi ile bir zorunluluk olarak ortaya çıkınca tekrar tartışılmaya başlanmıştır. Covid19 salgınının eğitim alanında yarattığı tahribat için telafi metodu olarak uzaktan eğitim ön plana çıkmış ancak burada uzaktan ölçme ve tabi değerlendirme ile ilgili kaygılar da artmıştır. Bu kaygılar temel eğitim düzeyinde, daha adil bir ölçmenin nasıl yapılacağından kaynaklanırken, yükseköğretimde artan çevrimiçi sınavlar sınav güvenliği tartışmasını başlatmıştır. Ayrıca bu dönemde yükseköğretimde pek çok farklı yöntemler de kullanılmıştır. Kitlesel eğitiminin yapılmadığı ve nispeten daha az öğrencisi olan programlar ödev ve proje gibi alternatif ölçme yöntemlerini kullanırken bir kısım programlarda çevrimiçi sınav metotları tercih edilmiştir. Uzaktan ölçme değerlendirmede teknik/yöntemler, altyapı uygunluğu, zaman ve güvenlik dengesi, erişilebilirlik ve yönetilebilirlik kavramları ile birlikte değerlendirilmektedir. Aşırı güvenlik tedbirleri ve altyapı kaynaklarının yetersizliğinin öğrenci mağduriyeti yarattığına yönelik bu süreçte pek çok itiraz

gelmiştir. Öte yandan güvenliği ve geçerliği tartışmalı sınav uygulamaları ve buna bağlı bağıl değerlendirme yönteminin kullanımına yönelik de öğrenci itirazının oluştuğu gözlenmektedir. Ayrıca doğru bir planlama yapıldığı ve tüm altyapının yeterli olduğu kabul edilmiş olsa bile güvenlik için kullanılan kişisel verilere erişim ile ilgili kurumların talebi öğrencilerde kaygıyı arttırmakta ve güvenlik-özgürlük dengesi de başka bir tartışmayı tetiklemektedir.

Bu araştırma kapsamında İstanbul üniversitesi açıköğretim/uzaktan eğitim programlarındaki 2020-2021 eğitim öğretim yılında yürütülen çevrimiçi ölçme değerlendirme sürecinin aşamaları ve beraberinde yürüyen güvenlik tartışmaları ele alınmıştır.

Yöntem

Bu çalışma uzaktan ölçme değerlendirme yöntemlerine ve bu alanda özellikle sınav güvenliğinin nasıl sağlandığına yönelik genel bir bakış sunmak ve İstanbul Üniversitesi AUZEF çevrimiçi sınav uygulaması ve çevrimiçi sınav için geliştirilen güvenlik tedbirlerinin avantaj ve dezavantajları ile birlikte irdelenmesi amacıyla yapılmış nitel bir durum çalışmasıdır. Davey'e göre (1991) durum çalışmaları gerçekte ortamda neler olduğuna bakma, sistematik bir biçimde verileri toplama, analiz etme ve sonuçları ortaya koyma yoludur (akt. Aytaçlı, 2012). Bu tür durum çalışmalarının daha sonraki araştırmacılar için bir veri kaynağı olması beklenir.

Buna göre İstanbul Üniversitesi Açık ve Uzaktan Eğitim Fakültesi örnekleminde pandemi süresince yapılan çevrimiçi sınav uygulaması ölçme yöntem ve teknikleri bakımından incelenmiştir. Sınav uygulaması için kullanılan yazılım ve bu bağlamda sınav güvenliği için alınan tedbirler ve ardışık iki eğitim-öğretim dönemi için bu tedbirlerdeki değişim gözlemlenmiştir. Bunun yanında Türkiye'deki açıköğretim programlarında pandemi sürecindeki mevcut uygulamalar araştırılmış ve bu alanda yapılan çalışmalar ile birlikte çevrimiçi sınav uygulamasında ölçme değerlendirme yöntemleri ve uygulanan güvenlik tedbirleri karşılaştırmalı olarak analiz edilmiş ve sunulmuştur. Çalışmanın yürütüldüğü kurumun Pandemi öncesi ve sonrasında hiç bir değişim yapmadan uyguladığı bağıl değerlendirme sisteminin işleyişine yönelik bir dersin pandemi öncesi ve pandemi sürecinde harf notu aralıkları karşılaştırılmıştır. Bu karşılaştırmada örneklem olarak kullanılan program ve ders, öğrenci itirazının yoğunluğuna göre seçilmiştir.

Bu çalışmada veriler kurumun çevrimiçi sınav uygulamasının gerçekleştirildiği yazılımdan, doküman incelemesi ve araştırmacıların yaptığı gözlemler ile elde edilmiştir. Ayrıca bağıl değerlendirme sisteminin işleyişine yönelik öğrenci itirazları İÜ AUZEF Talep yönetim sisteminden, ders geçme puan aralıklarına yönelik veriler İÜ AKSiS sisteminin veri tabanından elde edilmiştir. Araştırmacılar aynı zamanda sürecin içinde yer alan birer gözlemcidir. Çalışmanın geçerliliğini arttırmak amacıyla birden fazla veri kaynağı kullanılmış ayrıca oluşturulan sonuç raporları için hem uzman görüşüne başvurulmuş hem de farklı zamanlarda araştırmacılar tarafından sonuçlar tekrar gözden geçirilmiştir.

Sonuçlar

Çalışma sonucunda kitle eğitiminde -en az 1000 kişinin bulunduğu programlar- ölçme süreci ağırlıklı olarak çevrimiçi uygulanan çoktan seçmeli testler ile yapıldığı, araştırma örnekleminde yer alan kurumda, öğrenme yönetim sistemi dışında kurum içinde geliştirilen bir yazılım aracılığı ile sürecin yürütüldüğü ve yazılımda kurumun ihtiyaçlarına göre özelleştirmeler yapıldığı görülmektedir. Çevrimiçi sınavın ilk uygulandığı 2019 bahar döneminde 10 günlük bir sınav süresi ile öğrencilere sınavlara geniş bir zaman aralığında girme hakkı verilirken, 2020 bahar döneminde grup kopya girişimlerinin azaltılması amacıyla her bir dersin sınavının 1 gün ile sınırlandırıldığı, tüm sınavların ise 1 hafta içinde tamamlandığı, sınavlara girişin öğrenci numarası ve şifresi ile yapıldığı, kamera ya da ek bir güvenlik yazımının kullanılmadığı görülmüştür. Ancak sınav yazılımının tam ekran modundan çıkılması ile uyarı verdiği, birden fazla cihaz ile sisteme giriş yapıldığında ve durumun devamı halinde sınavın iptal olduğu, ayrıca her öğrenci ekranının özelleştirildiği ve özellikle medya yolu ile ekran/soru paylaşılması halinde kaynak öğrencinin tespit edilebildiği bu yol ile de sınavlarda kopya girişimlerine tedbir alındığı görülmektedir.

Yapılan çevrimiçi sınavlarda çoktan seçmeli testlerden oluşan geniş bir soru havuzunun bulunduğu, her bir soru için kapsam geçerliğini ve sınav güçlüğünü istenen düzeyde tutmak amacı ile soruya ait bölüm/konu bilgisi ile soru güçlüğü ve soru kullanım sayısı bilgisinin tutulduğu ve bu bilgilere göre her öğrencinin ekranına farklı bir sorunun gelmesinin mümkün olduğu görülmüştür. Dolayısı ile çevrimiçi sınavlarda sabit bir sınav formu olmadığı gibi farklı sorular farklı sıralarda öğrencilerin karşısına çıkmaktadır. Soru ile karşılaşan öğrencinin soruya cevap vererek veya boş bırakarak geçebildiği, boş bırakılan sorular için tekrar cevap verme imkânının tanınmadığı görülmüştür.

Çevrimiçi sınavlarda değerlendirme sistemi değişmediği için birtakım tartışmalar yapılmaktadır. Çalışma kapsamında incelenen çocuk gelişimi programına ait derslerde harf notu aralıklarının klasik sınav uygulaması aralıklarına göre yüksek olduğu, bağıl değerlendirme ile birlikte düşünüldüğünde yükselen geçme notlarının bazı öğrencilerde kaygıya ve itirazlara sebep olduğu görülmüştür.

Bu çalışmanın uzaktan ölçme ve değerlendirme yöntemlerinde mevcut uygulamalara bir projeksiyon tutarak bilimsel uygulama örneklerinin geliştirilmesi için yol göstermesi, ayrıca çevrimiçi sınav, bağıl değerlendirme ve "e-sinavda" güvenlik tartışmalarına bir bakış açısı getirmesi beklenmektedir.

Kaynaklar

Acar-Güvendir, M. ve Özer-Ozkan, Y. (2021). Uzaktan eğitimin değerlendirmeye yansımaları: Çevrim içi sınavlar mı sınıf içi sınavlar mı? *Dijital Ölçme ve Değerlendirme Araştırmaları Dergisi, 1*(1), 22-34. https://doi.org/10.29329/dmer.2021.285.2

Aksu Dünya, B., Aybek, E. C. ve Şahin, M. D. (2021). Yükseköğretimde uzaktan ölçme ve değerlendirme deneyimleri: Üç devlet üniversitesinden bir örnek. *Ahi Evran Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 7*(1), 232 – 244. https://doi.org/10.31592/aeusbed.804016

Akyürek, M. İ. (2020). Uzaktan eğitim: Bir alanyazın taraması. *Medeniyet Eğitim Araştırmaları Dergisi, 4*(1), 1-9. https://dergipark.org.tr/tr/pub/mead/issue/56310/711904

Anadolu Üniversitesi Açıköğretim, İktisat ve İşletme Fakülteleri Öğrenci Değerlendirme Sistemi Esasları (2020). Anadolu Üniversitesi, 12 Mayıs 2020. https://www.anadolu.edu.tr/acikogretim/yonetmelikler-ve-esaslar-yonergeler

Atılgan, H., Yurdakul, B. ve Öğretmen, T. (2012). Öğrenci başarısının belirlenmesinde bağıl ve mutlak değerlendirme üzerine bir araştırma. *İnönü University Journal of The Faculty of Education, 13*(2), 79-98. https://dergipark.org.tr/tr/pub/inuefd/issue/8696/108625

Aytaçli, B. (2012). Durum çalışmasına ayrıntılı bir bakış. *Eğitim Bilimleri Dergisi, 3*(1), 1-9.

Balta, Y. (2013). An Examination on Various Measurement and Evaluation Methods Used in Online Distance Education. *Journal of Turkish Studies, 3*(8), 37-45. https://doi.org/10.7827/TurkishStudies.4271

Baran, Ö. H. (2020). Açık ve uzaktan eğitimde ölçme ve değerlendirme. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi (AUAd), 6*(1), 28-40.

Benson, R., and Brack, C. (2010). Online learning and assessment in higher education (1st ed.). Cahndos Publishing. https://www.elsevier.com/books/online-learning-and-assessment-in-higher-education/benson/978-1-84334-577-0

Bozkurt, A., & Uçar, H. (2018). E-öğrenme ve e-Sınavlar: Çevrimiçi ölçme değerlendirme süreçlerinde kimlik doğrulama yöntemlerine ilişkin öğrenen görüşlerinin incelenmesi. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 14*(2), 745-755. https://doi.org/10.17860/mersinefd.357339

Cabi, E. (2016). Uzaktan eğitimde e-değerlendirme üzerine öğrenci algıları. *Yükseköğretim ve Bilim Dergisi, 6*(1), 94-101. https://doi.org/10.5961/jhes.2016.146

Can, E. (2020). Coronavirüs (Covid-19) Pandemisi ve pedagojik yansımaları: Türkiye'de açık ve uzaktan eğitim uygulamaları. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi, 6*(2), 11-53.

Dikmen, S. ve Bahçeci, F. (2020). Covid-19 Pandemisi sürecinde yükseköğretim kurumlarının uzaktan eğitime yönelik stratejileri: Fırat üniversitesi örneği. *Turkish Journal of Educational Studies, 7*(2), 78-98. https://doi.org/10.33907/turkjes.721685

Duman, B. (2011). Sınıf öğretmeni adaylarının bağıl değerlendirmeye ilişkin görüşleri. *Education Sciences, 6*(1), 536-548.

Kınalıoğlu, İ. H. ve Güven, Ş. (2011, 2-4 Şubat). *Uzaktan eğitim sisteminde öğrenci başarısını ölçülmesinde karşılaşılan güçlükler ve çözüm önerileri.* XIII. Akademik Bilişim Konferansı, İnönü Üniversitesi, Malatya.

Kim, N., Smith, M. J., and Maeng, K. (2008). Assessment in online distance education: A comparison of three online programs at a university. *Online Journal of Distance Learning Administration, 11*(1), 1-16. https://www.westga.edu/~distance/ojdla/spring111/kim111.html

Sari, H. (2020). Evde kal döneminde uzaktan eğitim: Ölçme ve değerlendirmeyi neden karantinaya almamalıyız? *Uluslararası Eğitim Araştırmacıları Dergisi, 3*(1), 121-128. https://dergipark.org.tr/tr/pub/ueader/issue/55302/730598

Solak, H. İ., Ütebay, G., ve Yalçın, B. (2020). Uzaktan eğitim öğrencilerinin basılı ve dijital ortamdaki sınav başarılarının karşılaştırılması. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi, 6*(1), 41-52. https://dergipark.org.tr/en/download/article-file/1179715

Yılmaz, R. (2017). Problems experienced in evaluating success and performance in distance education: A case study. *Turkish Online Journal of Distance Education, 18*(1), 39-39. https://doi.org/10.17718/tojde.285713

Değişen madde fonksiyonuna farklı bir bakış açısı: Göz izleme

Münevver Başman, Emine Burcu Tunç, Soner Kotan ve Müge Uluman Mert

Anahtar kelimeler: Değişen madde fonksiyonu, göz izleme, süre metriği, uzamsal metrik, sayısal metrik

Giriş

Değişen Madde Fonksiyonu (DMF), ölçülmesi istenen değişken açısından bireylerin yeteneklerine göre eşleştirilmesi ve daha sonra farklı gruplardaki bu bireylerin maddeyi farklı yanıtlama olasılıklarına sahip olduklarının, istatistiksel olarak ortaya konulmasıdır (Camilli ve Shepard, 1994; Zumbo, 1999). DMF madde yanlılığını belirlemek için bir ön adımdır ve Educational Test Service –ETS- tarafından 1986'da geliştirilmiş ve yanlılık analizlerinde bir standart haline gelmiştir (Roever, 2005). DMF kapsamında yapılmış çok sayıda araştırma mevcuttur ancak DMF'nin kaynağıyla ilgili uzmanlar tarafından ortak bir görüşe varılamamaktadır (Karami ve Nodoushan, 2011).

Yapılmış olan çalışmalarda, çeviri problemleri ve kültürel farklılıklar (Asil, 2010; Girl ve Khaliq, 2001), maddelerin çoktan seçmeli veya açık uçlu oluşu (Feng, 2008; Henderson, 1999; Qian, 2011; Zenisky ve diğ., 2003), farklı puanlama modelleri (Gelin ve Zumbo, 2003; Henderson, 2001; Tunç ve Kutlu, 2018) ve madde içerikleri (Liu ve Wilson, 2009; Mendes-Barnett ve Ercikan, 2006; Ong ve diğ., 2011) DMF nedenleri arasında gösterilmiştir. DMF üzerinde etkisi olduğu merak edilen alanlardan bir diğeri ise göz hareketlerinin izlenmesidir.

Göz izleme, gözün izleme örüntüsünün kaydedilmesini, gösterilen farklı ekranlara bakarken harcanan zamanı, görsel ilginin dağıldığı noktaların tam olarak neresi olduğunun belirlenmesini sağlar. Genellikle gözün bir noktaya 100ms den fazla sabitlenmesine sabitleşme ve sabitleşmeden sabitleşmeye atlanmasına ise sekme denilir (Zhang and Shen, 2001; Zhao and Zuo, 2006). Genellikle, gözün sabitleştiği nokta ilgiyi yansıtıyorken, sabitleşme süresi ise ilginin miktarının göstergesidir. Sabitleşme süresi ve yerleri bireylerin okuma stratejilerini, sahip oldukları ön bilgilerini ya da deneyimlerini yansıtır (Hyönä ve diğ., 2002).

Göz izleme, birçok farklı çalışma alanında kullanılmaktadır. Bunlar arasında; okuma biçimleri (Paulson ve Jenry, 2002; Rayner ve diğ., 2006), bilgi işlem (Radach ve Kennedy, 2004), birey-bilgisayar etkileşimi (Jacob ve Karn, 2003), sanal öğrenme (e-learning) (Liu ve Zhu, 2012) gösterilebilir. Özellikle

göz izleme teknolojisinin kullanıcı dostu olmasıyla göz izleme çalışmaları eğitim araştırmacılarının ilgisini çekmeye başlamıştır (Tsai ve diğ., 2012). Bununla birlikte test geliştirmeyle ilgili çalışmalar da bulunmaktadır (Stephens ve Sreenivasan, 2002). Ancak yapılmış olan araştırmalar incelendiğinde, DMF nedenlerinin göz izlemeyle araştırıldığı bir çalışmaya ulaşılamamıştır. Bu bağlamda, DMF nedenlerinin göz izleme yöntemi kullanılarak araştırılmasıyla, farklı bir bakış açısı kazandıracağı düşünülmektedir.

Bu araştırmanın amacı, DMF ve DMF göstermeyen maddeler kapsamında eşleştirilmiş bireyler için göz izlemeye ait süre, uzamsal ve sayısal metrikleri incelemektir.

Yöntem

Bu araştırmanın amacı, DMF'yi göz izlemeye ait süre, uzamsal ve sayısal metrikler kapsamında incelenmektir. Bu bağlamda araştırmanın tarama modeli niteliği taşıdığı ifade edilebilir. Çalışma grubunu, Marmara Üniversitesi Atatürk Eğitim Fakültesinin farklı programlarında öğrenim gören 435 kadın ve 134 erkekten oluşan toplam 619 lisans öğrencisi oluşturmaktadır. Araştırmanın amacı doğrultusunda, 619 kişiden aynı yetenek düzeyinde olan kadın ve erkek çiftleri eşleştirilmiştir. Göz izlemeye gönüllü olarak katılmayı kabul eden 9 kadın ve 9 erkek olmak üzere, toplamda 18 öğrenci göz izleme sürecine dahil edilmiştir.

Araştırmanın gerçekleştirilebilmesi için, farklı programlarda öğrenim gören öğrencilere ortak uygulanabilecek, ilgili konu alanında yeterliliklerini belirlemede işe koşulabilecek ve hâlihazırda geçerlik ve güvenirliği test edilmiş bir ölçme aracına ihtiyaç duyulmuştur. Bu bağlamda daha önce Akalın (2014) tarafından çalışılmış ve DMF'li madde belirlenmiş olan KPSS Genel Yetenek Sözel Testinde yer alan 12 madde kullanılmıştır. Hazırlanan test öğrencilere uygulandıktan sonra, belirlenen öğrenciler göz izleme laboratuvarına alınmıştır.

Çalışmanın göz izleme aşaması, Marmara Üniversitesi Atatürk Eğitim Fakültesi'nde bulunan İnsan-Bilgisayar Etkileşimi Laboratuvarında gerçekleştirilmiştir. Öncelikle Experiment Center isimli yazılım kullanılarak 12 maddeden oluşan form hazırlanmıştır. Toplam 18 öğrencinin bu soruları okumaları ve cevaplamaları sırasındaki göz hareketleri, SMI RED 500 model göz izleme arabirimi ile birlikte çalışan iViewX ve Experiment Center isimli yazılımlar kullanılarak kayıt edilmiştir. Her kayıt öncesinde, göz hareketlerinin daha yüksek doğrulukla kaydı için izleme cihazı kalibre edilmiştir. Son olarak BeGAze isimli analiz yazılımı kullanılarak, öğrencilerin göz hareketlerine ait sayısal, uzamsal ve süre metrikleri çıkarılmıştır.

Sonuçlar

Bu araştırmada DMF ve DMF göstermeyen maddeler kapsamında doğru ve yanlış yanıt verenler için süre metrikleri, uzamsal metrikler ve sayısal metrikler kapsamında farklılık olması beklenmektedir. Bu doğrultuda, odaklanma sayısının, toplam odaklanma süresinin, ortalama odaklanma süresinin, maksimum odaklanma süresinin, minimum odaklanma süresinin, sıçrama sayısının, toplam sıçrama süresinin, ortalama sıçrama süresinin, maksimum sıçrama süresinin, minimum sıçrama süresinin, toplam

sıçrama genişliğinin, ortalama sıçrama genişliğinin farklılaşması düşünülmektedir. İlgili sonuçlarla birlikte öğrencilerin okuduğunu anlama metinlerini anlamak için gösterdikleri bilişsel çaba, gerçekleştirdikleri bilişsel işlem sürecinin de değişmesi beklenmektedir. Böylelikle DMF çalışmalarına farklı bir bakış açısı kazandırmak hedeflenmektedir.

Kaynaklar

Akalın, Ş. Y. (2014). *Kamu Personeli Seçme Sınavı genel yetenek testinin madde yanlılığı açısından incelenmesi* (Tez No. 381796) [Doktora Tezi, Ankara Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Asil. M. (2010). *Uluslararası Öğrenci Değerlendirme Programı (PISA) 2006 öğrenci anketinin kültürler arası eşdeğerliğinin incelenmesi* (Tez No. 258301) [Doktora tezi, Hacettepe Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Camilli G., and Shepard L. A. (1994). *Methods for identifying biased test items* (Vol. 4). Sage Publications, Inc.

Feng. Y. (2008). *Difference in gender differential item functioning patterns across item format and subject area on diploma examinations after change in administration procedure* (Publication No. 304409289) [Doctoral dissertation, University of Alberta]. ProQuest Dissertations & Theses Global.

Gelin, M. N., and Zumbo. B. D. (2003). DIF results may change depending on how an item is scored: An illustration with the center for epidemiological studie depression (CES-D) scale. *Educational and Psychological Measurement, 63*, 65-74. https://doi.org/10.1177/0013164402239317

Girl. M. J., & Khaliq. S. N. (2001) Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement, 38*(2), 164-187. https://doi.org/10.1111/j.1745-3984.2001.tb01121.x

Henderson. D. L. (1999). *Investigation of differential item functioning in exit examinations across item format and subject area.* (Publication No. 304543522) [Doctoral dissertation, University of Alberta] ProQuest Dissertations & Theses Global.

Henderson D. L. (2001, 10-14 April). *Prevalence of gender DIF in mixed format high school exit examinations.* Paper presented at the Annual Meeting of the American Educational Research Association. Seattle, WA.

Hyönä, J., Lorch, R. F., Jr., & Kaakinen, J. K. (2002). Individual differences in reading to summarize expository text: Evidence from eye fixation patterns. *Journal of Educational Psychology, 94*(1), 44–55. https://doi.org/10.1037/0022-0663.94.1.44

Jacob, R. J., & Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *the mind's eye* (pp. 573-605). North-Holland.

Karami, H., and Nodoushan M. A. S. (2011). Differential item functioning (DIF): Current problems and future directions. *International Journal of Language Studies. 5*(4), 133-142. https://files.eric.ed.gov/fulltext/ED521872.pdf

Liu, M., & Zhu, Z. (2012). A case study of using eye tracking techniques to evaluate the usability of e-learning courses. *International Journal of Learning Technology, 7*(2), 154-171. https://doi.org/10.1504/IJLT.2012.047980

Mendes-Barnett, S., & Ercikan, K. (2006). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education, 19*, 289-304. https://doi.org/10.1207/s15324818ame1904_4

Ong. Y.M., Williams. J. S., & Lamprianou, I. (2011). Exploration of the validity of gender differences in mathematics assessment using differential bundle functioning. *International Journal of Testing, 11,* 271-293. https://doi.org/10.1080/15305058.2011.555574

Paulson, E. J., & Henry, J. (2002). Does the Degrees of Reading Power assessment reflect the reading process? An eye-movement examination. *Journal of Adolescent & Adult Literacy*, *46*(3), 234-244. https://www.jstor.org/stable/40017130

Qian. J. (2014). An investigation of position effects in large-scale writing assessments. *Applied Measurement in Education, 38*(7), 518–534. https://doi.org/10.1177/0146621614534312

Radach, R., & Kennedy, A. (2004). Theoretical perspectives on eye movements in reading: Past controversies, current issues, and an agenda for future research. *European journal of cognitive psychology*, *16*(1-2), 3-26. https://doi.org/10.1080/09541440340000295

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific studies of reading, 10*(3), 241-255. https://doi.org/10.1207/s1532799xssr1003_3

Roever, C. (2005). That's not fair! Fairness, bias, and differential item functioning in language testing. *SLS Brownbag, 9*(15), 1-14.

Stephens, R., & Sreenivasan, B. (2002). Analysis of substitution test performance using eye movement and video data. *Applied neuropsychology*, *9*(3), 179-182. https://doi.org/ 10.1207/S15324826AN0903_6

Tsai, M. J., Hou, H. T., Lai, M. L., Liu, W. Y., & Yang, F. Y. (2012). Visual attention for solving multiple-choice science problem: An eye-tracking analysis. *Computers and Education*, *58*(1), 375-385. https://doi.org/10.1016/j.compedu.2011.07.012

Tunç, E. B. ve Kutlu, Ö. (2018). İki ve çok kategorili puanlanan maddelerde değişen madde fonksiyonlarının karşılaştırılması. *Başkent University Journal of Education*, *5*(1), 40-50. http://buje.baskent.edu.tr/index.php/buje/article/view/104

Liu. L. O. & Wilson. M. (2009). Gender differences in large-scale math assessments: PISA trend 2000 and 2003. *Applied Measurement in Education, 22*(2), 164- 184. https://doi.org/10.1080/08957340902754635

Zenisky A. L., Hambleton R. K., & Robin F. (2003). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment, 9*(1-2), 61-78, https://doi.org/10.1080/10627197.2004.9652959

Zhang, G., and Shen, M. (2001). Eye tracking techniques in usability testing. *Journal of Ergonomics, 7*(4), 9–14.

Zhao, X., & Zuo, H. (2006). Eye tracker and eye tracking techniques. *Computer Engineering and Applications*, *12*(12), 55.

Zumbo. B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores.* Directorate of Human Resources Research and Evaluation. Department of National Defense.

Gelişen zihin yapısının akademik başarıya etkisinin aracılık modelleriyle PISA 2018 bağlamında incelenmesi

Özge Arıcı ve Özge Altıntaş

Anahtar kelimeler: Gelişen zihin yapısı, PISA, akademik başarı, aracılık etkisi, tekli aracılık modeli

Giriş

İnsanın psikolojik özellikleriyle ilgili merak, Platon, Aristo ve diğer Yunan düşünürlerine kadar uzanmaktadır. Bu düşünürler, akıl, bellek, öğrenme, güdü gibi insan doğasıyla ilgili günümüzdeki psikologların da ilgilendiği pek çok konu hakkında düşünmüşlerdir (Schultz ve Schultz, 2011). İnsana ilişkin bu türden özelliklerin geliştirilebilir mi yoksa sabit ve değiştirilemez mi olduğu tartışması ise oldukça eskidir. Örneğin, zekâ sabit bir özellik midir yoksa geliştirilebilir midir? Bazı insanların potansiyellerini gerçekleştirip diğerlerinin bunu gerçekleştirememesinin nedenlerine ilişkin çalışmalar gerçekleştiren Carol S. Dweck, sabit (fixed) yerine gelişen (growth) bir zihin yapısını (mindset) vurgulayan iki kavrama dayalı bir kuram geliştirmiştir (Dweck, 2000, 2006).

Gelişen zihin yapısına sahip bir birey, kendi yeteneğinin ve zekâsının zamanla gelişebileceğine inanırken sabit zihin yapısına sahip bir birey, bu özelliklerinin deneyimle değiştirilemez olduğuna inanmaktadır. Bireye ilişkin özelliklerin doğuştan belirlendiğini varsayan sabit zihin yapısının aksine gelişen zihin yapısına göre, çaba, uygun stratejiler, yönlendirmeler ve ortamlar yoluyla bu özellikler gelişebilmektedir. Bu zihin yapısına sahip bir bireyin, zorluklardan kaçınan ve çoğunlukla onay isteyen sabit zihin yapısına sahip bir bireyden daha üstün başarı seviyelerine ulaşmak için zorlukları birer fırsat olarak görme ve başarısızlıklarından öğrenme olasılığı ise daha yüksektir (OECD, 2021).

Gelişen zihin yapısına sahip öğrencilerin öğrenme stratejilerinde esnek ve yenilikçi olmaları, daha yüksek düzeyde motivasyon göstermeleri, yaşam boyu öğrenen ve bağımsız bireyler olmaları beklenmektedir. 21. yüzyılda bilginin hızla değişen doğası nedeniyle bu oldukça gerçekçi bir beklentidir (Marzano ve Heflebower, 2012). Öğrenciler yeteneklerinin geliştirilebilir olduğuna inandıklarında, gelecekteki akademik etkinliklerinin sonuçları üzerinde daha fazla düşünmekte; dolayısıyla öğrenmelerine daha fazla odaklanmaktadırlar. Bir başka anlatımla, gelişen zihin yapısına sahip

öğrencilerin gelecekte akademik açıdan daha başarılı olacaklarına olan inançlarının, yüksek beklentiye ve akademik başarı için daha fazla çalışma isteğine yol açtığı belirtilmektedir (Yeager ve Dweck, 2012).

PISA uygulamalarında, 21. yüzyıl becerilerine sıklıkla vurgu yapılarak öğrencilerin belirli bir içeriğe ilişkin bilgilerinin ve rutin olarak gerçekleştirdikleri bilişsel becerilerinin değerlendirilmesinin yanı sıra, öğrencilerin içsel ve kişilerarası yeterlilikleri de eleştirel bir biçimde değerlendirilmektedir (OECD, 2019a, 2019b). Bu kapsamda, PISA 2018 uygulamasında ele alınan kavramlardan biri de gelişen zihin yapısıdır. PISA 2018 sonuçlarına göre gelişen zihin yapısı öğrencilerin okuma becerilerindeki başarı puanlarını pozitif şekilde etkilemektedir (OECD, 2019c, 2021). Gelişen zihin yapısının akademik başarı üzerindeki etkisi alanyazında yapılan diğer çalışmalarla da örtüşmektedir (Claro ve diğ., 2016; McCutchen ve diğ., 2016; Outes ve diğ., 2017; Paunesku ve diğ., 2015; Yeager ve diğ., 2019).

Dweck (2006), gelişen zihin yapısının etkisinin farklı akademik disiplinler için farklı olabileceğini belirtmektedir. Bunun yanı sıra, öğrencilerin zihin yapılarının özellikle matematik ve fen başarısında etkili olduğunu ortaya koyan çalışmalar da yaygınlaşmaktadır. Zekâ ya da matematik ve fen becerilerinin değişmez özellikler olduğuna inanan öğrenciler (sabit zihin yapısına sahip) ile bu özelliklerin geliştirilebilir olduğuna inanan öğrenciler (gelişen zihin yapısına sahip) matematik ve fen başarısı açısından karşılaştırıldıklarında; gelişen zihin yapısına sahip olanların anlamlı düzeyde avantajlı oldukları belirtilmektedir (Blackwell ve diğ., 2007; Dweck, 2008).

Bu bağlamda, gelişen zihin yapısının akademik başarıya etkilerinin derinlemesine incelenmesi önem taşımaktadır. Bunun yanında, gelişen zihin yapısının akademik başarı üzerindeki etkisinin birtakım aracı değişkenler yoluyla incelenmesi de önerilmektedir. Başarısızlık korkusu, göreve yönelik motivasyon ve özyeterlik, gelişen zihin yapısının akademik başarı üzerindeki etkisinin incelenmesi sürecinde dolaylı etkilerinin araştırılmasının önerildiği değişkenlerdendir (Dweck ve Yeager, 2019; Yeager ve Dweck, 2012, 2020; akt., OECD, 2021). Buradan hareketle, öğrencilerin gelişen zihin yapısına ilişkin özelliklerinin akademik başarıya doğrudan ve dolaylı etkilerinin aracılık modelleriyle incelenmesine gerek duyulmuştur. Buna göre araştırmanın amacı, PISA 2018 sonuçlarına göre, Türkiye'deki öğrencilerin gelişen zihin yapısına ilişkin özelliklerinin okuma becerileri, matematik okuryazarlığı ve fen okuryazarlığı başarı puanlarına doğrudan etkileri ile başarısızlık korkusu, göreve yönelik motivasyon ve özyeterlik aracı değişkenleri yoluyla dolaylı etkilerinin aracılık modelleriyle belirlenmesidir.

Yöntem

Bu araştırmada, PISA 2018 sonuçlarına göre Türkiye'deki öğrencilerin gelişen zihin yapısına ilişkin özellikleri ile okuma becerileri, matematik okuryazarlığı ve fen okuryazarlığı başarı puanları arasındaki etkileşimin başarısızlık korkusu, göreve yönelik motivasyon ve özyeterlik değişkenleri aracılığıyla belirlenmesi amaçlandığından genel tarama modellerinden yararlanılmıştır (Büyüköztürk ve diğ., 2018). Çalışma, PISA 2018 Türkiye örneklemi üzerinde yürütülmüştür. PISA 2018'e 79 ülke ve ekonomide, 15 yaş grubundaki 32 milyon öğrenciyi temsilen 600.000'den fazla öğrenci katılmıştır.

Türkiye'deki PISA 2018 uygulamasına ise, İBBS Düzey 1'e göre 12 bölgeyi temsil eden 186 okul ve 6890 öğrenci katılmıştır (MEB, 2019).

Gelişen zihin yapısı, başarısızlık korkusu, göreve yönelik motivasyon ve özyeterlik değişkenlerine ilişkin veriler PISA 2018 öğrenci anketinden; okuma becerileri, matematik okuryazarlığı ve fen okuryazarlığı başarı puanlarına ilişkin veriler başarı testlerinden elde edilmiştir. Araştırma kapsamında, öğrencilerin gelişen zihin yapısına ilişkin özellikleri bağımsız değişken; okuma becerileri, matematik okuryazarlığı ve fen okuryazarlığı başarı puanları bağımlı değişken olmak üzere; başarısızlık korkusu, göreve yönelik motivasyon ve özyeterlik değişkenlerinin aracılık etkisi gösterip göstermediği kurulan tekli aracılık modelleriyle incelenmiştir.

Önerilen modellere ilişkin aracılık etkisi çözümlemeleri bootstrap yöntemiyle yapılmıştır. Bunun nedeni, bootstrap yönteminin aracılık çözümlemelerinde dağılıma ilişkin bir bilgi olmadığında ya da normal dağılım sayıltıları ihlal edildiğinde klasik hiyerarşik regresyon modellerine göre daha iyi sonuçlar vermesidir. Bununla birlikte güven aralıklarının belirlenmesinde daha duyarlı olmasıdır (Shrout ve Bolger, 2002). Bu kapsamda üretilen bootstrap örneklemlerinin sayısı 5000'dir. Çözümlemelerde, Preacher ve Hayes (2004) ile Hayes (2013) tarafından SPSS için geliştirilen ve Andrew Hayes'in web sitesinde de (http://www.afhayes.com) yer alan PROCESS makrolarından yararlanılmıştır.

Sonuçlar

PISA 2018 sonuçlarına göre Türkiye'deki öğrencilerin gelişen zihin yapısına ilişkin özelliklerinin okuma becerileri, matematik okuryazarlığı ve fen okuryazarlığı başarı puanlarına doğrudan ve dolaylı etkilerinin belirlenmesine yönelik yürütülen bu araştırmada, başarısızlık korkusu, göreve yönelik motivasyon ve özyeterlik aracı değişkenleri ele alınmıştır. Doğrudan ve dolaylı etkiler önerilen tekli aracılık modellerinin test edilmesine yönelik aracılık çözümlemeleriyle belirlenmiştir. Buna göre, gelişen zihin yapısının akademik başarı üzerindeki etkilerinin incelendiği tekli aracılık modellerinin tamamında, öğrencilerin gelişen zihin yapısına ilişkin özelliklerinin okuma becerileri, matematik okuryazarlığı ve fen okuryazarlığı başarısına doğrudan etkilerinin manidar olduğu belirlenmiştir. Başka bir anlatımla, öğrencilerin gelişen zihin yapısına ilişkin inançları arttıkça okuma becerileri, matematik okuryazarlığı ve fen okuryazarlığı başarı puanlarında da artış olmaktadır.

Gelişen zihin yapısının öğrencilerin okuma becerilerine, matematik okuryazarlığına ve fen okuryazarlığına etkisinin başarısızlık korkusu aracılığında incelendiği tekli aracılık modellerinde dolaylı ve toplam etkilerin tümü manidardır. Yapılan etki büyüklüğü hesaplamalarına göre, gelişen zihin yapısının okuma becerileri üzerindeki toplam etkisinin %6'sı başarısızlık korkusu aracı değişkeni tarafından oluşturulan dolaylı etkiyle açıklanmaktadır. Bu oran matematik okuryazarlığı için, %4; fen okuryazarlığı için ise, %5'tir.

Özyeterlik aracı değişkeni ile kurulan tekli aracılık modellerinin tümünde de aracılık etkileri manidardır. Yapılan etki büyüklüğü hesaplamalarına göre okuma becerileri, matematik okuryazarlığı ve fen okuryazarlığı alanlarındaki başarı puanlarının her biri için gelişen zihin yapısının toplam etkilerinin

%4'ü özyeterlik değişkeni tarafından oluşturulan dolaylı etkiler aracılığıyla açıklanmaktadır. Öte yandan, göreve yönelik motivasyon değişkeninin matematik okuryazarlığı ve fen okuryazarlığı üzerinde aracılık etkileri manidardır ve etki büyüklükleri sırasıyla %2 ve %1'dir. Ancak, öğrencilerin okuma becerileri üzerinde bu değişkene ilişkin manidar bir aracılık etkisi saptanmamıştır.

Bu çalışmanın iki önemli çıktısı bulunmaktadır. Bunlardan ilki, gelişen zihin yapısı ile akademik başarı arasındaki ilişkiyi aracılık modelleriyle aydınlatmak ve Türk alanyazınına bu kapsamda katkı sağlamaktır. İkincisi ise, gelişen zihin yapısının akademik başarıyı nasıl etkilediğinin yorumlanmasında; başarısızlık korkusu, göreve yönelik motivasyon ve özyeterlik değişkenlerinin rolünü vurgulamaktır. Konunun önemi, OECD'nin (2021) "*Sky's the limit: Growth mindset, students, and schools in PISA-Sınır gökyüzü: PISA'da gelişen zihin yapısı, öğrenciler ve okullar*" raporunda vurgulanmış ve bu konunun sonraki PISA uygulamalarında daha ayrıntılı çalışılacağı belirtilmiştir. Aynı raporda, öğrencilerin gelişen zihin yapılarının geliştirilmesiyle, ekonomik yoksulluğun öğrencilerin akademik başarısı üzerindeki olumsuz etkilerinin potansiyel olarak azaltabileceği de vurgulanmıştır. Bu vurgu, gelişen zihin yapısının başarı üzerindeki etkisine ilişkin bu araştırmadan elde edilen sonuçlarla da ilişkilendirildiğinde, öğrencilerin gelişen zihin yapısını geliştirmeye yönelik çalışmaların Türkiye'deki okul uygulamalarına dahil edilmesi bir gereklilik olarak ortaya çıkmaktadır.

Kaynaklar

Blackwell, L. S., Trzesniewski, K. H., and Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development, 78*(1), 246-263. https://doi.org/10.1111/j.1467-8624.2007.00995.x

Büyüköztürk, Ş, Kılıç-Çakmak, E., Akgün, Ö. E., Karadeniz, Ş. ve Demirel, F. (2018). *Eğitimde bilimsel araştırma yöntemleri* (25. baskı). Pegem Akademi.

Claro, S., Paunesku, D., & Dweck, C. S. (2016). Growth mindset tempers the effects of poverty on academic achievement. *Proceedings of the National Academy of Sciences, 113*(31), 8664-8668. https://doi.org/10.1073/pnas.1608207113

Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality, and development.* Psychology Press Taylor & Francis Group.

Dweck, C. S. (2006). *Mindset: The new psychology of success.* Random House.

Dweck, C. S. (2008). *Mindsets and math/science achievement.* Prepared for the Carnegie Corporation of New York-Institute for Advanced Study Commission on Mathematics and Science Education. https://www.growthmindsetmaths.com/uploads/2/3/7/7/23776169/mindset_and_math_sciscie_achievement_-_nov_2013.pdf

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach.* Guilford Press.

Marzano, R. J. & Heflebower, T. (2012). *Teaching and assessing 21st century skills.* Marzano Research.

McCutchen, K. L., Jones, M. H., Carbonneau, K. J., & Mueller, C. E. (2015). Mindset and standardized testing over time. *Learning and Individual Differences 45*, 208-213. https://doi.org/10.1016/j.lindif.2015.11.027

MEB (2019). PISA 2018 Türkiye ön raporu. *Eğitim Analiz ve Değerlendirme Raporları Serisi, No: 10*. MEB Yayınları. http://www.meb.gov.tr/meb_iys_dosyalar/2019_12/03105347_PISA_2018_Turkiye_On_Raporu.pdf

OECD (2019a). *PISA 2018 results (Volume I): What students know and can do*. PISA, OECD Publishing. https://doi.org/10.1787/5f07c754-en

OECD (2019b). *PISA 2018 assessment and analytical framework*. PISA, OECD Publishing. https://dx.doi.org/10.1787/b25efab8-en

OECD (2019c). *PISA 2018 results (Volume III): What school life means for students' lives*. PISA, OECD Publishing. https://doi.org/10.1787/acd78851-en

OECD (2021). *Sky's the limit: Growth mindset, students, and schools in PISA*. PISA, OECD Publishing. https://www.oecd.org/pisa/growth-mindset.pdf

Outes, I., Sanchez, A., & Vakis, R. (2017). *Growth mindset at scale – Increasing school attainment by affecting the mindset of pupils and teachers*. [Project]. World Bank. https://riseprogramme.org/sites/default/files/inline-files/Outes-Leon,%20Ingo,%20Sanchez,%20Alan,%20Vakis,%20Renos.%20%20Project-%20Growth%20Mindset%20at%20Scale.pdf

Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S., & Dweck, C. S. (2015). Mind-Set interventions are a scalable treatment for academic underachievement. *Psychological Science*, *26*(6), 784-793. https://doi.org/10.1177/0956797615571017

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods. Instruments & Computers, 36*(4), 717-731. https://doi.org/10.3758/BF03206553

Schultz, D. P., & Schultz, S. E. (2011). *A history of modern psychology* (10th ed.). Cengage Learning.

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods, 7*(4), 422-445. https://doi.org/10.1037/1082-989X.7.4.422

Yeager, D. S., & Dweck, C. S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educational Psychologist*, *47*(4), 302-314. http://dx.doi.org/10.1080/00461520.2012.722805

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., Carvalho, C. M. … Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature, 573*, 364-369. https://doi.org/10.1038/s41586-019-1466-y

Bayesçi ve frekansçı faktör analizi yöntemlerinin karşılaştırılması

Mehmet Taha Eser ve Gökhan Aksu

Anahtar kelimeler: Bayes, faktör analizi, ölçme modeli

Giriş

İstatistiğin gelişim sürecinde temel anlamda Klasik (veya Frekansçı, Berkeley istatistigi) yaklaşım ve Bayesyen yaklaşım olmak üzere iki farklı felsefi yaklaşımın ağırlık sahibi olduğu göze çarpmaktadır. Pek çok konu ve kavramın ele alınışında ve işlenişinde Frekansçı ve Bayesyen yaklaşımın birbirine alternatif olduğu görülmektedir. Bayesçi yaklaşımın kendi disiplini olan alternatif bir yaklaşım olduğu göz önünde bulundurulduğunda birçok istatistiksel kavram bu yaklaşım kapsamında farklı şekillerde yorumlanmaktadır (Ekici, 2005; Ekici, 2009; Etz ve Vandekerckhove, 2016).

Bayes teoremi matematiksel istatistiğin önemli bir teoremidir. Bu teorem; herhangi bir durumun modelini oluşturmada evrensel doğruları ve gözlemleri kullanarak sonuçlar üretmeyi amaçlar. Kesinlik içermeyen bir bilginin tahmininde, gözlemleri ve sübjektif görüşleri kullanması bu yaklaşımı diğer klasik istatistiksel yöntemlerden ayıran en önemli özelliğidir (Box ve Tiao, 1992; Congdon, 2003; Ekici, 2009)

Frekansçı ve Bayesçi yaklaşım zaman içerisinde kendi içlerinde gelişerek ve kendine has özellikleri belirginleşerek birbirlerinden teorik olarak farklılık göstermişlerdir. Frekansçı yaklaşım tümdengelim, Bayesçi yaklaşım ise tümevarım yöntemiyle paralellik göstermektedir. Genellikle Frekansçı yaklaşım nedensellik ilkesinin tanımlayıcı yorumuna; Bayesçi yaklaşım ise nedensellik ilkesinin olasılıklı yorumuna daha yakın görünmektedir (Ekici, 2009; Kruschke, 2010). Klasik yaklaşımlarda analizler normal dağılım varsayımı altında gerçekleştirilir. Bu varsayıma göre kovaryans matrisinin dağılımı ilgili örneklem boyutu büyükse verilerin örneklemdeki dağılımı normal dağılıma yaklaşır. Ancak araştırmacılar için bu varsayımların sağlanması pek mümkün olmayabilir. Özellikle tıp ve psikoloji gibi alanlarda yürütülen çalışmalarda büyük örneklem sayısına ulaşmak ya da davranış ve sosyal bilimler gibi eksik gözlemlerin olduğu araştırmalarda çok değişkenli normallik varsayımını sağlamak pek mümkün olmayabilir. Bu durumda parametre ve standart hata tahminleri yanlı sonuçlar verme eğiliminde olacaktır (Ekici, 2009; Ozechowski, 2014; Erkan, 2019). Fakat Bayesçi yaklaşımın temel alındığı analizlerde ham veriler ve önsel bilgiler kullanılarak sonsal dağılımlar elde edilir. Bu sayede klasik yaklaşımların temel alındığı analizlerin yetersiz kaldığı durumlarda Bayesçi yaklaşım en uygun çözüm sağlanmaktadır. Frekansçı yaklaşım genel olarak sıfır hipotezinin (boş hipotez) test edilmesine odaklanmaktadır ve çoğu zaman p değerinin hatalı

kullanımı sonucunda bilimin tekrarlanabilirlik krizini arttırmaktadır. Frekansçı yaklaşım kapsamında boş hipotez sadece bir sonraki teste kadar hayatta kalmaktadır. Bu, frekansçı yaklaşım için önemli bir sınırlılıktır ve bu durum araştırmacıların kendilerini "boş hipotezin yanlış tarafında" bulmalarına neden olmaktadır (Bernardo ve Smith, 1994; Lee, 2004). Bayesçi yaklaşımının bu ve bu gibi sorunların üstesinden gelmenin bir yolu olduğu konusunda genel bir kabul söz konusudur (Szucs ve Ioannidis, 2016; Etz ve Vandekerckhove, 2016).

Açımlayıcı faktör analizi, Thurnstone (1934)'un ulaşmaya çalıştığı mükemmel basit yapıyı keşfetme anlamında geliştirilmiş klasik bir yöntem olarak kabul edilmektedir (Gorsuch, 1983, 2003). Klasik açımlayıcı faktör Analizlerinin gerçekleştirilmesine ilişkin adımların neredeyse tümü belirli bir dereceye kadar araştırmacının keyfiliğindedir (Conti ve diğ., 2014). Klasik Faktör analizi faktör modelinin boyutunun seçilmesi, maddelerin faktörlere tahsis edilmesi, faktör yüklerinin tahmin edilmesi ve birden fazla faktöre yüklenen maddelerin çıkartılması olmak üzere temel olarak dört adımdan oluşmaktadır. Aynı zamanda gizil yapının boyutunun seçimine, faktörlerin çıkartılmasına ve döndürülmesine ilişkin çeşitli yöntemler söz konusudur (Jenrich, 2001; Gorsuch, 2003).

Bayesçi faktör analizi ile araştırmacının amacı teorileri ve önceki inançları (prior) daha iyi yansıtan bir analiz gerçekleştirmektir. Analiz, Markov Chain Monte Carlo (MCMC) yöntemleri yardımıyla gerçekleştirilmektedir. MCMC yöntemleri, karmaşık dağılımlardan türetilen gözlemleri simüle eden ve yüksek boyutlu modellerle çalışılmasına olanak sağlayan modellerdir. Modeldeki parametreler ve gizil değişkenler denklemsel olarak sistematik bir şekilde karşılığı olan sonsal dağılımdan simüle edilir. MCMC'de bilinmeyen parametreler ve gizil değişkenler için tam sonsal dağılımlar çıkarılabileceğinden büyük örneklem varsayımlarına da ihtiyaç duyulmaz (Ozechowski, 2014). Dolayısıyla Bayes kestirimi küçük örneklem büyüklüklerinde de iyi performans göstermektedir ve aynı zamanda Maksimum Olabilirlik (ML) tahmininden farklı olarak, analizlerde Bayes kestirimi kapsamında çok değişkenli normal dağılım varsayımı aranmamaktadır. Bayes kestirimi hesaplama açısından, Maksimum Olabilirlik tahmininden daha karmaşık olan modellerin tahmin edilmesine olanak tanımaktadır (Asparouhov ve Muthen, 2010; Schmitt, 2011). MCMC kapsamında en fazla başvurulan yöntemler Metropolis-Hastings algoritması temelli MCMC örnekleme ve Gibbs örneklemedir.

Yöntem

Araştırmanın kapsamında frekansçı yönteme dayalı klasik açımlayıcı faktör analizi ve Bayesçi yönteme dayalı açımlayıcı faktör analizi sonuçlarını karşılaştırılması amaçlanmaktadır. Bu amaç kapsamında iki farklı yöntemle elde edilen analiz sonuçları arasındaki benzer ve farklı yönlerin ortaya çıkartılması amaçlandığından çalışma nicel araştırma yöntemlerinden karşılaştırmalı çalışmalar modeli altında değerlendirilmektedir (Mills, Bunt ve Brujin, 2006).

Araştırmanın çalışma grubunu, beş farklı liseden 15-18 yaş aralığında yer alan 778 öğrenci oluşturmaktadır. Araştırma kapsamında veri toplama aracı olarak Buss ve Perry Saldırganlık Ölçeği kısa formu kullanılmıştır (Bryant ve Smith, 2001). Formun Türkçe versiyonunun cinsiyet, yaş, ölçme

eşdeğerliği ve değişen madde fonksiyonu anlamında psikometrik özellikleri Kuzucu ve Sarıot-Ertürk (2020) tarafından incelenmiş ve sonuçlar Saldırganlık Ölçeği Kısa Formu'nun Türk kültürü için dört faktörden oluşan, kabul edilebilir düzeyde güvenirliğe sahip, cinsiyetler arası ölçüm farkı olmayan, ergen ve yetişkinler için saldırganlık ölçümünde kullanılabilecek bir ölçme aracı olduğunu göstermiştir.

Araştırma kapsamında R, Jamovi ve Lisrel programlarından yararlanılmıştır. R programı kapsamında gerçekleştirilen Bayesçi açımlayıcı faktör analizi için BayesFM adlı paket kullanılmıştır (Piatek, 2019). Buss ve Perry Saldırganlık Ölçeğinin kısa formunun geliştirildiği orijinal yayında ölçeğe ilişkin dört alt boyutun birbiriyle ilişkili olduğu ve açımlayıcı faktör analizi için faktör çıkarma yöntemi olarak temel eksen faktörleme (principal axis factoring) ve döndürme yöntemi olarak promax yöntemini kullanmışlardır. Doğrulayıcı faktör analizi için ise kestirim yöntemi olarak en çok olabilirlik yönteminin kullanıldığından ve faktör yük değeri olarak kesme puanı 0.4'ün temel alındığından bahsedilmektedir (Buss ve Perry, 1992; Bryant ve Smith, 2001). Araştırmada gerçekleştirilen klasik açımlayıcı faktör analizi kapsamında temel eksen faktörleme yöntemi temel alınarak karşılaştırma anlamında ilişkili alt boyutlar için farklı döndürme yöntemleri kullanılmıştır. Araştırma kapsamında gerçekleştirilen analizlerde faktör yük değeri için kesme değerinin .50 olmasına karar verilmiştir.

Araştırma kapsamında KMO değeri .81 bulunmuş ve Bartlett testinin istatistiksel olarak manidar olduğu (χ^2=2558, sd=66, p< .05) belirlenmiştir. 778 bireyden elde edilen bilgiler sonucu oluşturulduğu göz önünde bulundurulduğunda çalışma grubu büyüklüğünün yeterli düzeyde olduğu söylenebilir.

Sonuçlar

Araştırma kapsamında, farklı döndürme yöntemleri altında elde edilen faktör sayılarının değişmediği ve ölçme aracının dört alt boyutlu bir yapıya sahip olduğu belirlenmiştir.

Aynı faktör çıkarma yöntemi altında çalışma kapsamında ele alınan beş farklı faktör döndürme yöntemine göre elde edilen açıklanan varyans, ortalama hataların karekökü (RMSEA), Tucker-Lewis indeksi (TLI) ve göreli ki-kare (pseudo chi-square) değerlerinin aynı olduğu belirlenmiştir.

Çalışmada hem frekansçı hem de Bayesci faktör analizi ile ölçme aracının dört alt boyutlu bir yapıya sahip olduğu bulgusu benzerlik göstermektedir. Yöntemler arasında elde edilen farklılık ise ölçekten çıkarılması gereken madde sayısı ile ilgilidir. Elde edilen bu farklılıklara ilişkin hangi sonuçların rapor edilmesi gerektiğini belirleme noktasında doğrulayıcı faktör analizi gerçekleştirilmiştir.

Her üç model için de CFI, RFI, NFI, GFI ve AGFI değerlerinin aynı olması ve göreli ki-kare ile RMSEA değerleri göz önünde bulundurulduğunda Bayesci faktör analizi yöntemiyle oluşturulan ölçme modelinin diğer modellere göre daha iyi bir model olduğu söylenebilir. Bu durumda 12 maddelik modelin tercih edilmesi gerektiği sonucuna ulaşılmıştır (Bollen ve Long, 1993). Aynı zamanda hem frekansçı hem de Bayesci faktör analizi yöntemleri ile kurulan üç farklı modele ilişkin uyum iyiliği istatistiklerinin benzer olduğu belirlenmiştir.

Kaynaklar

Asparouhov, T., & Muthén, B. (2010). *Bayesian analysis of latent variable models using Mplus* (version 5). http://www.statmodel.com/download/bayesadvantages18.pdf

Bernardo J. M., & Smith., A. F. M. (1994). *Bayesian theory.* Wiley.

Bollen, K. A., & Long, J. S. (1993) *Testing structural equation models.* Sage.

Box, G.E.P., & Tiao, G.C. (1992). *Bayesian inference in statistical analysis.* John Wiley & Sons.

Bryant, F. B., & Smith, B. D. (2001). Refining the architecture of aggression: A measurement model of the buss-perry aggression questionnaire. *Journal of Research in Personality, 35*(2), 138-167. https://doi.org/10.1006/jrpe.2000.2302

Buss, A. H., & Perry, M. (1992). The aggression questionnaire. *Journal of Personality and Social Psychology, 63*(3), 452. https://doi.org/10.1037/0022-3514.63.3.452

Congdon, P. (2003). Modelling spatially varying impacts of socioeconomic predictors on mortality outcomes. *J Geograph Syst 5,* 161–184. https://doi.org/10.1007/s10109-003-0099-7

Ekici, O. (2005). *Bayesçi regresyon ve Win- BUGS ile bir uygulama* (Tez No. 217581) [Yüksek lisans tezi, İstanbul Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Ekici, O. (2009). İstatistikte Bayesçi ve klasik yaklaşımın farklılıkları. *Balıkesir Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 12*(21), 89-101.

Erkan, G. (2019). *Klasik ve bayesci yapısal eşitlik modellerinde parametre tahminlerinin karşılaştırılması: sıralı kategorik verilerle bir uygulama* (Tez No. 547271) [Yüksek lisans tezi, Hacettepe Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Etz, A., and Vandekerckhove, J. (2016). A bayesian perspective on the reproducibility project: Psychology. *PLoS One, 11*(2), e0149794. https://doi.org/10.1371/journal.pone.0149794

Gorsuch, R. (1983). *Factor analysis* (2nd ed.). Lawrence Erlbaum Associates.

Gorsuch, R. L. (2003). Factor analysis. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology*, (2th ed., pp. 143–164). John Wiley & Sons Inc.

Conti, G., Frühwirth-Schnatter, S., Heckman, J. J., & Piatek, R. (2014). Bayesian exploratory factor analysis. *Journal of Econometrics, 183*(1), 31-57. https://doi.org/10.1016/j. jeconom.2014.06.008.

Jennrich, R. I. (2001). A simple general procedure for orthogonal rotation. *Psychometrika, 66,* 289-306. https://doi.org/10.1007/BF02294840

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences, 14*(7), 293–300. https://doi.org/10.1016/j.tics.2010.05.001

Kuzucu, Y., and Sarıot-Ertürk, Ö. (2020). Psychometric properties of Turkish version of aggression questionnaire short form in adolescents and adults. *Journal of Measurement and Evaluation in Education and Psychology, 11*(3), 243-265. https://doi.org/10.21031/epod.683176

Lee, P. M. (2004) *Bayesian statistics: An introduction.* Wiley.

Mills, M., Bunt, G. G., & Brujin, J. (2006). Comparative research persistent problems and promising solutions. *International Sociology, 21*(5), 619-631. https://doi.org/10.1177/0268580906067833

Ozechowski T. J. (2014). Empirical Bayes MCMC estimation for modeling treatment processes, mechanisms of change, and clinical outcomes in small samples. *Journal of Consulting and Clinical Psychology, 82*(5), 854-867. https://doi.org/10.1037/a0035889

Piatek, R. (2019). *Bayes FM: Bayesian inference for factor modeling* (version 0.1.3) [Computer software]. https://cran.r-project.org/package=BayesFM

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment, 29*(4), 304-321. https://doi.org/10.1177/0734282911406653

Szucs, D., and Ioannidis, J. P. (2016). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology, 15*(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797

Thurstone, L. L. (1934). The vectors of mind. *Psychological Review, 41*(1), 1-32. https://doi.org/10.1037/h0075959

Yükseköğretim kurumları kalite göstergelerinin ağ analizi ile incelenmesi

Akif Avcu

Giriş

Üniversiteler, yüzyıllar boyunca topluma hizmet eden potansiyel profesyonelleri, iş adamlarını, siyasi liderleri, dini ve sosyal âlimleri yetiştirmek, değerlerini zenginleştirmek ve kaynaklarını geliştirmek için çok önemli bir role sahip olmuştur (Mustard, 1998). Dolayısıyla yükseköğretimin toplumun gelişmesinde hayati bir rol oynadığı söylenebilir. Bu sebeple, daha fazla ve daha iyi yükseköğretim kurumları olmadan, gelişmekte olan ülkelerin küresel bilgiye dayalı ekonomiden yararlanmanın giderek daha zor olacağını belirtilmiştir (World Bank, 2000).

Günümüzde üniversiteler mali kısıtlamalar, artan rekabet, artan hesap verebilirlik talepleri ve gelişen teknolojilerin getirdiği zorluklarla karşı karşıyadır (Gappa ve diğ., 2007). Yükseköğretim kurumları ve bu kurumların paydaşları kanıt ve veriye dayalı yöntemlerle performans değerlendirmesine yönelik artan bir talep vardır. Bu nedenle, yükseköğretim kurumları birtakım somut göstergeler ile performanslarını değerlendirmektedir. Bu göstergeler, üniversite sıralamalarını başta olmak üzere farklı birçok parametreyi içerebilir. Kalite göstergelerine dayalı bu değerlendirmeler kurumlara ve paydaşlara çıktı sağladığı gibi, üniversite karşılaştırmalarında ve yükseköğretime yönelik alınan kararlarda kullanılmaktadır. Mevcut durumda ise bu göstergelerin neler olması gerektiği tartışılsa da (Vos, 1996), yükseköğretim kurumları kalite göstergelerine daha fazla önem verme konusunda artan beklentilerle karşılaşmaktadır (Gover ve Loukkola, 2018). Genel hatları itibariyle, kalite göstergeleri üç sınıfa ayrılabilir: eğitim girdileri, eğitim çıktıları ve eğitim süreçleri.

Sosyal bilimlerde göstergelerin ağ analizi ile incelenmesi gittikçe popüler olan bir çalışma alanıdır (örn. Zhou ve diğ., 2017; Dalampira ve Nastis, 2020). Üniversite kalite göstergelerin de birbirleriyle ilişkili olduğu bilinmektedir (Federkeil, 2008) ve karmaşık etkileşimler ağı ile birbirine bağlanmıştır. Buradan hareketle, ülkemizdeki üniversiteler özelinde yükseköğretimde kalite göstergelerinin ağ analizi ile incelenmesinin bu etkileşimlerin daha iyi anlaşılmasına yardımcı olacağı düşünülmektedir. Bu bilgiler ışığında, gerçekleştirilen bu çalışmanın amacı, Türkiye'de bulunan yükseköğretim kurumlarının kalite göstergelerinin ağ analizi ile incelenmesi ve en önemli ve en az önemli ağların ortaya çıkartılmasıdır.

Yöntem

Gerçekleştirilen bu çalışmada Yükseköğretim Kurumuna bağlı Yükseköğretim Kalite Kurulu tarafından sunulan yükseköğretim kurumları kalite göstergelerine ait veriler kullanılmıştır. Kalite göstergeleri verileri, Kuruma Ait Bilgiler (örn. öğrenci sayısı, öğretim elemanı sayısı), Kalite Güvence Sistemi (örn. stratejik hedeflerin gerçekleşme yüzdesi, sıralamalar, değişim programlarına katılım), Eğitim Öğretim, Araştırma Geliştirme ve Toplumsal Katkı alt başlıklarında toplam 82 gösterge için 207 yükseköğretim kurumundan elde edilen sayısal bilgileri içermektedir. Gerçekleştirilen bu çalışmada analizler 133 yükseköğretim kurumu için 2020 yılına ait 19 farklı kalite gösterge değişkeni ile gerçekleştirilmiştir. Veri setine dâhil edilen değişkenler ve üniversiteler belirlenirken verilerin en fazla %5 kayıp olmasına ve değişkenlerin benzer olmamasına (örneğin üniversite sıralamaları ile ilgili benzer değişkenler vardır) dikkat edilmiştir. %5 altında kayıp veri ise kayıp veri yapısı incelendikten sonra doğrusal regresyon yöntemi ile ataması gerçekleştirilmiştir.

Verilerin analizi R (R core team, 2020) programında bulunan *qgraph* (Epskamp ve diğ., 2012) paketi ile gerçekleştirilmiştir. Analiz sürecinde öncelikli olarak verilerin çok değişkenli normalliği incelenmiştir. Daha sonrasında yükseköğretim kurumlarına ait gelişim göstergelerinin ağ yapısı kestirilmiştir. Ağ yapısının kestirilmesi sürecinde düzenlileştirilmiş kısmı korelasyonları kullanılmıştır. Kestirilen ağ yapısı içerisindeki en önemli ve en az önemli göstergeler güç, yakınlık ve arasında olma merkezilik ölçütleri ile değerlendirilmiştir. Ayrıca, ağ yapısına en az katkı sağlayan göstergeler Zhang'ın kümeleme katsayısı (Zhang ve Horvath, 2005) ile belirlenmiştir. Son olarak, göstergelerin ne düzeyde bir ağ oluşturabildiğinin belirlenmesi için Küçük Dünya (Small Wordness) indeksi hesaplanmıştır.

Sonuçlar

Elde edilen bulgular, yükseköğretimdeki kalite göstergeleri arasındaki geçişkenliğin yüksek olduğunu ve bu göstergelerin ağ yapısına katkı sağladıkları görülmüştür. Ağ içerisinde idari ve araştırmaya dair stratejik hedeflerin gerçekleştirilme yüzdelerinin ilişkili olduğu, yayın sürecinde uluslararası işbirlikleri ile Türkiye sıralamalarının ilişkili olduğu görülmüştür. Güç merkezilik indisi en yüksek olan kalite göstergeleri uluslararası yayınlardaki ortaklıklar ve üniversitelerin Türkiye sıralamaları olarak belirlenmiştir. Aynı zamanda yayınlardaki uluslararası ortaklıkların yakınlık ve arasında olmak merkezilik indisleri için de yüksek düzeyde indeks değerlerine sahip olduğu görülmüştür. Bu bulgu, yayınlardaki uluslararası ortaklıkların kalite göstergeleri ağı üzerindeki en etkili gösterge olduğunu göstermiştir. Ayrıca, Zhang'ın kümeleme katsayısı incelendiğinde, kümeleme indis değeri en düşük göstergelerin haftalık ortalama ders saatleri ile yabancı uyruklu öğrenci oranları olduğu görülmüştür. Yani bu göstergeler kalite göstergeleri ağına en az katkı sağlayan değişkenlerdir.

Kaynaklar

Dalampira, E. S., & Nastis, S. A. (2020). Mapping sustainable development goals: A network analysis framework. *Sustainable Development*, *28*(1), 46-55. https://doi.org/10.1002/sd.1964

Epskamp, S., Cramer, A. O., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). Qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, *48*(1), 1-18. https://doi.org/10.18637/jss.v048.i04

Federkeil, G. (2008). Rankings and quality assurance in higher education. *Higher Education in Europe*, *33*(2-3), 219-231. https://doi.org/10.1080/03797720802254023

Gappa, J. M., Austin, A. E., & Trice, A. G. (2007). *Rethinking faculty work: Higher education's strategic imperative*. Jossey-Bass.

Gover, A. & Loukkola, T. (2018*). Enhancing quality: From policy to practice (Enhancing Quality through Innovative Policy & Practice/EQUIP).*

Mustard, F. (1998). *The nurturing of creativity: The role of higher education*. Oxford University Press. Karachi.

R Core Team (2020). *R: A language and environment for statistical computing* (version 3.0.1) [Computer Software]. http://www.R-project.org/

Vos, R. (1996). *Educational indicators: What's to be measured? Integration and regional programs department, inter-American institute for social development*, Inter-American Development Bank.

World Bank (2000). *Report on higher education in the developing countries: Peril and promise*. Oxford University Press. https://documents1.worldbank.org/curated/en/345111467989458740/pdf/multi-page.pdf

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4, Article17. https://doi.org/10.2202/1544-6115.1128

Zhou, X., Moinuddin, M., & Xu, Z. (2017). *Sustainable Development Goals Interlinkages and Network Analysis: A practical tool for SDG integration and policy coherence*. Institute for Global Environmental Strategies (IGES), Kanagawa, Japan.

Uzaktan öğretimde uygulanan elektronik portfolyoların genellenebilirlik ve çok yüzeyli Rasch ölçme modeli ile değerlendirilmesi

İsmail Karakaya, Nazira Tursynbayeva ve Umur Öç

Anahtar kelimeler: Elektronik portfolyoların değerlendirilmesi, farklılaşan puanlayıcı davranışları, genellenebilirlik kuramı, çok yüzeyli Rasch ölçme modeli

Giriş

Mektupla başlayan uzaktan eğitim uygulamaları radyo, televizyon, telefon, bilgisayar ve internet teknolojilerinin gelişmesi ve değişen dünya koşulları (Pandemi vs.) sayesinde eğitimde önemli bir yer almıştır (Çallı ve diğ., 2002). Uzaktan eğitime geçiş sürecinde olası olumsuzlukları ortadan kaldırmak adına internet temelli farklı uygulamalar denenmiştir. Ancak bilgi edinmenin yanında karar verme, eleştirel düşünme, problem çözme gibi üst düzey zihinsel becerilerin geliştirilmesi konusunda yeterli ilerleme sağlanamamıştır (Barak ve diğ., 2007). Üst düzey becerilerin kazandırılması ya da geliştirilmesi ders içeriklerinin temel amacı haline gelmiştir (Boddy ve diğ., 2003; Riedler ve Eryaman, 2016; Watts ve diğ., 1997). Üst düzey zihinsel becerilerin kazandırılması bir süreçtir ve bu süreç bireyden bireye değişmektedir. Kutlu ve diğ. (2014) üst düzey zihinsel becerilerin bireylerin bilişsel, duyuşsal ve psikomotor özelliklerinin tümüne atıfta bulunduğunu belirtmişlerdir. Üst düzey zihinsel beceriler performans görevleri, portfolyo gibi süreç odaklı ve tamamlayıcı ölçme araçlarıyla ölçülebilirler. Uzaktan eğitim sürecinde öğrencilerin üst düzey zihinsel becerilerini geliştirmelerine destek olan ve öğrencilerini gelişim sürecini etkin bir şekilde yansıtan ölçme araçlarından biri de elektronik portfolyodur (Egan, 2012; Jenson, 2011). Elektronik portfolyo sayesinde hem süreç hem de sonuç gözlemlenebilirken elektronik portfolyo eğitimde her kademede kullanıma uygundur (Barker, 2005). Elektronik portfolyonun eğitimin her aşamasında kullanılabilir olması ve bireylerin üst düzey bilişsel becerilerini geliştirmelerine yardımcı olması gibi önemli avantajları olsa da portfolyoların puanlama sürecinde katılık, cömertlik, halo etkisi, merkezi eğilim gibi sorunlar meydana gelebilmektedir (Bahar ve diğ., 2006; Chang ve diğ., 2011; Hung, 2012).

Bu çalışmanın amacı, yükseköğretimde öğrenci başarısının belirlenmesinde kullanılan öğrenci elektronik portfolyoların birden fazla puanlayıcı tarafından değerlendirilmesi bağlamında genellenebilirlik kuramı ve çok yüzeyli rasch ölçme Modeli kullanılarak değerlendirilmesidir. Bu amaç

doğrultusunda öz değerlendirme, akran değerlendirme ve birden fazla öğretim üyesinin değerlendirilmeleri incelenmiştir.

Yöntem

Bu çalışma Genellenebilirlik Kuramı'nın Gazi Üniversitesi Gazi Eğitim Fakültesi 2020-2021 Bahar döneminde ölçme ve değerlendirme dersini alan 44 öğrencinin elektronik portfolyolarının değerlendirilmesi sonucu elde edilen verilere uygulanmasını temel aldığı ve elektronik portfolyo sisteminin kullanılabilirliğini ortaya çıkarmaya çalıştığı için betimsel araştırma kapsamındadır.

Araştırmada öğretim üyeleri ve öğrenciler olmak üzere iki farklı katılımcı grubu bulunmaktadır. Bu kapsamda öğrencilerin öz ve akran değerlendirmeleri, öğretim üyelerinin de rubrik değerlendirme sonuçları araştırmaya dahil edilmiştir. Öğrenciler, Gazi Üniversitesi Eğitim Fakültesi 2020-2021 bahar döneminde Ölçme ve Değerlendirme Dersi'ni alan toplamda 44 Türk dili ve Edebiyatı ve Biyoloji bölümü öğrencisinde oluşmaktadır. 20 öğrencinin verisi ise, puanlama esnasında eksiklikler olduğundan dolayı analiz dışı bırakılmıştır. 24 öğrenci verisiyle analize devam edilecektir.

Çalışmada puanlayıcı kaynaklı varyansı belirlemek amacıyla öncelikle EduG (Swiss Society for Research in Education Working Group, 2010) programından faydalanılmıştır. Puanlayıcı kaynaklı varyans yüksek çıktığı için problemli puanlayıcı davranışını belirlemek amacıyla FACETS (Linacre, 2017) bilgisayar programı kullanılmıştır. Tüm puanlayıcılar tüm öğrenci e-portfolyo dosyalarını puanladığından tamamen çaprazlanmış desen kullanılmıştır. Çaprazlanmış desenlerde norm temelli karşılaştırmalar kullanıldığında puanlayıcıların cömertliği ya da katılığı öğrencilerin sınav sonucundaki sıralamasında farklılığa sebep olmayacaktır. Bu problemli puanlayıcı davranışı tüm öğrencileri etkileyeceğinden durum dengelenecektir. Puan ortalamasının alındığı düşünüldüğünde puan dağılımındaki kayma sabit olacaktır. Bu sayede cömertlik ya da katılık dışında başka bir problemli puanlayıcı davranışı olmazsa norm temelli karşılaştırmaların tercih edildiği çaprazlanmış desenler oldukça güvenilir olacaktır (Wolfe, 2004).

Çalışmada kullanılan veriler de Genellenebilirlik Kuramı'nda tümüyle çaprazlanmış desen olarak tasarlanmıştır. Verilerden desene ait genellenebilirlik ve güvenirlik katsayıları hesaplanmıştır. Ayrıca problemli puanlayıcı davranışlarından olan katılık ve cömertlik de incelenmiştir. Yapılan çalışmada öncelikle puanlayıcılardan kaynaklanan varyansı belirlemek için EduG (Swiss Society for Research in Education Working Group, 2010) programından faydalanılmıştır. Sonrasında puanlayıcılardan kaynaklanan varyansın yüksek olduğu tespit edilmiş ve problemli puanlayıcı davranışlarını belirlemek amacıyla FACETS (Linacre, 2017) bilgisayar programı kullanılmıştır.

Sonuçlar

Yapılan çalışmada Gazi Üniversitesi Eğitim Fakültesi 2020-2021 Bahar döneminde Ölçme ve Değerlendirme Dersi'ni alan 24 öğrencinin elektronik portfolyolarının değerlendirilme sonuçları incelenmiştir. Sonuçlar değerlendirilirken tümüyle çaprazlanmış desen b x ö x p (b: birey, ö: ölçüt, p:

puanlayıcı) uygulanmıştır. Verilerde 5 puanlayıcı, 7 ölçüt ve 24 birey bulunmaktadır. Bu araştırmada birey, ölçütler ve puanlayıcılar olmak üzere üç yüzey bulunduğundan her bir yüzey için ölçüm değerleri ve uygunluk istatistikleri incelenmiştir.

Değerlendirme sonucunda elektronik portfolyolar değerlendirilirken 124 anlamlı etkileşimin yanlılık değerlerinin işaretleri incelendiğinde yanlılık değerleri pozitif işaretli olan 71 etkileşimde farklılaşan puanlayıcı cömertliği, yanlılık değerleri negatif işaretli olan 51 etkileşimde ise farklılaşan puanlayıcı katılığı olduğu görülmektedir. Değerlendirmede bulunan puanlayıcıların istatistiksel olarak farklı puanlayıcı davranışları gösterdikleri, öğretim üyelerinin yapmış oldukları puanlamaların ve öğrencilerin yapmış oldukları puanlamaların kendi içlerinde benzer puanlayıcı davranışı gösterdikleri belirlenmiştir. Öğrenci ve öğretim üyelerinin puanlamaları arasında fark olmasının temel sebebi öğrencilerin yapmış oldukları öz ve akran değerlendirmelerinin benzer puanlanmış olması (birçoğu tam puan) olması ile açıklanabilir. Tespit edilen farklılaşan puanlayıcı davranışlarını olabildiğince en aza indirebilmek geçerli ve güvenilir sonucu elde edebilmek için puanlayıcılara puanlayıcı eğitimi verilmesi ve planlanmaktadır.

Kaynaklar

Bahar, M., Nartgün, Z., Durmuş, S. ve Bıçak, B. (2006). *Geleneksel-alternatif: Ölçme ve değerlendirme öğretmen el kitabı.* Pegem Akademi.

Barak, M., Ben-Chaim, D. and Zoller, U. (2007). Purposely teaching for the promotion of higher-order thinking skills: A case of critical thinking. *ResSciEduc, 37*, 353–369.

Barker, K. C. (2005). *E-portfolio for the assessment of learning.* http://www.futured.com/documents/FuturEdePortfolioforAssessmentWhitePaper.pdf

Boddy, N., Watson, K., & Aubusson, P. (2003). A trial of the five Es: A referent model for constructivist teaching and learning. *Research in Science Education, 33*, 27–42.

Chang, C. C., Tseng, K. H., Chou, P. N., & Chen, Y. H. (2011). Reliability and validity of Web-based portfolio peer assessment: A case study for a senior high school's students taking computer course. *Computers and Education, 57*(1), 1306-1316. https://doi.org/10.1016/j.compedu.2011.01.014

Çallı, İ., Bayram, Y. ve Karacadağ, M. C. (2002, 23-24 Mayıs). Türkiye'de Uzaktan Eğitimin Geleceği ve E-Üniversite, Açık ve Uzaktan Eğitim Sempozyumu, Anadolu Üniversitesi.

Egan, J. P. (2012). E-portfolio formative and summative assessment: Reflections and lessons learned. *Proceedings of Informing Science & IT Education Conference (InSITE), 12,* 417-422. http://proceedings.informingscience.org/InSITE2012/InSITE12p417-422Egan0119.pdf

Hung, S. T. A. (2012). A wash back study on e-portfolio assessment in an English as a foreign language teacher preparation program. *Computer Assisted Language Learning, 25*(1), 21–36. https://doi.org/10.1080/09588221.2010.551756

Kutlu, Ö., Doğan, C. D. ve Karakaya, İ. (2014). *Öğrenci başarısının belirlenmesi/ performansa ve portfolyoya dayalı durum belirleme.* Pegem Akademi.

Linacre, J. M. (2017). *A user's guide to FACETS: Rasch-model computer programs.* MESA Press.

Riedler-Eryaman, M., & Eryaman M. Y. (2016). Complexity, diversity and ambiguity in teaching and teacher education: practical wisdom, pedagogical fitness and tact of teaching. *International Journal of Progressive Education, 12*(3), 172-186

Watts, M., Jofili, Z., & Bezerra, R. (1997). A case for critical constructivism and critical thinking in science education. *Research in Science Education, 27*(2), 309–322.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46*, 35-51. http://psycnet.apa.org/record/2004-19990-003.

Ortaöğretime geçiş sınavının özel öğrenme güçlüğü olan öğrencilere göre ölçme değişmezliğinin incelenmesi

Selma Şenel

Giriş

Özel gereksinim bireylere uygulanan testlerden elde edilen sonuçların, özel gereksinimli olmayan bireylerin test sonuçları ile karşılaştırılabilir olması test puanlarının geçerliği için kritik önem arz eder. Uluslararası düzeyde yapılan çok sayıda araştırma bu puanların geçerli ve karşılaştırılabilir olması için önemli bir gayret olduğunu göstermektedir (Bolt ve Ysseldyke, 2008a; Elliott ve diğ., 2018; Gregg ve Nelson, 2012; Reise, 1990; Rogers ve diğ., 2019). Özellikle bireyin yaşamında önemli olan bazı kararlara dayanak olan puanların elde edildiği geniş ölçekli testlerde, engel gruplarına göre ölçme değişmezliği ön plana çıkmaktadır.

Testin bireylere, ölçtüğü özellik dışındaki bir özelliğe göre farklı bir grupta yer alması dolayısıyla, farklı davranmaması gerekir. Bu nedenle, bireyin engel durumu ile ilgili bir ölçme yapılmıyorsa, engel durumu test sonuçlarına etki etmemesi gerekir. Bireyin engeli ya da özel gereksiniminin test sonuçlarının etkilememesi için okuyucu, ek süre, işaretleyici vb. test düzenlemeleri uygulanır (National Center for Learning Disabilities, 2005). Ancak, yapılan araştırmalar, test düzenlemelerinin engel gruplarına göre ölçme değişmezliğini sağlamadığını göstermektedir (Knickenberg ve diğ., 2020; Şenel, 2021). Bu nedenle geniş ölçekli testlerin engel gruplarına göre ölçme değişmezliğinin incelenmesi bir gereklilik olmaktadır.

Özel öğrenme güçlüğü bireyin dinleme, konuşma, okuma, yazma, heceleme, dikkat yoğunlaştırma, matematik, akıl yürütme, motor ve organizasyon becerilerini olumsuz etkileyen yapısal bir sorun olarak ifade edilmektedir. Özel öğrenme güçlüğü, zekâ düzeylerinin orta ya da orta üstünde olmasına rağmen bazı akademik alanlardaki beklenmedik performans düşüklüğü ile karşımıza çıkmaktadır. Okuma güçlükleri "disleksi", yazma güçlükleri "disgrafi", matematik güçlükleri ise "diskalkuli" olarak isimlendirilmektedir. Ancak bu güçlüklerin daha ziyade birlikte gözlendiği bilinmektedir. Uluslararası istatistiklere göre, özel öğrenme güçlüğü olan öğrenciler, tüm özel gereksinimli öğrenciler içerisinde en yüksek orana sahiptir. Okul çağındaki nüfusun %5 ile %15'nin özel öğrenme güçlüğü yaşadığı belirtilmektedir (Bolt ve Ysseldyke, 2008b; Elliott ve diğ., 2018; First, 2013; Grigorenko ve diğ., 2019; Rogers ve diğ., 2019). Ancak Türkiye'de tanılama süreçlerindeki aksayan yönler nedeniyle bu gereksinim grubunun oranının çok daha düşük rapor edildiği bilinmektedir.

Öğrenme güçlüğü olan öğrencilerde ek süre düzenlemesi en sık kullanılan düzenleme olup, bireylerin bu düzenleme ile daha iyi performans gösterdikleri gözlenmiştir (Camara ve diğ., 2005; Gregg ve Nelson, 2012). Türkiye'de de, özel öğrenme güçlüğü olan öğrenciler tek kişilik salonlarda, 20 dakika ek süre ile sınava alınmaktadırlar. Öğrenciler ek olarak, okuyucu ve kodlayıcı talep edebilmektedirler (Milli Eğitim Bakanlığı, 2018).

Bu araştırmada Türkiye'de, ortaöğretime geçiş sınavının özel öğrenme güçlüğü gösterip göstermeme durumuna göre ölçme değişmezliğinin incelenmesi amaçlanmıştır. Özel gereksinimli öğrenciler içerisinde en büyük grubun özel öğrenme güçlüğü olanlar olması, öğrencilerin girdikleri ilk geniş ölçekli testin ortaöğretim sınavını olması bu araştırmanın çerçevesinin belirlenmesinde çıkış noktası olmuştur. Bunun yanında, özel öğrenme güçlüğü olan öğrencilere uygulanan ulusal düzeydeki geniş ölçekli testler için, ölçme değişmezliğinin araştırılmamış olması bu araştırmayı önemli kılmaktadır.

Orta Öğretime Geçiş Sınavı: İlk defa 2018 yılında başlayarak resmî ve özel ortaokullar, imam hatip ortaokulları ve geçici eğitim merkezlerinin (GEM) 8'inci sınıflarında öğrenim gören öğrencilerin fen liseleri, sosyal bilimler liseleri, proje uygulayan eğitim kurumları ile mesleki ve teknik Anadolu liselerinin Anadolu teknik programlarına öğrenci yerleştirilmesi amacıyla Bakanlıkça yapılan merkezî sınavıdır (MEB, 2018).

Yöntem

Araştırmada betimsel bir yaklaşımla yürütülmüştür. Araştırmada var olan bir durumun olduğu gibi ortaya koyulması amaçlandığından tarama modelinde bir çalışmadır. Çalışma grubu 2017-2018 eğitim öğretim yılında Merkezi Sınavı alan özel öğrenme güçlüğü olan öğrenciler (n=994) ve 4986 özel gereksinimi olmayan bireyden seçkisiz olarak seçilen özel öğrenme güçlüğü olan öğrenci sayısına denk sayıda (n=1000) normal gelişim gösteren ortaokul 8.sınıf öğrencilerinden oluşmaktadır. Özel öğrenme güçlüğü olan öğrenciler, analizlerde odak grubu oluşturmuştur.

Farklı bir özel gereksinimi ile (görme engeli, zihinsel engeli, işitme engeli, süreğen hastalığı olan, bedensel engelli, dikkat eksikliği ve hiperaktivite bozukluğu, ince motor becerilerinde yetersizlikler) birlikte öğrenme güçlüğü olan öğrenciler araştırma dışında (n=37) tutulmuştur. Ders muafiyeti olanlar, bazı derslerin sınavlarına girmeyenler, İngilizce dışında yabancı dil sınavına girenler araştırma dışında tutulmuştur. Çalışma grubunu oluşturan özel öğrenme güçlüğü olan ve sınavda ek süre düzenlemesi ile alan 994 öğrenci ve eşdeğer olarak seçilen çalışma grubu. 4986 özel gereksinimi olmayan bireyden 1000 kişilik veri random olarak çekilmiştir. Çalışma grubu toplamda 1994 kişiden oluşmaktadır.

Araştırma verileri Milli Eğitim Bakanlığı Ölçme Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü'nden resmi süreçler izlenerek alınmıştır. Merkezi Sınav Milli Eğitim Bakanlığı tarafından ilk kez 2018 yılında sekizinci sınıf öğretim programları esas alınarak uygulanmıştır. Çoktan seçmeli 90 maddeden oluşan sınav, sözel ve sayısal olmak üzere iki ayrı oturumda uygulanan iki bölümden oluşmaktadır. Sözel alan Türkçe (n=20), Din Kültürü ve Ahlak Bilgisi (n=10), T.C. İnkılap Tarihi ve Atatürkçülük (n=10) ile yabancı dil (n=10) alt testlerinden, toplam 50 maddeden oluşmaktadır. Sayısal

alan ise matematik(n=20) ve fen bilimleri (n=20) alt testlerinden, toplamda 40 maddeden oluşmaktadır. Sözel alanın süresi 75 dakika; sayısal alanın süresi 60 dakikadır (MEB, 2018).

Ölçme değişmezliğini incelemek üzere, öncelikle her bir alt testteki maddeler için Değişen Madde Fonksiyonu (DMF) incelenmiştir. DMF analizinde Mantel Haenszel ve Lord'un ki karesi yöntemleri kullanılmıştır. Lord'un Ki karesi yöntemi MTK'ya dayalı bir teknik olduğu için MTK varsayımları test edilmiştir. Ek olarak, Yapısal Eşitlik Modellemesine dayalı tekniklerden Çoklu Grup Doğrulayıcı Faktör Analizi uygulanarak, alt testlerin yapısal, zayıf, güçlü ve katı değişmezlikleri de aşamalı olarak incelenmiştir. Analizler R'da *ltm, difR, lavaan* paketleri ile uygulanmıştır.

Sonuçlar

Araştırma sonucunda, Ortaöğretime Geçiş Sınavında, özel öğrenme güçlüğüne göre Fen bilimleri alt testinde 12 maddenin, Matematik alt testinde 10 maddenin, Türkçe alt testinde 4 maddenin, İngilizce alt testinde 4 maddenin, T.C. İnkılap Tarihi ve Atatürkçülük alt testinde 3 maddenin, Din Kültürü ve Ahlak Bilgisi alt testinde 1 maddenin her iki yönteme göre DMF gösterdiği gözlenmiştir. Buna göre, 90 maddelik testin 34 maddesi her iki yönteme göre DMF göstermektedir. En az bir yöntemde DMF gösteren madde sayısının 67 olması (%74) de dikkate değer bir bulgudur. Aynı zamanda, referans grup olan normal gelişim gösteren öğrenciler lehine, en az bir yönteme göre DMF gösteren madde sayısı da 36'dır.

Çoklu grup doğrulayıcı faktör analizi sonuçlarına göre, tüm alt testlerde yapısal değişmezlik sağlanmaktadır. Yapısal değişmezlik ölçme değişmezliğinin en temel olanıdır. Testlerin, gruplarda aynı faktör yapısına sahip olduğunu gösterir. Ölçme değişmezliğinin ikinci aşama incelemesi olan, zayıf değişmezlik tüm alt testlerde sağlanmamaktadır. Bu sonuca göre, sınavın tüm alt testlerinde, faktör yükleri gruplar arasında değişiklik göstermektedir. Zayıf değişmezlik sağlanmadığından, güçlü ve katı değişmezliği de sağlanmaz. Bu bulgular, sonuçların özel öğrenme güçlüğü olan ve olmayan öğrencilerin maddeleri benzer şekilde yanıtlamadıklarını göstermektedir. Özel öğrenme güçlüğü olan ve olmayan öğrencilerin madde ortalamalarının aynı olmadığı da belirtilebilir. Bu farklılıkların nedeninin madde yanlılığı olup olmadığına ilişkin uzman görüşlerine dayalı incelemeler yapılabilir.

Kaynaklar

Bolt, S. E., & Ysseldyke, J. (2008a). Accommodating students with disabilities in large-scale testing a comparison of differential item functioning (DIF) identified across disability types. *Journal of Psychoeducational Assessment*, *26*, 121–138. https://doi.org/10.1177/0734282907307703

Bolt, S. E., & Ysseldyke, J. (2008b). A comparison of differential item functioning (DIF). *Journal of Psychoeducational Assessment*, *26*(2), 121–138. https://doi.org/10.1177/0734282907307703

Camara, W. J., Copeland, T., & Rothschild, B. (2005). *Effects of extended time on the SAT ® I: Reasoning test score growth for students with learning disabilities.* https://files.eric.ed.gov/fulltext/ED562679.pdf

Elliott, S. N., Kettler, R. J., Beddow, P. A., & Kurz, A. (2018). *Handbook of accessible instruction and testing practices* (2nd ed.). Springer International Publishing. https://doi.org/10.1007/978-3-319-71126-3

First, M. B. (2013). *DSM-5 handbook of differential diagnosis*. American Psychiatric Pub.

Gregg, N., & Nelson, J. M. (2012). Meta-analysis on the effectiveness of extra time as a test accommodation for transitioning adolescents with learning disabilities: More questions than answers. *Journal of Learning Disabilities*, *45*(2), 128–138. doi:10.1177/0022219409355484

Grigorenko, E. L., Compton, D. L., Fuchs, L. S., Wagner, R. K., Willcutt, E. G., & Fletcher, J. M. (2019). Understanding, educating, and supporting children with specific learning disabilities: 50 years of science and practice. *American Psychologist*. doi:10.1037/AMP0000452

Knickenberg, M., Zurbriggen, C., Venetz, M., Schwab, S., & Gebhardt, M. (2020). Assessing dimensions of inclusion from students' perspective–measurement invariance across students with learning disabilities in different educational settings. *European Journal of Special Needs Education*, *35*(3), 287–302. doi:10.1080/08856257.2019.1646958

Milli Eğitim Bakanlığı (2018). *Sınavla öğrenci alacak ortaöğretim kurumlarına ilişkin merkezî sınav başvuru ve uygulama klavuzu*. https://www.meb.gov.tr/meb_iys_dosyalar/2020_05/06105923_BasYvuru_ve_Uygulama_KYlavuzu_2020_GuYncel.pdf

National Center for Learning Disabilities. (2005). *No Child Left Behind: Determining appropiate assessment accommodations for students with disabilities*. Schwab Learning. https://files.eric.ed.gov/fulltext/ED486451.pdf

Reise, S. P. (1990). A comparison of item and person-fit methods of assessing model data fit in IRT. *Applied Psychological Measurement*, *14*(2), 127–137. https://doi.org/10.1177/014662169001400202

Rogers, C. M., Thurlow, M. L., Lazarus, S. S., & Liu, K. K. (2019). *A summary of the research on effects of test accommodations: 2015-2016 (NCEO Report 412)*. http://www.nceo.info

Şenel, S. (2021). Assessing measurement invariance of Turkish "Central Examination for Secondary Education Institutions" for visually impaired students. *Educational Assessment, Evaluation and Accountability, 33,* 621-648. https://doi.org/10.1007/s11092-020-09345-5

Çeşitli BOBUT algoritmalarının alt yetenek düzeylerindeki performanslarının karşılaştırılmasına dönük bir simülasyon çalışması

Selma Şenel

Anahtar kelimeler: Özel gereksinimli öğrenciler, CAT algoritması, Monte Carlo simülasyonu, madde seçim yöntemi, düşük yetenek aralığı

Giriş

Bilgisayar Ortamında Bireye Uyarlanmış Test (BOBUT) yönteminin temel iddialarından biri ölçülen özellik bakımından uçlarda yer alan yeterliklerde geleneksel testlere göre daha kesin ve güvenilir sonuçlar üretmesidir. Ancak, BOBUT'ta de uç yeteneklerin kestiriminin orta yetenektekilere göre daha düşük kesinlikte olduğu, yanlı sonuçlar elde edilebildiği bilinmektedir (Babcock ve Weiss, 2009; Riley, Conrad, Bezruczko ve Dennis, 2007). Düşük yeterliklerde pozitif yanlılık, yüksek yeterlikteki bireylerde ise negatif yanlılık söz konusu olabilmektedir. BOBUT θ=1 yakınlarında daha kesin kestirimler üretmektedir (Magis ve diğ., 2018).

Bu durum, BOBUT'un tüm yeterlik düzeylerini hedefleyen yapısına ters düşmektedir. Bunun temel nedeni, madde havuzunda uç yetenekler için yeterli sayıda ve nitelikli maddeler olmaması olabilir. Nitekim BOBUT uygulamasının performansının temel belirleyicisi, madde havuzunun niteliği ve genişliğidir. BOBUT madde havuzlarında ortalama güçlükteki maddeler daha fazla yer alır. Bunun yanında, test giriş kuralı, madde seçme yöntemi, yetenek kestirim yöntemi, test sonlandırma kuralı gibi algoritma ögeleri de yetenek bakımından uçlarda yer alan bireyler için BOBUT performansını etkileyebilir.

BOBUT'un özel gereksinimli bireylere uygulanan geniş ölçekli testlerde önemli avantajları olduğu belirtilebilir. Bilgisayar tabanlı test düzenlemelerine elverişli olması bunlardan biridir. Bunun dışında kısa süren testlerle, ek süre düzenlemesine gerek kalmamaktadır. Daha çok bilgi sunması ve daha güvenilir test puanları sağlaması da tercih sebebi olabilmektedir (Şenel ve Kutlu, 2018a, 2018b; Stone ve Davey, 2011). Ancak özel gereksinimli öğrencilerin çoğunlukla, alt yetenek düzeylerinde yer aldığı bilinmektedir. BOBUT'un dezavantajlı durumda olan alt yetenek düzeylerinde, ortalama yeterlikteki bireylere göre daha yanlı test sonuçları üretmesi bir problem olarak değerlendirilebilir. Özellikle geniş ölçekli ve çok sayıda özel gereksinimli öğrencinin dâhil olduğu sınavlarda alt yetenek düzeyleri hakkında yeterli kesinlikte bilgi sunan algoritmaların tercih edilmesi önemlidir.

BOBUT performansına ilişkin göstergeler; test uzunlukları, standart hata değerleri, yanlılık(bias) ve RMSEA değerleridir. Daha az maddenin uygulandığı, standart hata değerleri ve RMSEA değerleri düşük, yanlılığı sıfıra yakın olan bir BOBUT'un iyi performans gösterdiğini belirtilebilir. j, yanıtlayıcıları ve N toplam yanıtlayıcı sayısını göstermek üzere *Yanlılık* ve *RMSE* değerlerine ilişkin eşitlikler Eşitlik 1 ve Eşitlik 2'de sunulmuştur. Eşitlik 1 ve 2'ye göre, bireyin kestirilen yetenek düzeyi ile gerçek yetenek düzeyi arasındaki fark ne kadar fazla ise BOBUT performansı o kadar düşük kabul edilir.

$$Yanlılık\ (bias) = \frac{\sum_{j=1}^{N}(\widehat{\theta_j}_j - \theta_j)}{N} \qquad \text{(Eşitlik 1)}$$

$$RMSE\ (Root\ Mean\ Squared\ Error) = \sqrt{\frac{\sum_{j=1}^{N}(\widehat{\theta_j}_j - \theta_j)^2}{N}} \qquad \text{(Eşitlik 2)}$$

Geniş bir yetenek ranjından bireylere uygulama yapılan BOBUT uygulamalarında, yetenek aralığına göre kestirimlerin kesinliğinde önemli değişim göstermeyen yöntemler seçilmesi gerektiğini söylemek yanlış olmayacaktır. Bu belirlemeden yola çıkarak, bu araştırmada alt yetenek düzeylerinde, çeşitli BOBUT algoritmalarının performanslarının karşılaştırılması amaçlanmıştır.

Yöntem

Araştırmada Monte Carlo simülasyonu ile, ideal kabul edilebilecek madde parametresi dağılımlarıyla (a parametresi, uniform dağılımda, 1-2 aralığında; b parametresi, uniform dağılımda, -3-+3 aralığında; c parametresi, uniform dağılımda, 0-0.20 aralığında) (Wainer ve diğ., 2000) 1000 maddelik bir madde havuzu oluşturulmuştur. Ortalaması 0, standart sapması 1 olan normal dağılım gösteren 1000 kişilik bir yetenek dağılımı oluşturulmuştur. Farklı madde seçme yöntemlerinin, yetenek kestirimi yöntemlerinin ve sonlandırma kurallarının uçlarda yer alan bireylerin kestirimindeki BOBUT performansları kıyaslanmıştır.

Kıyaslanacak yöntemlerin seçiminde, alanyazında sıklıkla tercih edilen ve önerilen yöntemlere gidilmiştir. Bayesçi Sonsal Beklenti Kestirimi (Bayesian Expected A Posteriori- EAP), En Yüksek Olabilirlik, (Maximum Likelihood-ML), En Yüksek Sonsal Kestirim (Maximum a Posteriori – MAP) en sık tercih edilen yetenek kestirim yöntemleri olduğu bilinmektedir. Bu nedenle yetenek kestirim yöntemleri olarak üç yöntem tercih edilmiştir. Madde seçme kuralı olarak, En Yüksek Fisher Bilgisi (MFI), en uygun b degeri (bOpt), en uygun yeterlik (thOpt), Kulbak-Leibler sapma kriteri, oransal (proportional) ve ilerlemeli (progressive) yöntemler kullanılmıştır. Standart hataya (SH) dayalı sonlandırma kuralının, alt yetenek düzeylerine ilişkin kestirimler için en güçlü yöntem olduğu belirtilmektedir (Babcock ve Weiss, 2009; Choi ve diğ., 2011). Test uzunluğunun sonlandırma kuralı olarak kullanılması alt yeterlik düzeylerinde önerilmediği ifade edilebilir. Ancak madde kullanım sıklığı ya da içerik kontrolü yapılmadan, tek boyutlu bir yapının testinde ideal uzunluk 15-20 olarak belirtilebilir (Babcock ve Weiss, 2009). Bu araştırmada ikinci bir sonlandırma kriteri olarak "20" madde uzunluğu da koşullara eklenmiştir (SH + 20). Toplamda 36 koşul kıyaslanmıştır.

Sonuçlar

Araştırma sonucunda, ortalama performans göstergeleri bakımından en iyi düzeyde (r[gerçek θ ile kestirilen θ arasındaki korelasyon] >= .95; Bias <= .01; RMSEA< .33; madde sayısı< 18) performans gösteren 11 koşul belirlenmiştir. Bu koşullar ve performans göstergeleri Tablo 1'de sunulmuştur.

Tablo 1

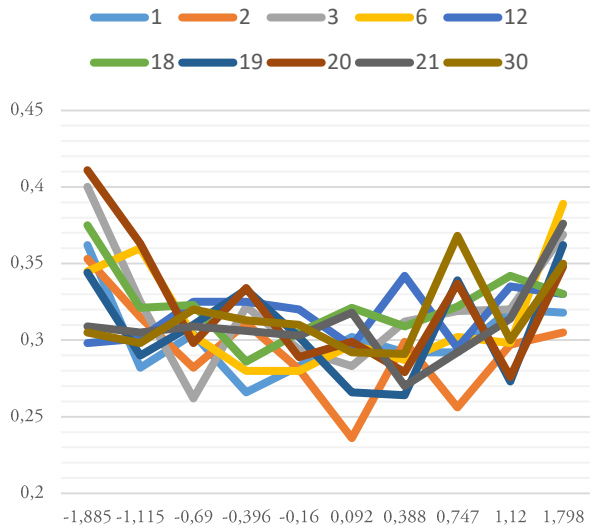
Koşullar ve Performans Göstergeleri

Koşul No	Yetenek kestirim	Madde Seçme Kuralı	Sonlandırma Kuralı	Ortalama Madde Sayısı	r	RMSEA	Yanlılık
1	EAP	MFI	SH	12.9	0.96	0.3032	0.0074
2	EAP	ilerlemeli	SH	14.8	0.96	0.295	0.0014
3	EAP	oransal	SH	17.8	0.95	0.3229	0.0034
4	EAP	thOpt	SH	21.9	0.95	0.3287	-0.0063
5	EAP	bOpt	SH	22	0.95	0.3268	-0.0123
6	EAP	KL	SH	13.3	0.95	0.3161	-0.0053
12	ML	KL	SH	14.6	0.96	0.3172	0.0112
18	BM	KL	SH	12.5	0.95	0.3242	0.0257
19	EAP	MFI	SH + 20	12.9	0.95	0.3101	-0.0053
20	EAP	Aşamalı	SH + 20	15	0.95	0.3261	0.0047
21	EAP	proportional	SH + 20	17.6	0.95	0.3112	-0.0015
24	EAP	KL	SH + 20	13.3	0.95	0.318	-0.0064
30	ML	KL	SH + 20	14.6	0.96	0.3156	-0.0012

Tablo 1'e göre, Bayesçi Sonsal Beklenti Kestirimi (Bayesian Expected A Posteriori- EAP) yetenek kestirim yönteminin ve Kullback- Leibler kriteri madde seçme kuralının genel ortalama bakımından diğer yöntemlere göre daha iyi performans gösterdiği belirtilebilir. Öne çıkan bu 11 koşulun, farklı yeterlik düzeylerindeki gücünü gözlemlemek açısından θ aralıklarına göre RMSEA değerinin değişim grafiği Şekil 1'de, θ aralıklarına göre yanlılık değerinin değişim grafiği Şekil 2'de sunulmuştur. Koşullara ilişkin numaralar Tablo 1'de sunulmuştur.

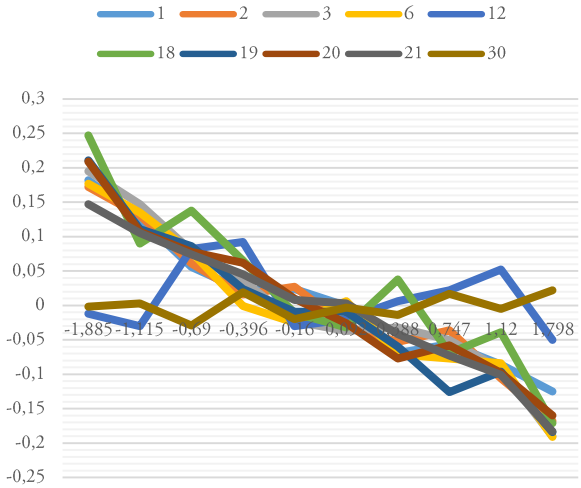
Şekil 1

θ Aralıklarına göre RMSEA Değerinin Değişimi



Şekil 2

θ Aralıklarına Göre Yanlılık Değerinin Değişimi



Şekil 1 ve 2'ye göre RMSEA ve yanlılık ortalaması düşük algoritmalar içerisinden alt yeterlik düzeylerinde en iyi performans göstererek, yeterlik düzeyleri boyunca BOBUT performansı açısından tutarlılık gösteren algoritmalar; En Çok Olabilirlik Yetenek Kestirim Yöntemi, Kullbak-Leibler bilgisi madde seçme kuralı, SH ve SH +20 madde sonlandırma kuralı olarak belirtilebilir.

Tartışma ve Sonuç

Araştırmanın bulgularına dayalı olarak, alt yetenek düzeylerindeki bireylerin hassasiyetle ölçülmesi amaçlanan testlerde, yöntem seçiminde BOBUT performansı güçlü olan En Çok Olabilirlik Yetenek Kestirim Yöntemi, Kullbak-Leibler bilgisi madde seçme yöntemlerine gidilebileceği ifade edilebilir. Bunun yanında, standart hata kuralı ve madde uzunluğu sınırı (20 madde) ile birlikte kullanılan standart hata madde sonlandırma kurallarının her ikisinin de performansının bernzer olduğu gözlenmiştir. Bundan sonraki araştırmalarda, alt yetenek düzeylerindeki bireylerin ve dezavantajlı grupların dahil olduğu BOBUT uygulamaları için; post-hoc simülasyonlarla, sonuçların gerçek veri ile uyumu incelenebilir.

Kaynaklar

Babcock, B., & Weiss, D. J. (2009, June 2). Termination criteria in computerized adaptive tests: Variable-length CATs are not biased. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC conference on computerized adaptive testing*. http://iacat.org/sites/default/files/biblio/cat09babcock.pdf

Choi, S. W., Grady, M. W., and Dodd, B. G. (2011). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement*, *71*(1), 37–53. https://doi.org/10.1177/0013164410387338

Magis, D., Yan, D., & von Davier, A. A. (2018). *Computerized adaptive and multistage testing with R: Using packages catR and mstR. Measurement: Interdisciplinary Research and Perspectives* (C. 16). https://doi.org/10.1080/15366367.2018.1520560

Riley, B. B., Conrad, K. J., Bezruczko, N., & Dennis, M. L. (2007). Relative precision, efficiency and construct validity of different starting and stopping rules for a computerized adaptive test: The GAIN substance problem scale. *Journal of Applied Measurement*, *8*(1), 48–64. https://pubmed.ncbi.nlm.nih.gov/17215565/

Şenel, S. ve Kutlu, Ö. (2018a). Comparison of two test methods for VIS: Paper-pencil test and CAT. *European Journal of Special Needs Education*, *33*(5), 631–645. https://doi.org/10.1080/08856257.2017.1391014

Şenel, S. ve Kutlu, Ö. (2018b). Computerized adaptive testing design for students with visual impairment. *Egitim ve Bilim*, *43*(194), 261–284. https://doi.org/10.15390/EB.2018.7515

Stone, E., & Davey, T. (2011). Computer-adaptive testing for students with disabilities: A review of the literature. *ETS Research Report Series*, *2011*(2), 1–24. https://doi.org/10.1002/j.2333-8504.2011.tb02268.x

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Lawrence Erlbaum Associates.

Bilgisayar ortamında bireye uyarlanmış test uygulamalarında maddeyi yeniden cevaplayabilme

Ömer Faruk Şen ve Hülya Kelecioğlu

Giriş

Teknoloji, psikometri ve eğitim alanlarındaki son gelişmelerle bireye uyarlanmış test uygulamalarının kullanım alanları hızla artmaktadır. Bilgisayar ortamında bireye uyarlanmış test (BBT) uygulamalarının geleneksel testlerden farklı olarak daha kullanışlı olması, kesin ve güvenilir sonuçlar vermesi, bireye özgü soruların yer alması, testte daha az madde ile bireyin yeteneğinin arzu edilen zaman diliminde belirlenmesi, anlık değerlendirmeye olanak sağlamasıyla popülaritesi günden güne artmaktadır. BBT uygulamalarının geleneksel testlere göre bazı üstünlükleri olmasına rağmen aynı şekilde bazı kısıtlamalara sebep olduğu da görülmektedir. BBT uygulamalarının uygulama sürecinde daha önceden cevaplanmış maddeyi yeniden gözden geçirmeye izin verilmemesi en önemli sınırlılıklarındandır (Vispoel ve diğ., 2000).

BBT uygulamalarında maddeyi yeniden gözden geçirmeye izin verilip verilmemesi gerektiği sorunu hem test düzenleyiciler hem de testi alan bireyler için büyük önem arz etmektedir. BBT uygulamaları ulusal çapta yapıldığı düşünüldüğünde maddeyi yeniden cevaplayabilmeye izin verilmediğinde testi alan bireyler için büyük bir strese ve sınav endişesine sebep olmakta ve bireylerin doğru cevabı bildiği halde hata yapmasına sebep olabilmektedir (Lunz ve diğ., 1992). Bu durum sonucunda bireylerin elde edecekleri test puanları yeteneklerinin gerçek bir göstergesi olmayacağı için testin güvenirliğini ve geçerliğini düşürecektir (Vispoel ve diğ., 2000). Bu nedenle BBT uygulamalarını alan bireylerin büyük bir çoğunluğu madde yanıtlamada yeniden cevaplayabilme esnekliğinin sağlanmasını arzu etmektedir (Bowles ve Pommerich, 2001). Ayrıca dikkatsizlik, anlık endişe ve hesaplama hataları gibi nedenlerden dolayı yanlış cevaplanan soruların tekrardan gözden geçirilmesine izin verilmesi ile bireyin yetenek kestirimindeki doğruluğunu artıracağı belirtilmektedir (Vispoel ve diğ., 2000).

Test uygulayıcılar ise maddenin yeniden gözden geçirilmesine izin verildiği durumlarda (G-BBT) onlara sınav süresinin uzaması gibi ekstra maliyetler doğuracağı ve testin geçerliğini düşüreceğini belirtmektedirler (Gershon ve Bergstrom, 1995). Araştırmacılar BBT uygulamalarında cevap değişikliği olursa bilgisayar tarafından seçilen bir sonraki maddenin artık en iyi madde olmayacağı için yetenek

kestirimini olumsuz etkileyeceğini ve testin etkililiğini azaltacağını belirtmektedir (Stocking, 1997; Vispoel ve diğ., 1999). BBT uygulamalarında maddenin yeniden cevaplanmasına izin verilmemesinin bir diğer nedeni de bireylerin test puanı şişirme stratejilerini kullanarak gerçek yetenek düzeyinden daha fazla puan alma ihtimalinin olmasıdır (Jensen, 2017). Bu test stratejilerinden biri olan Wainer stratejisini (Wainer, 1993) kullanan bir birey ilk sorudan itibaren bütün maddeleri bildiği halde yanlış cevaplayarak kolay maddelerin sorulmasını sağlamakta ve maddeyi yeniden gözden geçirmeye izin verildiğinde maddelerin hepsini doğru cevaplamaktadır. Böylece yetenek kestirimi olarak En Çok Olabilirlik (Maximum Likelihood) yöntemi kullanıldığında testten en yüksek puanı elde edebilmektedir. Bir diğer taraftan, bazı araştırmacılar G-BBT uygulamasının olası puan şişirmeye yönelik test stratejilerine nazaran yetenek kestiriminin daha sağlıklı olabileceğine vurgu yapmaktadır (Vispoel ve diğ., 2002).

BBT uygulamalarında sınırlılığı ortadan kaldırmak için uygulamada bireylerin ön bilgilerinden yararlanarak başlangıç değeri tanımlanabilir. Böylece BBT algoritmasının doğası gereği bireyin yeteneğine uygun bir başlangıç değeri tanımlandığında birey yetenek düzeylerine daha uygun maddelerle karşılaşacaktır (Stout ve diğerleri, 2003). Böylece birey cevap değişikliği yapsa bile bu değişikliğin ölçme kesinliği üzerindeki etkisi daha az olacaktır. Bu doğrultuda bu çalışmanın amacı başlangıç değeri tanımlanan BBT uygulamalarında maddeyi yeniden gözden geçirmeye izin verildiğinde ölçme kesinliği yetenek kestirim yöntemlerine göre değişmektedir. Bu amaç doğrultusunda aşağıdaki sorulara cevap aranmıştır.

1. Maddeyi yeniden cevaplamaya izin verilmediği BBT uygulamalarında ölçme kesinliği
   a. Başlangıç değeri olarak (-1,1) aralığında rastgele bir değer, regresyonla belirlenen bir değer tanımlandığında
   b. Yetenek kestirimi olarak Maksimum Olabilirlik Kestirimi (MLE), Sınırlandırılmış Maksimum Olabilirlik Kestirimi (MLEF) ve Beklenen Sonsal Dağılım (EAP N(0,1) ve EAP N(μ, σ)) seçildiğinde

   nasıl değişmektedir?
2. Maddeyi yeniden cevaplamaya izin verildiği BBT uygulamalarında ölçme kesinliği
   a. Başlangıç değeri olarak (-1,1) aralığından rastgele bir değer, regresyonla belirlenen bir değer tanımlandığında
   b. Yetenek kestirimi olarak Maksimum Olabilirlik Kestirimi (MLE), Sınırlandırılmış Maksimum Olabilirlik Kestirimi (MLEF) ve Beklenen Sonsal Dağılım (EAP N(0,1) ve EAP N(μ, σ)) seçildiğinde

   nasıl değişmektedir?

Yöntem

Bir hibrid simülasyon olarak planlanan bu çalışma, bilgisayar ortamında bireye uyarlanmış test uygulamalarında bireylere daha önce sunulan maddeleri incelemelerine izin verildiğinde, başlama

kuralının ve yetenek kestirim yönteminin ölçme kesinliği üzerindeki etkisini incelemeyi amaçlayan temel bir araştırmadır.

Çalışmanın odağı PISA'yı bir bireye uyarlanmış bir test bataryası olarak ele almaktır. PISA test bataryasında öncelikle bireylerin Matematik BBT uygulamalarını aldıkları varsayılmıştır. Buradan elde edilen sonuçlarla başlangıç değeri tanımlanan fen BBT uygulaması gerçekleştirilmiştir. Bu doğrultuda fen ve matematikten en az 10 madde cevaplamış bireyler araştırmaya dâhil edilmiştir. Çalışmanın örneklem büyüklüğü 13606'dür. Fen ve Matematik BBT uygulamalarının madde havuzu ikili puanlanan 171 fen ve 73 matematik sorusundan oluşmaktadır. Madde parametreleri olarak PISA raporunda belirtilen değerler kullanılmıştır (Tablo 1).

Madde parametreleri ile bireylerin yetenekler kestirilmiş ve bireylerin fen ve matematik düzeyleri arasındaki ilişki hesaplanmıştır (Fen = 0,577*matematik + 0,085). Monte Carlo simülasyon yöntemi ile bireylerin fen ve matematik maddelerine ilişkin eksik cevap matrisleri tamamlanmıştır. Bu veriler ışığında öncelikle Matematik BBT simülasyon çalışması gerçekleştirilmiştir. Bu uygulama sonucunda kestirilen matematik yetenekleri ile regresyon eşitliğinden yararlanarak Fen BBT uygulamasının başlangıç değerleri belirlemiştir.

Fen BBT uygulamasında ise iki farklı cevap seti kullanıldı. Monte Carlo simülasyon sonucu elde edilen tam cevap matrisi ile bireylere maddeyi gözden geçirmeye izin verilmeyen BBT simülasyonu gerçekleştirildi. Maddeyi gözden geçirmeye izin verilen modelde ise bireylerin yüksek puan etmek amacıyla önce tüm maddeleri bilerek yanlış cevapladıkları; maddeyi yeniden gözden geçirmeye izin verildiğinde ise tüm maddeleri doğru cevaplamaya çalıştıkları farz edilmiştir. Bireylerin maddeleri amaçlı yanlış cevaplama olasılıkları Cui ve diğer araştırmacıların (2018) varsayımı olan (P+2)/3 denklemi ile hesaplanmıştır. Maddeye gözden geçirmeye izin verilmediği durum referans model olarak ele alınmıştır.

Araştırma kapsamında maddeyi yeniden gözden geçirmeye izin verilen/ verilmeyen, başlangıç değeri belirlemede ön bilginin kullanıldığı/kullanılmadığı, son yetenek kestirimi olarak MLE, MLEF, EAP N(0,1) ve EAP N(μ, σ) kullanıldığı toplam 16 koşul birbiriyle karşılaştırılmıştır. Koşulların hepsinde test uzunluğu 25 madde, geçici yetenek kestirim yöntemi olarak önsel bilgileri dikkate alan EAP N(μ, σ) yöntemi, madde kullanım sıklığı da 0,25'dir. Araştırmada her bir koşul için ölçme kesinliğini belirlemek için RMSE ve yanlılık (bias) değerleri hesaplanmıştır.

Tablo 1

Madde parametreleri

	Parametre	N	Ortalama	Ss	Min	Mak
Matematik	a	73	1,139	0,389	0,220	2,303
	b	73	-0,093	0,771	-1,810	1,710
Fen	a	171	1,171	0,392	0,402	2,475
	b	171	0,025	0,557	-1,412	1,953

Sonuçlar

Maddeyi yeniden gözden geçirmeye izin verilen BBT uygulamalarında en kötü senaryo olan tüm cevapların değiştirildiğinde tüm koşullarda yanlılık ve RMSE değerlerinin artmaktadır. Ancak, ilişkili alt testler içeren PISA verilerinden yola çıkarak BBT uygulamasının başlangıç değerleri belirlendiğinde ise yanlılık ve RMSE değerlerinin referans model olan modelle (cevap değişikliğine izin verilmeyen ve başlangıç kuralı aralık olan model) benzer sonuçlar verdiği Tablo 2'de görülmektedir. Cevap değişikliğine izin verilen modelde son yetenek kestirimi olarak önsel bilgilerin de kullanıldığı EAP yetenek kestirim ile elde edilen değerlerin hata miktarı en azdır. ML ve MLEF yöntemlerinde ise bayes yöntemlerine nazaran hata miktarı daha fazladır. Alt testler arasında yüksek korelasyon içeren test bataryalarında regresyon eşitliği ile bireyin başlangıç değerini tanımlamak bireyin yetenek düzeyine uygun maddelerle karşılaşma olasılığı artırmaktadır. Böylece birey cevap değişikliği gerçekleştirse bile o madde birey hakkında bilgi verme düzeyi referans modele göre daha fazla olmaktadır. Ayrıca, regresyon eşitliği ile BBT uygulamasında başlangıç değerini tanımlamak Wainer stratejisine karşı direnç gösterdiği söylenebilir. Bir başka deyişle, birey Wainer stratejisini kullanmak istese bile yetenek düzeyine yakın maddelerle karşılaşacağı için bütün maddeleri yanlış cevaplama olasılığı ve ML yöntemi ile en yüksek puanı elde etme olasılığı da azalmaktadır. Her ne kadar önsel bilgilerin kullanıldığı EAP yöntemi en az hataya sahip olsa da yanlış tanımlanan ön bilgiler, test sürecinde uygulanan maddelerden elde edilen bilgileri baskılamakta ve yanlı sonuçlar elde edilmesine sebep olabilmektedir. Bu nedenle testin amacına göre ön bilgilerin kullanılıp kullanılmayacağına karar verilmesi gerekmektedir. Eğer test bataryasının amacı ulusal düzeyde sertifika ve sıralama amacı gütmeyip izleme veya durum tespiti gibi durumlarda başlangıç değeri kullanılabilir ve böylece bireylere cevap değişikliği hakkı verilebilir.

Tablo 2

Yetenek Kestirim Yöntemlerine Göre Yanlılık ve RMSE Değerleri

Koşullar		Yanlılık (Bias)				RMSE			
Cevap Değişikliğine	Başlangıç kuralı	ML	MLEF	EAP N(0,1)	EAP N(μ, σ)	ML	MLEF	EAP N(0,1)	EAP N(μ, σ)
İzin verilmediğinde	Aralık	0.00	0.00	0.00	0.00	0.20	0.20	0.18	0.18
	Regresyon	0.00	0.00	0.00	0.00	0.23	0.21	0.19	0.17
İzin verildiğinde	Aralık	0.05	0.04	0.02	0.02	0.46	0.37	0.26	0.26
	Regresyon	0.03	0.02	0.04	0.00	0.43	0.35	0.24	0.20

Kaynaklar

Bowles, R., and Pommerich, M. (2001, April). *An examination of item review on a CAT using the specific information item selection algorithm.* Paper presented at the annual meeting of the National Council of Measurement in Education, Seattle, WA.

Cui, Z., Liu, C., He, Y., & Chen, H. (2018). Evaluation of a new method for providing full review opportunities in computerized adaptive testing-computerized adaptive testing with salt. *Journal of Educational Measurement, 55*(4), 582-594. https://doi.org/10.1111/jedm.12193

Gershon, R., & Bergstrom, B. (1995, April). *Does cheating on CAT pay: NOT!* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC document reproduction service No. ED 392 844).

Jensen, M. G. (2017). *Extension of the item pocket method allowing for response review and revision to a computerized adaptive test using the generalized partial credit model* (Doctoral dissertation). The University of Texas at Austin, Texas, United State of America.

Lunz, M. E., Bergstrom, B. A., & Wright, B. D. (1992). The effect of review on student ability and test efficiency for computerized adaptive tests. *Applied Psychological Measurement, 16*, 41-51.

Stocking, M. L. (1997). Revising item responses in computerized adaptive tests: A comparison of three models. *Applied Psychological Measurement, 21*, 129-142.

Stout, W., Ackerman, T., Bolt, D., Froelich, A. G., and Heck, D. (2003). On the Use of Collateral Item Response Information to Improve Pretest Item Calibration. LSAC Research Report Series.

Vispoel, W. P., Clough, S. J., Bleiler, T., Hendrickson, A. B., and Ihrig, D. (2002). Can examinees use judgments of item difficulty to improve proficiency estimates on computerized adaptive vocabulary tests? *Journal of Educational Measurement, 39*, 311–330.

Vispoel, W. P., Henderickson, A. B., and Bleiler, T. (2000). Limiting answer review and change on computerized adaptive vocabulary test: Psychometric and attitudinal results. *Journal of Educational Measurement, 37*, 21-38.

Vispoel, W. P., Rocklin, T. R., Wang, R., & Bleiler, T. (1999). Can examinee use a review option to obtain positively biased ability estimates on a computerized adaptive test? *Journal of Educational Measurement, 36*, 141-157.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice, 15*, 15–20.

Türkçe öğretmen adaylarının farklı test maddesi yazma yeterliklerinin Rasch analiziyle incelenmesi

Ayfer Sayın ve Mehmet Şata

Anahtar kelimeler: Test maddesi, puanlayıcı nitelikleri, çok yüzeyli Rasch, geçerlik, güvenirlik

Giriş

Öğretim programlarında yer alan kazanımlara öğrencilerin ulaşabilmesinde öğretmen niteliğinin önemli bir yeri olduğu bilinmektedir. Öğrencilerin programdaki kazanımlara ne düzeyde ulaştıklarının geçerli ve güvenilir araçlarla belirlenmesi, öğrenciler hakkında karar alırken değerlendirildiği kadar öğrencilere etkili geri bildirim verme konusunda da önemlidir. Bu araştırmada Türkçe öğretmen adaylarının okuduğunu anlama becerilerinin ölçülmesine yönelik oluşturmuş oldukları farklı madde türlerindeki yeterliklerinin incelenmesi amaçlanmıştır. Araştırma kapsamında ayrıca öğretmen adaylarının oluşturmuş oldukları maddeleri değerlendirilen puanlayıcıların niteliklerinin puanlamaya olan etkisi de incelenmiştir.

Performans değerlendirme çalışmalarında puanlayıcı kaynaklı hata kaynaklarının sıklıkla ölçümlere karıştığı raporlanmıştır. Puanlayıcı niteliklerinin de öğrenci performansının değerlendirilmesinde önemli bir faktör olduğu görülmüştür. Örnek olarak Du ve diğ. (1996), MFRM aracılığıyla üniversite öğrencileriyle yaptıkları çalışmada bazı puanlayıcıların konu türüne göre farklılaşan puanlayıcı yanlılığı gösterdiklerini ve bu durumunda öğrencilerin yazma becerisi puanlarını etkilediğini belirlemişlerdir. Benzer şekilde farklı backgrounddaki puanlayıcılar (notive/nonnative) puanlama anahtarının gramer ölçütüne göre yanlılık göstermektedirler buna göre gramer ölçütü genellikle puanlayıcılar tarafından (örneğin notive/non native) en yanlı şekilde puanlanan ölçüttür (Eckes, 2005, 2008; Kondo-Brown, 2002; Schafer, 2008). Bu araştırmada ise öğretmen adaylarının hazırladıkları farklı madde türlerinin alan uzmanlığı ve ölçme değerlendirme uzmanlığı açısından farklılık gösterme durumu incelenmiştir. Ayrıca öğretmen adaylarının farklı madde yazma yeterliklerinin de incelenmesi amaçlanmıştır. Bu amaç doğrultusunda aşağıdaki araştırma sorularına yanıt aranmıştır:

1. Puanlayıcı nitelikleri öğretmen adaylarının farklı test maddesi yazma yeterliklerini değerlendirilmesi sürecinde farklılık göstermekte midir?
2. Türkçe öğretmen adaylarının farklı test maddesi yazma yeterlikleri farklılık göstermekte midir?

3. Puanlayıcının nitelikleri ile öğretmen adaylarının farklı test maddesi yazma yeterlikleri arasında nasıl bir etkileşim bulunmaktadır?

Yöntem

Bu araştırmada nicel araştırma yöntemlerinden ilişkisel tarama deseni kullanılmıştır. İlişkisel tarama modelinde amaç, iki veya daha fazla değişken arasındaki ilişkinin varlığının ve derecesinin (Karasar, 2009) herhangi bir müdahalede bulunulmadan (Büyüköztürk ve diğ., 2018) incelemektir. Araştırmada farklı madde türlerini (çoktan seçmeli, açık uçlu, kısa cevaplı ve doğru yanlış) içeren test geliştiren 84 Türkçe öğretmen adayı bulunmaktadır. Öğretmen adayları 10 hafta süren ölçme ve değerlendirme dersini aldıktan sonra madde hazırlama sürecine geçmiştir. Öğretmen adayları tarafından geliştirilen testlerdeki maddeler farklı niteliklere sahip üç puanlayıcı tarafından puanlanmıştır. Puanlayıcılardan biri ölçme ve değerlendirme alan uzmanı (Rater3), biri Türkçe alan uzmanı (Rater2) son puanlayıcı ise hem Türkçe hem de ölçme ve değerlendirme alan uzmanıdır (Rater1).

Öğretmen adaylarına ölçme ve değerlendirme dersinin test geliştirme konusu kapsamında yüz yüze verilen 12 saatlik eğitim sonrasında öncelikle öğretmen adayları, Türkçe öğretim programında okuduğunu anlama becerisindeki kazanımlarla belirtke tablolarını oluşturmuşlardır. Belirtke tablosu hazırlandıktan sonra öğretmen adaylarından doğru-yanlış, kısa cevaplı, çoktan seçmeli ve açık uçlu maddelerle ölçülmesi planlanan kazanımları yazmaları istenmiştir. Okuduğunu anlama becerisi, özelliğinin doğası gereği kullanılan metin türüne bağlı olarak şekillenmektedir. Metnin uzunluğu, anlatım özelliği, ifadeleri vb. oluşturulacak maddenin türünü ve düzeyini belirlemede doğrudan bir etkiye sahip olmaktadır (Sayın ve Takıl, 2021). Metin seçiminde dikkat edilecek hususlar doğrultusunda seçilen ya da yazılan ve düzenlenen metinlerden ilgili kazanımlar doğrultusunda maddeler, madde yazım formuna yazılmıştır. Formda ilgili kazanım, metin, soru yönergesi, maddeler ve cevap anahtarı olacak şekilde beş bölüm bulunmaktadır. Formun girişinde ayrıca her bir madde türünde soru yazım ilkeleri doğrultusunda dikkat edilecek hususlar yer almaktadır. Öğretmen adayları 5 doğru-yanlış, 5 kısa cevaplı, 3 çoktan seçmeli ve 1 açık uçlu olacak şekilde toplam 14 madde oluşturmuşlardır. Araştırma kapsamında Türkçe öğretmen adayları tarafından farklı madde türlerinden oluşturulan test, araştırmacılar tarafından her bir madde türü için geliştirilen holistik dereceli puanlama anahtarı kullanılarak puanlanmıştır.

Araştırma kapsamında öğretmen adaylarının soru yazma yeterliklerinin değerlendirilmesi amaçlandığından araştırmanın doğasına uygun olan çok yüzeyli Rasch analizi kullanılmıştır (Linacre, 2012). Bu araştırmada puanlayıcılar öğretmen adayları ve madde türü olmak üzere üç yüzey (değişkenlik kaynağı) bulunmaktadır. Araştırmadaki tüm değişkenlik kaynakları dikkate alınmış ve tüm puanlayıcıların tüm öğrencileri tüm madde yazma türleri ile değerlendirdiği tamamen çaprazlanmış desen kullanılmıştır. Veri analizi yapılırken Myford ve Wolfe (2003, 2004) tarafından belirtilen yönergeler dikkate alınmıştır.

Sonuçlar

Türkçe öğretmen adaylarının farklı madde türü yazma yeterliklerinin değerlendirilmesinin amaçlandığı bu çalışmada ilk olarak puanlayıcı niteliklerinin yapılan değerlendirme üzerindeki etkisi belirlenmeye çalışılmıştır. Yapılan analizler neticesinde puanlayıcı niteliğinin öğrenci performansını değerlendirilmesi sürecinde istatistiksel olarak farklılık oluşturduğu tespit edilmiştir. Puanlayıcı nitelikleri incelendiğinde hem Türkçe hem de ölçme ve değerlendirme alanında uzmanlığı bulunan puanlayıcının en güvenilir puanlamayı yaptığı görülmektedir. Daha sonra Türkçe eğitimi alan uzmanının güvenilir puanlama yaptığı; puanlayıcılar arasında en az güvenilir puanlamanın ise ölçme ve değerlendirme uzmanı tarafından gerçekleştirildiği belirlenmiştir. Öğretmen adaylarının farklı madde yazma yeterlikleri incelendiğinde ise Türkçe öğretmen adaylarının doğru-yanlış maddelerini hazırlama yeterliklerinin yüksek, çoktan seçmeli madde yazma yeterliklerinin de düşük olduğu sonucuna ulaşılmıştır.

Puanlayıcıların farklı madde türlerini değerlendirmesi sürecinde en fazla hangi madde türünde daha fazla beklenmedik puanları verdiklerini belirlemek amacıyla artık değerlerin standartlaştırılmış hâli incelenmiştir. Yapılan analizler neticesinde 51 tane uç değerin olduğu ve birinci puanlayıcının 11 (%21.57), ikinci puanlayıcının 19 (%37.25) ve üçüncü puanlayıcının ise 21 (%41.18) tane uç değere sahip olduğu bulunmuştur. Uç değerlerin daha çok hangi madde türünde ortaya çıktığı incelendiğinde ise çoktan seçmeli madde türünde 6 (%11.76), açık uçlu maddelerde 7 (%13.73), doğru yanlış maddelerde 8 (%15.69) ve kısa cevaplı maddelerde ise 30 (%58.82) tane olduğu tespit edilmiştir. Buna göre en fazla uç değerlerin görüldüğü madde türü olan kısa cevaplıda puanlayıcıların en az uyuşma gösterdiği veya puanlayıcı niteliklerinin en fazla etkilediği madde türü olduğu söylenebilir.

Kaynaklar

Büyüköztürk, Ş., Kılıç-Çakmak, E., Akgün, Ö., Karadeniz, Ş. ve Demirel, F. (2008). *Bilimsel araştırma yöntemleri.* Pegem Akademi.

Du, Y.,Wright, B. D., & Brown, W. L. (1996, April). *Differential facet functioning detection in direct writing assessment.* Paper presented at the Annual Meeting of the American Educational Research Association, New York, USA.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2,* 197-221. https://doi.org/10.1207/s15434311laq0203_2

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*, 155-185. doi:10.1177/0265532207086780

Karasar, N. (2012). *Bilimsel araştırma yöntemi.* Nobel Akademik Yayıncılık.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(1), 3-31.

Sayın, A. ve Takıl, B. (2021). Türkçe öğretmen adaylarının sosyal ve akademik yeterlik algılarının incelenmesi. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 17*(2), 868-883.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465-493.

Açımlayıcı faktör analizi ve madde tepki kuramına göre seçilen istatistik tutum ölçeği maddelerine uygulanan doğrulayıcı faktör analizi sonuçlarının karşılaştırılması[1]

Sinan Muhammet Bekmezci ve Nuri Doğan

Anahtar kelimeler: Yapı geçerliği, açımlayıcı faktör analizi, madde tepki kuramı, dereceleme ölçeği modeli, doğrulayıcı faktör analizi

Giriş

Ölçme ve değerlendirme işlemlerinden anlamlı sonuçlar elde edebilmek için kullanılan veri toplama araçlarından elde edilen sonuçların geçerlik ve güvenilirlik gibi psikometrik özelliklerin standartlara uygun olmasını sağlamak gerekmektedir. Önemli bir özellik olan geçerlik en klasik tanımıyla, bir testin ölçülmesi istenilen özelliği başka özelliklerle karıştırmadan ölçebilme derecesi olarak ele alınmaktadır (Ebel ve Frisbie, 1991). Literatürde sıklıkla kullanılan 4 tip geçerlik türü bulunmaktadır bunlar yordama geçerliği, uyum geçerliği, kapsam geçerliği ve yapı geçerliğidir. Test geliştirme sürecinde yapı geçerliğini etkileyebilen madde ayırt ediciliği, madde güçlüğü ve maddeler arası ilişkiler söz konusudur. Teste ve maddeye ilişkin istatistiksel sonuçları ortaya koymak için klasik test kuramı ve madde tepki kuramı başta olmak üzere farklı kuramlar temel alınarak çeşitli analizler yapılabilmektedir.

Bu konudaki alanyazın incelendiğinde belirli yaklaşımların kullanıldığı çalışmalara rastlanmakla birlikte karşılaştırma çalışmalarının daha az olduğu görülmektedir. Örneğin, Krishnan ve Idris (2018), "Bir Ölçme Aracının Kalitesini Artırmak için Kısmi Puan Modeli'nin Kullanımı" çalışmasında, örtük özelliği ölçme sürecinde testteki maddelerin birlikte iyi işleyip işlemediklerini kısmi puan modelini kullanarak belirlemeye çalışmışlardır. Karlin ve Karlin (2018)'in, Rasch modelinin kullanımını açıklamak ve birkaç maddesi aynı olan iki farklı sınavın geçerliğini incelemek amacıyla yaptıkları çalışmada, Rasch modelinin testlerin geçerliği ortaya koymada olduğu kadar, öğrencilerin değerlendirilmesine de öncelik sağladığını belirtmişlerdir. Wei ve diğ. (2014) ise, öğrencilerin bilim modellerini anlama düzeylerini ölçme amacıyla KTK ile geliştirilen ölçme aracını KTK' nın sınırlılıklarını düşünerek, Rasch modeli ile incelemişlerdir ve aracın güvenirliği ve geçerliği için kanıt sunmanın yanında aracın iyileştirilebilmesi için bazı yönlere vurgu yapmışlardır. Bununla beraber İlhan ve Güler (2017), çalışmalarında likert tipi

[1] Çalışma "Psikometrik Özelliklerin Belirlenmesinde Veri Madenciliği, Yapay Sinir Ağları ve Faktör Analizi Sonuçlarının Karşılaştırılması" adlı tezin bir parçasıdır.

ölçeklerde KTK ve Rasch analizinden elde edilen yetenek kestirimlerini göreli uyum, mutlak uyum ve dağılım özellikleri açısından karşılaştırdıkları çalışmalarında hesaplanan yetenek ölçüleri arasında yüksek korelasyon (.95, $p< .05$) ve iki kuramla kestirilen yetenek ölçüleri arasında yüksek bir göreli uyum olduğunu, iki kurama göre kestirilen yetenek puanları arasında mutlak bir uyum bulunmadığını, Rasch analizinde kestirilen yetenek puanlarının KTK'dan elde edilen yetenek puanlarına kıyasla daha simetrik ve daha sivri bir dağılıma sahip olduğunu bulmuşlardır. İlhan ve Güler (2018)'de, açık uçlu maddelerde KTK ile çok yüzeyli Rasch modeline (ÇYRM) göre hesaplanan güçlük indekslerinin karşılaştırılma amacıyla yaptıkları çalışmada iki kurama göre kestirilen güçlük indeksleri arasında pozitif yönde, güçlü ve anlamlı bir korelasyonun ($r= .999$, $p< .001$) bulunduğu, her iki kurama göre de başarı testindeki 10 maddenin kolaydan zora doğru ayrı şekilde sıralandığını belirlemişlerdir. İlhan (2016)' da, açık uçlu sorularla yapılan ölçmelerde KTK ve ÇYRM göre hesaplanan yetenek kestirimlerinin karşılaştırdığında hesaplanan yetenek kestirimleri arasındaki göreli uyumun son derece yüksek olduğu, iki kurama göre hesaplanan yetenek kestirimlerine ait ortalamalar arasında anlamlı fark bulunduğu ve dolayısıyla mutlak bir uyumun söz konusu olmadığını; ayrıca, ÇYRM'de rapor edilen yetenek kestirimlerinin ölçüt geçerliğinin KTK' dan elde edilen yetenek kestirimlerine kıyasla daha yüksek olduğunu belirtmiştir, Uysal (2015)'de, "Araştırma Özyeterlik Ölçeğinin Psikometrik Özelliklerinin Klasik Test Kuramı ve Madde Tepki Kuramına Göre İncelenmesi" isimli yüksek lisans tez çalışmasında, araştırma özyeterlik ölçeğinin madde (madde ayırıcılık gücü, maddenin ölçtüğü özellik düzeyi) ve ölçek (ölçek puanı, güvenirlik ve geçerlik) özelliklerini KTK ve MTK' ya göre incelemiş, madde ayırıcılık güçleri arasında pozitif yönde yüksek düzeyde bir ilişki, ölçülen özellik düzeyleri açısından iki kuram arasında negatif yönde yüksek düzeyde bir ilişki bulunmuş ve ölçeğin iki kurama göre de kestirilen güvenirlik katsayılarının benzer olduğu ve ölçek puanlarının güvenilir olduğunu, belirtmişlerdir.

İlgili çalışmalar incelendiğinde, Rasch model başta olmak üzere farklı MTK modellerinin klasik yaklaşımlara alternatif olarak ele alındığı görülmektedir. Çalışmalar madde güçlük parametrelerinin, madde ayırıcılık parametrelerinin karşılaştırılmasına, farklı test ve madde tiplerinde güvenirlik-geçerlik çalışmalarının yapılmasına odaklanmakla birlikte, özellikle Türkiye'de likert tipi ölçeklerin geçerlik analizlerinin yapılmasında MTK modellerinin kullanıldığı yeterince çalışma görülmemiştir. Bu nedenle çalışmada MTK kapsamındaki dereceleme ölçeği modelinin kullanışlılığının incelenmesi amaçlanmaktadır.

Yöntem

Bu çalışma, açımlayıcı faktör analizi ve MTK modellerinden dereceleme ölçeği modeli ile tespit edilen yapı geçerliği kanıtlarının birbirleriyle farklarını, benzerliklerini ortaya çıkarmayı amaçlanmaktadır. Bu araştırma farklı tekniklerden elde edilecek ölçek yapılarının ortaya konulması açısından betimsel (Grove, Burns ve Gray, 2012), farklı tekniklerden elde edilen ölçek yapılarına ait doğrulayıcı faktör analizi uyum indeslerinin karşılaştırılması açısından ilişkisel bir araştırmadır (Büyüköztürk ve diğ., 2017).

Çalışmanın amacı tutum ölçeğine farklı tekniklerle madde seçmenin testin psikometrik sonuçlarına etkisini araştırmak olduğundan araştırma bir çalışma grubu üzerinden yürütülmüştür. Çalışma grubu istatistik dersi almış veya almamış olan lisans, yüksek lisans düzeyindeki üniversite öğrencilerinden oluşmaktadır. Çalışma grubuna istatistik tutum ölçeği pandemi koşullarından dolayı çevrimiçi olarak uygulanmıştır. Çevirimiçi ulaşılan katılımcı sayısı 808 kişidir ve ölçme aracındaki soru sayısı düşünüldüğünde yeterli büyüklükte olduğu söylenebilir. Katılımcılar demografik değişkenlere göre incelendiğinde İstatistik Tutum Ölçeğinin uygulandığı grubun %74'ü kadın, %26'sı erkektir. Katılmcıların %68'i daha önce bir istatistik dersi aldığını, %32'si ise istatistik dersi almadığını belirtmiştir. Katılımcıların "Akademik alanınızı aşağıdakilerden hangisi daha iyi tanımlar?" sorusuna %36'sının sayısal, %43'ünün eşit ağırlık, %18'inin sözel, %3'ünün dil cevabı verdiği görülmüştür.

İstatistiğe yönelik tutum düzeyini belirlemek amaçlı hazırlanan bu ölçme aracı 40 sorudan oluşan bir 'İstatistik Tutum Ölçeği' deneme formudur. Ölçme aracındaki maddelerden 20 tanesi olumsuz olduğundan bu maddelerde ters puanlama yapılmıştır. Ölçek içerisindeki maddelerden 9 ile 40. maddeler birbirinin aynısı, 1 ve 31. maddeler ise birbirinin tam zıddı ifadeye sahiptir. Bu maddeler yanıtlayıcıların samimi yanıt verip vermediklerini kontrol etmek amaçlı kullanılmıştır ve maddelerle ilgili yapılan işlemler bulgular kısmında açıklanmıştır.

Araştırmada ilk aşamada ölçeğin faktör yapısını ortaya koymak ve boyutluluk üzerine bulguları ortaya koyabilmek için polikorik korelasyon matrisine ve temel eksenler kestirimine dayanan açımlayıcı faktör analizi (AFA) yapılmıştır. AFA sonucunda boyut sayısına karar vermek için paralel analiz tekniği sonuçlarından yararlanılmıştır. AFA çerçevesinde madde faktör yükü, açıklanan varyans oranı, ve boyutluluk yakınsama indeksi incelenerek ölçeğin psikometrik özellikleri belirlenmeye çalışılmıştır.

MTK analizleri kapsamında sıralı çok kategorili modeller için kullanılabilen Masters (1982) tarafından geliştirilen "Kısmi Puanlama Modeli" ile David Andrich (Andrich, 1978a, 1978b, 1978c) ve Earling Andersen (Andersen, 1977) tarafından geliştirilen "Dereceleme Ölçeği Modeli" kullanılarak analizler yapılmıştır. MTK kapsamındaki bu analizlerden model veri uyumunun daha iyi olduğu belirlenen dereceleme ölçeği modeli sonuçları AFA sonuçları ile karşılaştırmada kullanılmıştır.

Sonuçlar

İki kurama göre ölçek uygulamasının güvenirlik ve geçerlikleri açısından problemli olabilecek maddeler incelenmiş ve madde seçimi yapılmaya çalışılmıştır. Açımlayıcı faktör analizi ve dereceleme ölçeği modeli sonuçları dikkate alınarak ölçeğin tek boyutlu bir yapıya sahip ölçek elde edilmeye çalışılmıştır. Açımlayıcı faktör analizine göre tek boyutlu 32 maddelik bir nihai ölçek yapısı elde edilirken, dereceleme ölçeği modelinden madde-veri uyumuna dayanarak tek boyutlu yapıya sahip 25 maddeli bir nihai ölçek oluşturulmuştur. İki kurama göre elde edilen ölçekler için hesaplanan Cronbach Alpha iç tutarlık güvenirlik değerlerinin de oldukça yüksek bir değer olduğu görülmüştür. İki nihai ölçek yapısına göre yapılan doğrulayıcı faktör analizi uyum indeksleri sonuçlarına göre de iki tekniğe göre elde edilen formların uyun değelerinin yüksek ve uyumsuzluk (hata) değerlerinin küçük olduğu görülmüştür.

Kaynaklar

Grove, S. K., Burns, N., and Gray, J. R. (2012). *The practice of nursing research: Appraisal, synthesis, and generation of evidence* (7th ed.). Elsevier.

Büyüköztürk, Ş., Çakmak, E. K., Akgün, Ö. E., Karadeniz, Ş. ve Demirel, F. (2017). *Bilimsel araştırma yöntemleri.* Pegem Akademi

Ebel, R. L., & Frisbie, D.A. (1991) *Essentials of educational measurement.* Prentice – Hall International, Inc.

İlhan, M. (2016). Açık uçlu sorularla yapılan ölçmelerde klasik test kuramı ve çok yüzeyli rasch modeline göre hesaplanan yetenek kestirimlerinin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 31*(2), 346-368. https://doi.org/10.16986/HUJE.2016015182

İlhan, M. ve Güler, N. (2017). Likert tipi ölçeklerde klasik test kuramı ile rasch analizinden elde edilen yetenek kestirimleri arasındaki uyumun test edilmesi. *Ege Eğitim Dergisi, 18*(1), 244-265. https://doi.org/10.12984/egeefd.289576

İlhan, M., and Güler, N. (2018). A comparison of difficulty ındices calculated for openended ıtems according to classical test theory and many facet rasch model. *Eurasian Journal of Educational Research, 75*, 99-114. https://doi.org/10.14689/ejer.2018.75.6

Karlin, O., and Karlin S. (2018). Making better tests with the rasch measurement model. *InSight: A Journal of Scholarly Teaching, 13*, 76-100. https://files.eric.ed.gov/fulltext/EJ1184946.pdf

Krishnan, S., and Idris, N. (2018). Using partial credit model to improve the quality of an instrument. *International Journal of Evaluation and Research in Education*, *7*(4), 313-316. http://doi.org/10.11591/ijere.v7i4.15146

Wei, S., Liu, X., and Jia, Y., (2014). Using rasch measurement to validate the instrument of students' understanding of models in science (SUMS). *International Journal of Science and Mathematics, 12*, 1067-1082. https://doi.org/10.1007/s10763-013-9459-z

Öğretmenlerin eğitimde teknoloji kullanımına yönelik tutumları ile teknolojiyi kullanma becerilerinin incelenmesi

Özge Özbek

Anahtar kelimeler: Beceriler, beceri ölçeği, tutum, tutum ölçeği, öğretmenlerin teknoloji kullanımı, öz –yeterlik inancı,

Giriş

Öğretmenler, eğitim programlarının uygulayıcıları olduklarından eğitim sisteminin ve eğitim hizmetlerinin kalitesini belirleme ve sistemin başarılı olup olmaması ile ilgili sorumluluk sahibidirler (Mahiroğlu, 2009). Öğretmenlerin bu aşamada büyük bir rol oynadığı söylenebilir.

Öğretmenlerden beklenen, yeniçağda gelişmekte olan teknolojiyi derslerinde kullanabilir hâle getirebilmeleridir. Yapılan araştırmalar sonucu bilgisayarın öğretimde etkili kullanılmasını öğretmenlerin bilgisayarı mesleki anlamda faydalı bulup bulmamaları, öğretimde kullanmaları durumunda öz-yeterlilik algılarıyla ilgili becerilerinin etkisi yer almaktadır. Teknoloji bilgisi açısından öğretmenlerin nitelikli açıdan yetişememesi, çağı yakalayamayan öğretmenlerin öğrencilere ulaşamaması, teknolojiye karşı olumsuz tutumları başlıca sorunlar arasındadır.

Bu sorunlar çerçevesinde öğretmenlerin teknolojiyi kullanma becerileri ve teknolojiye yönelik tutumları arasındaki ilişkinin incelenmesi amaçlanmıştır. Bu amaç doğrultusunda oluşturulan alt problemler aşağıdaki gibidir:

1. Öğretmenlerin teknolojiyi kullanma becerileri ve tutumları ne düzeydedir?
2. Öğretmenlerin teknolojiyi kullanma becerileri ile teknolojiyi kullanımlarına yönelik tutumları arasında anlamlı bir ilişki var mıdır?
3. Öğretmenlerin teknolojiyi kullanma becerileri ve teknolojiyi kullanımlarına yönelik tutumları cinsiyetlerine göre farklılık göstermekte midir?
4. Öğretmenlerin teknolojiyi kullanma becerileri ve teknolojiyi kullanımlarına yönelik tutumları mezun olunan fakülteye göre farklılık göstermekte midir?
5. Öğretmenlerin teknolojiyi kullanma becerileri ve teknolojiyi kullanımlarına yönelik tutumları deneyimlerine göre farklılık göstermekte midir?

6. Öğretmenlerin teknolojiyi kullanma becerileri ve teknolojiyi kullanımlarına yönelik tutumları çalıştıkları okul türüne göre farklılık göstermekte midir?

Araştırmanın örneklemi basit tesadüfi örnekleme yöntemi ile belirlenmiştir. Türkiye genelinde çoğu Ankara ilinde çalışan farklı branşlardaki öğretmenler tarafından oluşturulmuştur. Uygulama sürecinde demografik özelliklerin yer aldığı anketler sonucu ölçekleri cevaplayan öğretmenlerin 233'ü (%77.9) kadın, 66'sı (%22.1) erkek olmak üzere toplam 299 öğretmen ile yapılmıştır. Araştırmada önceden tarafımca çalışılmış olan öğretmenlerin demografik özelliklerinin ve eğitimde bilişim teknolojilerine yönelik tutumlarının ölçüldüğü ölçeğe ait bilgiler kullanılmış olup yenileme çalışmasıyla birlikte geçerlik ve güvenirliği tekrardan incelenmiştir. Tutumların ölçülmesinin dışında alan taraması yapılarak hazırlanan bu tutumlarla ilişkili olabilecek "Eğitimde Teknolojiyi Kullanma Becerileri Ölçeği de(ETKBÖ)" bu çalışmada yer almıştır. Ölçeklere ait yenileme çalışması yapıldıktan sonra son hâlleriyle birlikte uzman görüşü alınmış olup bu görüşler doğrultusunda düzeltmeler yapılmış, uygulama sürecine geçilmiştir. Öğretmenlerin Teknoloji Kullanımına Yönelik Tutum Ölçeği (ÖTKYTÖ) 27 madde ile oluşturulmuştur. Bu maddelerden 9'u olumsuz ve 18'i olumlu maddedir. Her bir madde 1-5 arası puanlanmış olup ETKBÖ beşli likert tipi ölçek olarak hazırlanmıştır. Örneklemimizden toplanan veriler SPSS 20'ye girildikten sonra yapılan analizler sonucu ayırt ediciliği düşük maddeler çıkarılmış olup ölçeğimiz 7'si olumsuz, 18'i olumlu maddeler üzerinden analizler gerçekleştirilmiştir. Ayrıca güvenirlik analizi yapılmış olup Cronbach alfa değeri .89'dur. KMO değeri ve Barlett testi yapılarak faktör analizi yapılması uygun görülmüştür. Faktör analizi ile birlikte Varimax döndürme yapılarak boyutlar incelenmiş olup ölçeğimizin tek boyutlu olduğuna karar verilmiş ve maddelerin korelasyon değerlerine bakılarak madde sayısı 22'ye düşürülmüştür. Ölçeğin son hâli ile Cronbach alfa değeri .91 çıkmıştır. ETKBÖ ilgili alan taraması yapıldıktan sonra 15 maddeden oluşacak şekilde öğretmenlerde olması beklenen olası becerileri içermektedir. Ölçek dörtlü likert tipi ölçek olarak hazırlanmıştır. Güvenirlik analizi sonucu Cronbach alfa değeri .91'dir. KMO değeri ve Barlett testi yapılarak faktör analizi yapılması uygun görülmüştür. Faktör analizi ile birlikte Equamax döndürme yapılarak boyutlar incelenmiş olup ölçeğimizin iki boyutlu olduğuna karar verilmiştir. Ölçeğe ait boyutlar "genel beceriler" ve "yeni nesil teknolojik beceriler" olarak yapılandırılmıştır.

Öğretmenlerin teknolojiye yönelik tutum ve becerileri öğretmenlerin cinsiyetlerine, çalıştıkları kuruma, meslekteki deneyim yıllarına ve mezun oldukları fakülte çeşitlerine göre ilişkileri bağımsız t-testi ve ANOVA kullanılarak belirlenmiştir. Analizler sonucu öğretmenlerin teknoloji kullanımına yönelik tutumlarının beceriler ile doğru orantılı olduğu, anlamlı düzeyde kestirildiği gözlenmiştir. Her ne kadar beceriler ile tutumlar arasında cinsiyet farkı anlamlı düzeyde çıkmasa da alt boyut olan genel beceriler ve yeni nesil teknoloji becerileri incelendiğinde anlamlı farklılıklar yer aldığı belirlenmiştir. İlgili diğer sonuçlar çalışmamızda yer almakta olup anlamlı farklılıkların görülmediği değişkenler de yer almaktadır.

Likert tipi ölçeklerde seçenek farklılıklarının maddelerin psikometrik özelliklerine etkisi

Nuri Doğan, Meltem Yurtçu ve Ceylan Gündeğer

Anahtar kelimeler: Seçenek farklılığı, psikometrik özellikler, likert tipi ölçekler

Giriş

Ölçme ve değerlendirme birçok fiziki ölçümlerde daha net sonuçlar elde edilmekte ve gerçeğe daha yakın sonuçlara ulaşılmaktadır. Ancak insanı konu alan birçok ölçme de ise gözlemlenemeyen faktörlerin de bir araya gelmesinden kaynaklı olarak ölçülmek istenen özelliğe ulaşılması biraz güçleşmektedir. Bu yüzden ölçme sürecini gerçekleştirirken hataları minimize edecek olan veya amaç doğrultusunda en uygun olan ölçme aracını seçmek oldukça önem arz etmektedir. Bireylerin duyuşsal becerilerinin ölçülmesinde ise sıklıkla Likert tipi ölçeklerden yararlanılmaktadır. Bu ölçeklerde yer alan maddeler önemli olduğu gibi ölçeklerde yer alan kategorilerde önem arz etmektedir. Dolayısı ile bu kategoriler tercih edilirken görünüş geçerliği sağlaması yanında incelenecek konu ile ilgili özelliği ortaya çıkartmada bireylerin algılarını yönlendirmeyecek bir formatta olması yanlı sonuçlar elde edilmemesi için önemli bir adımdır. Dolaylısı ile ölçek geliştirme sürecinde araştırmacıların bu adıma da gereken önemi vermesi gerekmektedir. Böylece geliştirilen ölçeklerin geçerlik ve güvenirlik düzeyleri artırılabilecektir.

Özellikle literatürde yapılan birçok araştırmada farklı kategori derecelerine sahip Likert tipi ölçeklerin orta kategorilerinin birbirlerinden farklılık gösterdiği ve hepsinin aynı anlamda nitelendirilemeyeceği sonucu elde edilmektedir. Bu sonuçtan yola çıkarak bu araştırmada, aynı ölçeğin faklı kategori isimleri kullanılan formlarının bireylerin algılarını nasıl etkilediğini incelemek amaçlanmıştır.

Yöntem

Bir ölçeğin kategorileri seçeneklerine göre farklılaşan formları arasındaki farklılıkları ortaya koymak amaçlandığından araştırma tarama modelinde bir araştırmadır. Bu yönü araştırmanın türü nedensel karşılaştırma araştırmaları olarak belirlenmiştir.

Araştırmanın ulaşılabilir örneklemi Aksaray Üniversitesi, Hacettepe Üniversitesi ve İnönü Üniversitesi'nde öğrenim gören üniversite öğrencileri oluşturmaktadır. Formlardaki kayıp ve eksik veriler

ile birlikte tekrar eden veriler data setinden çıkarıldıktan sonra bütün formları yanıtlamış olan 377 bireye ait veriler üzerinde analizler gerçekleştirilmiştir.

Bu araştırmada, Baykul (1990) tarafından geliştirilen *Matematikle İlgili Düşünceler Ölçeği*'nin kısa formu temel alınmıştır. Orijinal formunda 15 olumlu 15 olumsuz madde bulunan ölçeğe ait güvenirlik katsayısı 0.96 olarak elde edilmiştir (Nartgün, 2002). Ölçekteki maddelerin aynı yapıyı temsil ettiği göz önüne alınarak ölçeğe ait ilk etapta 15 madde ele alınmıştır. Bu maddelerin temel alınan ölçeğe ait matematik ile ilgili olduğu ve öğrencilerin matematiği kullanma becerilerine katkı sağlayacağı düşünülen iki madde eklenmiş ve sadece derse yönelik olarak ifade edilen iki madde ölçek formundan çıkarılmıştır. Nihai hale getirilen ölçek formu dört farklı seçenek kategorisine göre düzenlenmiştir. Bu kategoriler;

Form1 için: 1 Kesinlikle Katılmıyorum, 2-Katılmıyorum, 3-Fikrim Yok, 4- Katılıyorum, 5-Kesinlikle Katılıyorum

Form2 için: 1 Kesinlikle Katılmıyorum, 2-Katılmıyorum, 3-Kararsızım, 4- Katılıyorum, 5-Kesinlikle Katılıyorum

Form3 için: 1-Kesinlikle Katılmıyorum, 2-Katılımıyorum, 3-Ne katılıyorum ne katılmıyorum, 4-Katılıyorum, 5- Kesinlikle Katılıyorum

Form4 için: Kesinlikle Katılmıyorum (1) (2) (3) (4) (5) Kesinlikle Katılıyorum

seçenekleri yer almaktadır. KMO ve Barttlett değerleri dikkate alındığında yapının faktörlenebilir olduğu gözlemlenmektedir. Aynı zamanda faktör analizi sonucunda tüm formlar tek boyutlu bir yapı sergilemektedir. Ölçeğe ait dört forma ait güvenirlikler 0.965 ile 0.969 arasında değişmektedir.

Araştırmada sıralı düzeyde verilmiş olan cevap kategorilerinin madde parametrelerini kestirmek üzere Generelazied Partial Credit Model (GPCM) ve Graded Responce Model (GRM) birlikte kullanılmıştır. Her iki modele göre madde parametreleri kestirildikten sonra parametreler formlara göre karşılaştırılmıştır. Karşılaştırmalar Kruskall Wallis testi ile, ikili karşılaştırmalar ise Mann Whitney U testi ile yapılmıştır.

Verilerin analizi R istatistik programından yürütülmüş (R Development Core Team, 2013) olup mirt (Chalmers, 2012), TAM (Robitzsch, Kiefer ve Wu, 2021). FSA (Ogle, Wheeler ve Dinno, 2021), EFAtools (Steiner ve Grieder, 2020) paketlerinden yararlanılmıştır.

Sonuçlar

Araştırmada kullanılan GPCM ve GPM modellerinin veri ile uyumunu incelemek üzere model veri uyumu incelemesi yapılmıştır. Modellere ait uyum indeksleri Tablo1 te verilmiştir.

Tablo 1

Model Veri Uyumu

Model/Form	"Likelihood"	-2LL	"AIC"	"BIC"
"grmModelForm1"	-5424.388	-166.704	10988.776	11264.033
"gpcmModelForm1"	-5507.740		11155.481	11430.738
"grmModelForm2"	-5355.097	-192.152	10850.194	11125.451
"gpcmModelForm2"	-5451.173		11042.347	11317.604
"grmModelForm3"	-5513.108	42.04	11166.217	11441.474
"gpcmModelForm3"	-5492.088		11124.177	11399.434
"grmModelForm4"	-5932.855	59.278	12005.710	12280.967
"gpcmModelForm4"	-5903.216		11946.433	12221.690

Model veri uyumları incelendiğinde Form1 ve Form 2 için GRM nin model uyumunun daha fazla olduğu sonucu elde edilirken Form3 ve Form 4 için GPCM veri seti ile daha uyumlu olduğu sonucu gözlemlenmektedir. Her iki modele göre dört form için madde parametreleri elde edilmiştir.

GPCM'ye göre kestirilen madde parametreleri dikkate alındığında "a" ayırt edicilik parametresinin Form1 de 0.80 ile 3.92; Form2 de 0.76 ile 4.71; Form3 te 0.87 ile 5.24; Form 4 te 0.63 ile 5.58 değerleri arasında değiştiği gözlemlenmiştir. "b1" sınır parametre değerinin Form1 de -2.06 ile -0.89; Form2 de -1.89 ile -0.8; Form3 te -1.79 ile -0.74; Form 4 te -1.23 ile -0.64 değerleri arasında değiştiği gözlemlenmiştir. "b2" sınır parametre değerinin Form1 de -0.50 ile 0.53; Form2 de -0.65 ile 0.58; Form3 te -0.75 ile 0.17; Form 4 te -0.82 ile 0.27 değerleri arasında değiştiği; "b3" sınır parametre değerinin Form1 de -1.01 ile 0.21; Form2 de -1.11 ile 0.39; Form3 te -0.88 ile 0.44; Form 4 te -0.80 ile 0.33 değerleri arasında değiştiği; "b4" sınır parametre değerinin ise Form1 de 0.31 ile 1.50; Form2 de 0.42 ile 1.42; Form3 te 0.43 ile 1.47; Form 4 te -0.08 ile 1.2 değerleri arasında değiştiği gözlemlenmiştir. GPCM modeline göre marjinal güvenirlikler en küçük değer 0.866 (Form1 için) en yüksek değer ise 0.890 (Form3 için) olarak elde edilmiştir.

GRM ye göre elde edilen madde parametreleri dikkate alındığında "a" ayırt edicilik parametresinin Form1 de 2.09 ile 4.85; Form2 de 1.85 ile 5.38; Form3 te 2.00 ile 5.95; Form 4 te 1.66 ile 6.35 değerleri arasında değiştiği gözlemlenmiştir. Sınır parametre değerleri dikkate alındığında ise "b1" sınır parametre değerinin Form1 de -2.02 ile -0.91; Form2 de -1.91 ile -0.85; Form3 te -1.9 ile -0.81; Form 4 te -1.51 ile -0.77 değerleri arasında değiştiği gözlemlenmiştir. "b2" sınır parametre değerinin Form1 de -0.90 ile 0.15; Form2 de -0.86 ile 0.06; Form3 te -0.91 ile -0.01; Form 4 te -0.89 ile -0.05 değerleri arasında değiştiği; "b3" sınır parametre değerinin Form1 de -0.48 ile 0.48; Form2 de -0.45 ile 0.5; Form3 te -0.39 ile 0.51; Form 4 te -0.44 ile 0.43 değerleri arasında değiştiği; "b4" sınır parametre değerinin ise Form1 de 0.39 ile 1.45; Form2 de 0.47 ile 1.4; Form3 te 0.48 ile 1.46; Form 4 te 0.25 ile 1.22 değerleri arasında değiştiği gözlemlenmiştir. GRM modeline ait olarak testlerin marjinal güvenirlikleri 0.934 (Form1 için) ile 0.939 (Form3 için) aralığında değişmektedir.

Modellere göre kestirilmiş olan aynı parametrelerin formlara göre farklılaşıp farklılaşmadığı Kruskall Wallis testi ile incelenmiştir.

Tablo 2

GPCM ve GRM Modelinde Aynı Parametrelerin Formlara Göre Farklılaşması

	GPCM			GRM		
	χ^2	sd	p	χ^2	sd	p
Formlardaki a değerleri	2.72	3	.437	1.18	3	.759
Formlardaki b1değerleri	14.90	3	.002*	5.12	3	.163
Formlardaki b2 değerleri	4.05	3	.256	0.13	3	.988
Formlardaki b3 değerleri	3.12	3	.373	0.70	3	.873
Formlardaki b4 değerleri	11.44	3	.010*	8.23	3	.042*

GPCM e göre b1 ve b4 parametresinin; GRM e göre b4 parametresinin formlara göre farklılık gösterdiği sonucu elde edilmiştir. Bu farklılığın hangi formlar üzerinde olduğunu incelemek için Mann Whitney U testinden yararlanılmıştır.

Tablo 4

Fark Bulunan Parametrelerin Formlara Göre Karşılaştırılması

	GPCM			GRM	
	Comparison	Z	P		
b1	Form1 - Form2	-0.02317449	0.9815110882		
	Form1 - Form3	-0.93856679	0.3479532084		
	Form2 - Form3	-0.91539230	0.3599857166		
	Form1 - Form4	-3.34871362	0.0008118767*		
	Form2 - Form4	-3.32553913	0.0008824771*		
	Form3 - Form4	-2.41014683	0.0159461016*		
b4	Form1 - Form2	-0.3012684	0.763209871	-0.24333213	0.80774811
	Form1 - Form3	-0.1622214	0.871131492	-0.18539591	0.85291852
	Form2 - Form3	0.1390469	0.889413063	0. 05793622	0.95379943
	Form1 - Form4	2.5955427	0.009444166*	2.18998919	0.02852502 *
	Form2 - Form4	2.8968111	0.003769767*	2.43332132	0.01496102 *
	Form3 - Form4	2.7577642	0.005819817*	2.37538510	0.01753065 *

Formlara göre b1 ve b4 parametresi incelendiğinde 4. form ile diğer formlardaki bu değerlerin farklılaştığı gözlemlenmektedir. b1 parametresi için Form4 te kestirilen katsayının diğer formlardakinden anlamlı düzeyde yüksek olduğu, b4 parametresi dikkate alındığında ise Form4 te kestirilmiş katsayının diğer formlardakinden anlamlı düzeyde düşük olduğu sonucu elde edilmiştir. Formlara göre b4 parametresinin form4 ile diğer formlarda anlamlı düzeyde farklılık gösterdiği sonucu elde edilmiştir. Form4 te kestirilmiş olan b4 katsayının diğer formlardakinden düşük olduğu sonucu elde edilmiştir.

Sonuçlar

Araştırmada Likert tipi ölçeklerde aşamalı olarak verilen cevap kategorileri için uygun olan GPCM ve GRM modelleri teme alınmıştır. Form1 ve Form2 nin GRM modeli ile, Form3 ve Form4 ün ise GPCM modeli ile daha uyumlu oldukları sonucu elde edilmiştir.

İki modelde de Form4 ün diğer formlardan farklı olarak "a" değerinin geniş bir ranja, "b1" değerinin ise dar bir ranja sahip olduğu gözlemlenmiştir. Bu sonuçlar Form4 teki kategorileri öğrencilerin algılarını farklılaştırdığının bir göstergesi olarak ele alınabilir.

Marjinal güvenirlik katsayı her iki model için en fazla bilgi veren Form3 olduğu gözlemlenmiştir.

Her iki modelde de Form 4 tedi parametrelerin diğer formlarla farklılaşma gösterdiği sonucu elde edilmiştir. Bu durumda Form4 teki kategorilerin öğrencileri işaretleme konusunda yönlendirdiği anlamına gelmektedir. Dolayısı ile ölçek geliştirme sürecinde daha çok bireylerin sadece duygularını yansıtacak formlar hazırlarken dereceleme ölçek kategorilerine yer verilmemesine dikkat edilmelidir.

Kaynaklar

Nartgün, Z. (2002). *Aynı tutumu ölçmeye yönelik likert tipi ölçek ile metrik ölçeğin madde ve ölçek özelliklerinin klasik test kuramı ve örtük özellikler kuramına göre incelenmesi* (Tez No. 113510) [Doktora tezi, Hacettepe Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models.* Sage.

Robitzsch, A., Kiefer, T., and Wu, M. (2021). *TAM: Test analysis modules* (version 3.7-16) [Computer Software]. https://cran.r-project.org/package=TAM.

Steiner, M., & Grieder, S. (2020). EFA tools: An R package with fast and flexible implementations of exploratory factor analysis tools. *Journal of Open Source Software*, *5*(53), 2521. https://doi.org/10.21105/joss.02521

R Core Team (2020). *R: A language and environment for statistical computing* (version 3.0.1) [Computer Software]. http://www.R-project.org/

Açımlayıcı faktör analizinde şans başarısının uyum indeksi, bilgi kriteri ve iç tutarlığa etkisi

Gökhan Kumlu ve Nuri Doğan

Anahtar kelimeler: Açımlayıcı faktör analizi, bilgi kriteri, iç tutarlık, şans başarısı, uyum indeksi

Giriş

Uluslararası ve Ulusal düzeyde yapılan sınavlarda sıklıkla tercih edilen seçmeli testler (Turgut ve Baykul, 2012), öğrencilere hazır olarak verilen cevaplardan birini seçerek yanıtladıkları test türüdür. Bu tür testler arasında çoktan seçmeli testler, doğru yanlış testleri ve eşleştirmeli testler sayılabilir (Doğan, 2019). Seçme gerektiren testlerde kolay uygulanabilme ve puanlanabilme, objektiflik, daha fazla soru ile daha fazla davranışın yoklanabilmesi gibi avantajlarının olmasının yanında, üst düzey becerileri ölçebilen soruların hazırlamanın zorluğu, doğru cevabı şansla bulma gibi dezavantajları da bulunmaktadır (Başol, 2018).

Seçmeli testlerde cevaplayıcının doğru cevabı tamamen şansla bulması sonucu elde ettiği puanı şansla kazanması, şans başarısı olarak tanımlanır. Şans başarısı, madde ve test puanlarında hataya sebep olmasından dolayı test puanlarının güvenirliğinin ve geçerliğinin düşmesi sorununa yol açar (Turgut ve Baykul, 2012). Şans başarısının test puanların üzerindeki etkisine yönelik alanyazında yapılan çalışmalar incelendiğinde Zimmerman ve Williams (1965) şans başarısının test puanlarının hatasını artıracağını, Mattson (1965) şans başarısının test puanlarına ilişkin güvenirliği azaltacağını ve standart hatayı artıracağını, Lord (1963) şans başarısına ilişkin hatadan düzeltilmiş puanların daha geçerli olacağını belirtmişlerdir.

Şans başarısının test ve madde puanları üzerinde önemli düzeyde etkisinin bulunması, dolaylı olarak testin yapısını da etkilemektedir. Cukadar (2019), ölçüm göstergelerinin altında yatan faktör sayısını belirlemede şans başarısının açımlayıcı faktör analizi, paralel analiz, Kaiser kuralı ve Cattell's scree testinin performansı üzerindeki etkisini araştırmak amacıyla bir çalışma yürütmüştür. Çalışma sonuçları, dört yöntemin de faktör sayısını belirlemede şans başarısının olmadığı duruma göre şans başarısının varlığında daha kötü performans gösterdiğini göstermiştir. Verilerin boyutluluğunu belirlemek için temel boyutluluk testi (DIMTEST; Stout, Froelich ve Gao, 2001), normal-ogive harmonik analizi robust yöntemi (NOHARM; McDonald, 1962) ve test puanlaması, madde istatistikleri ve madde faktör analizi

(TESTFACT; Monahan ve diğ., 2007) üzerinde şans başarısının etkisini araştırmaya yönelik çalışmalar da gerçekleştirilmiştir.

Alan yazında şans başarısının test ve madde puanları üzerine etkisine ilişkin çok fazla sayıda çalışmaya rastlanırken (Doğan ve diğ., 2007; Frary, 1985), testin yapısı üzerindeki etkisine yönelik çalışmaların daha sınırlı sayıda kaldığı gözlenmektedir (Stone ve Yeh, 2006; Tate, 2003). Bu çalışmada şans başarısı düzeyinin artmasıyla açımlayıcı faktör analizi sonucunda tek boyutlu veri yapısı için elde edilen uyum indeksleri, bilgi kriterleri ve iç tutarlık değerlerine etkisinin araştırılması amaçlanmıştır. Bu bağlamda şans başarısı artışının RMSR (Root Mean Square Residual), GFI (Goodness of Fit), CAF (common part accounted for), RMSEA (Root Mean Square Error of Approximation), TLI (Tucker Lewis Index), CFI (Comparative Fit Index), MFI (MacDonald Fit Index), BIC (Bayesian Information Criterion), AIC (Akaike Information Criterion), CAIC (Consistent Akaike Information Criterion), SABIC (Sample-Size Adjusted BIC) uyum indeksleri, bilgi kriterleri ile iç tutarlık değerlerine etkisi araştırılmıştır.

Yöntem

Araştırma farklı düzeylerdeki şans başarısının açımlayıcı faktör analizi sonucunda elde edilen uyum indekslerine, bilgi kriterlerine ve testin güvenirliğine etkisini ortaya çıkarmak amacıyla simülatif bir çalışma olarak yürütülmüştür. Simülasyon yönteminde şans başarısı ile birlikte örneklem büyüklüğü, test uzunluğu, madde ve yetenek parametreleri kontrol edilmiştir. Tek boyutlu veriler ile yürütülen bu araştırmanın simülasyon koşullarına yönelik bilgiler Tablo 1'de verilmiştir.

Tablo 1

Araştırmanın Simülasyon Koşulları

Test Uzunluğu	Örneklem Büyüklüğü	a parametresi (Madde ayırt edicilik düzeyi)	c parametresi (Şans başarısı düzeyi)
15	125	μ=1.20; σ=0.20	μ=0.10; σ=0.05
30	250	μ=2.00; σ=0.20	μ=0.20; σ=0.05
60	500		μ=0.30; σ=0.05
	1000		μ=0.40; σ=0.05
			μ=0.50; σ=0.05

Tablo 1'e göre test uzunluğu 3, örneklem büyüklüğü 4, madde ayırt edicilik düzeyi 2, şans başarısı düzeyi 5 farklı koşul olmak üzere toplamda 120 koşul ele alınmıştır. Alan yazında yapılan araştırmalarda simülatif çalışmalara ilişkin sonuçların tutarlı ve genellenebilir olması açısından 100 tekrarın yapılmasının yeterli olduğunun belirtilmesinden dolayı (Cao, 2008; Kim ve Kolen, 2006), bu araştırmada her veri seti için 100 tekrar yapılmıştır. Dolayısıyla araştırma 120 (koşul) x 100 (tekrar) = 12000 veri seti üzerinde yürütülmüştür.

Verilerin üretilmesi aşamasında, araştırmanın simülasyon koşulları çerçevesinde 3 farklı test uzunluğu ve 4 farklı örneklem büyüklüğü dikkate alınmıştır (15 madde, 30 madde ve 60 madde; 125, 250, 500 ve 1000). Örneklem büyüklüklerine ilişkin bireylerin yetenek dağılımları ortalaması 0 standart sapması 1 olan N(0,1) normal dağılımdan üretilmiştir. Test maddeleri üç parametreli lojistik modele (3PLM) göre üretilmiştir. Madde ayırt edicilik düzeyini gösteren a parametresi değerleri 2 farklı düzeyde 1,00 ile 1,50 arasında değişen ortalaması 1,20 ve standart sapması 0,20 ile değerleri 1,75 ile 2,25 arasında değişen ortalaması 2,00 ve standart sapması 0,20 olan tek biçimli (uniform) dağılımdan üretilmiştir. Güçlük düzeyini gösteren b parametresi -2 ve +2 arasında değerler alacak biçimde ortalaması 0, standart sapması 0,65 olan normal dağılımdan üretilmiştir. Şans başarısı düzeyini gösteren c parametresi 5 farklı düzeyde (i) değerleri 0,05 ile 0,15 arasında değişen ortalaması 0,10 ve standart sapması 0,05, (ii) değerleri 0,16 ile 0,25 arasında değişen ortalaması 0,20 ve standart sapması 0,05, (iii) değerleri 0,26 ile 0,35 arasında değişen ortalaması 0,30 ve standart sapması 0,05, (iv) değerleri 0,36 ile 0,45 arasında değişen ortalaması 0,40 ve standart sapması 0,05, (v) değerleri 0,46 ile 0,55 arasında değişen ortalaması 0,50 ve standart sapması 0.05 olan tek biçimli (uniform) dağılımdan üretilmiştir. Verilerin üretilmesinde ve analizlerinde R yazılımının "psych" ve "EFA.dimensions" paketlerinden yararlanılmıştır.

Sonuçlar

İlk iki koşul için yapılan analiz sonrası tek boyutlu veri yapısı için uyum indeksleri, bilgi kriterleri ve iç tutarlık değerlerine ilişkin elde edilen ön sonuçlar Tablo 2 ve Tablo 3'te verilmiştir.

i) madde ayırt edicilik düzeyi μ=1,20; σ=0,20, test uzunluğu 15 madde, örneklem büyüklüğü 125 için,

Tablo 2

Birinci Koşul

Uyum indeksleri, bilgi kriterleri ve iç tutarlık	Şans başarısı düzeyi				
	0.10	0.20	0.30	0.40	0.50
RMSR	0.111	0.117	0.123	0.132	0.144
RMSEA	0.183	0.180	0.181	0.193	0.212
GFI	0.867	0.799	0.709	0.613	0.501
CAF	0.575	0.592	0.621	0.626	0.653
TLI	0.423	0.348	0.259	0.175	0.082
CFI	0.505	0.441	0.365	0.292	0.213
MFI	0.238	0.254	0.246	0.211	0.152
BIC	36.918	30.099	31.218	82.239	167.598
AIC	291.466	284.647	285.766	336.787	422.147
CAIC	-53.082	-59.901	-58.782	-7.761	77.598
SABIC	521.450	514.631	515.750	566.771	652.131
CRα	0.849	0.808	0.760	0.716	0.662
Omega	0.852	0.812	0.766	0.724	0.674

ii) madde ayırt edicilik düzeyi μ=1,20; σ=0,20, test uzunluğu 15 madde, örneklem büyüklüğü 250 için;

Tablo 3

İkinci Koşul

Uyum indeksleri, bilgi kriterleri ve iç tutarlık	Şans başarısı düzeyi				
	0.10	0.20	0.30	0.40	0.50
RMSR	0.078	0.083	0.089	0.095	0.105
RMSEA	0.112	0.110	0.112	0.117	0.131
GFI	0.933	0.892	0.825	0.749	0.646
CAF	0.538	0.543	0.564	0.565	0.585
TLI	0.695	0.634	0.532	0.440	0.307
CFI	0.739	0.687	0.599	0.520	0.406
MFI	0.570	0.582	0.570	0.544	0.467
BIC	-121.611	-129.928	-119.883	-95.347	-15.929
AIC	195.321	187.004	197.048	221.584	301.002
CAIC	-211.611	-219.928	-209.883	-185.347	-105.929
SABIC	445.862	437.545	447.590	472.125	551.544
CRα	0.857	0.818	0.769	0.729	0.684
Omega	0.859	0.821	0.773	0.734	0.690

Tablo 2 ve 3'teki uyum indeksleri, bilgi kriterleri ve iç tutarlık değerlerine göre, şans başarısı düzeyinin artması ile birlikte model veri uyumunun ve iç tutarlık değerlerinin azaldığı görülmektedir. Diğer yandan örneklem büyüklüğünün artmasıyla şans başarısının etkisinin azalma eğilimi gösterdiği söylenebilir.

Dolayısıyla tüm koşullar analiz edildiğinde şans başarısının etkisinin gözlenmesi yanında örneklem büyüklüğü, test uzunluğu ve madde ayırt edicilik düzeyinin artmasıyla tek boyutlu veri yapısı için model veri uyumunu ve iç tutarlık değerlerini daha az etkilemesi beklenmektedir.

Kaynaklar

Başol, G. (2018). *Eğitimde ölçme ve değerlendirme.* Pegem Akademi.

Cao, L. (2008). *Mixed-format test equating: Effects of test dimensionality and common item sets* (Publication No. 3341415) [Doctoral dissertation, University of Maryland]. ProQuest Dissertations & Theses Global.

Cukadar, I. (2019). *An evaluation of four methods for determining the number of factors underlying measurement indicators under the presence of guessing effects* (Publication No. 13809098) [Doctoral dissertation, The Florida State University]. ProQuest Dissertations & Theses Global.

Doğan, N. (2019). Geleneksel ölçme ve değerlendirme teknikleri I: Yanıtı seçmeyi gerektiren ölçme araçları. N. Doğan (Ed.), *Eğitimde ölçme ve değerlendirme* içinde (s.113-138). Pegem Akademi.

Doğan, N., Çetin, S. & Gelbal, S. (2007). *The use of self-assessment in improving quessing.* The Jubilee Paper presented at the Iternational Scientific Conference under the heading "Science, Education and Time as Our Concern" November 30 – December 1, 2007, Plovdiv University, Smolyan, Bulgaria.

Frary, R. B. (1985). Multiple-choice versus free-response: A simulation study. *Journal of Educational Measurement, 22*(1), 21-31. https://doi.org/10.1111/j.1745-3984.1985.tb01046.x

Kim, S., & Kolen, M.J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education, 19*(4), 357-381. https://doi.org/10.1207/s15324818ame1904_7

Lord, F. M. (1963). Formula scoring and validity. *Educational and Psychological Measurement, 23*, 663-672. https://doi.org/10.1177/001316446302300403

Mattson, D. (1965). The effects of guessing on the standart error of measurement and the reliability of test scores. *Educational and Psychological Measurement, 25*, 727-730.

McDonald, R. P. (1962). A general approach to non-linear factor analysis. *Psychometrika, 27*(4), 397-415. https://doi.org/10.1007/BF02289646

Monahan, P. O., Stump, T. E., Finch, H., & Hambleton, R. K. (2007). Bias of exploratory and cross-validated DETECT index under unidimensionality. *Applied Psychological Measurement, 31*(6), 483-503. https://doi.org/10.1177/0146621606292216

O'Connor, B. P. (2021). *EFA.dimensions: Exploratory factor analysis functions for assessing dimensionality* (version 0.1.7.2). https://cran.r-project.org/web/packages/EFA.dimensions/index.html

Revelle, W. (2021). *psych: Procedures for psychological, psychometric, and personality research* (version 2.1.6) [Computer Software]. https://cran.r-project.org/web/packages/psych/index.html

Stone, C. A. & Yeh, C. C. (2006). Assessing the dimensionality and factor structure of multiplechoice exams: An empirical comparison of methods using the multistate bar examination. *Educational and Psychology Measurement, 66*(2), 193-214. https://doi.org/10.1177/0013164405282483

Stout W., Froelich A.G., & Gao F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 357-375). Springer.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of response to test items. *Applied Psychological Measurement, 27*, 3, 159-203. https://doi.org/10.1177/0146621603027003001

Turgut, M. F. ve Baykul, Y. (2012). *Eğitimde ölçme ve değerlendirme* (4. baskı). Pegem Akademi.

Zimmerman, D. W., and Williams, R. H. (1965). Effect of chance success due to guessing on error of measurement in multiple-choice tests. *Psychological Reports, 16*, 1193-1196. https://doi.org/10.2466/pr0.1965.16.3c.1193

Madde tepki kuramına dayalı test eşitlemede ortak madde oranının ve madde ayırt ediciliğinin eşitleme hatasına etkisi

Yıldız Yıldırım, Tuba Gündüz ve F. Gül İnce Aracı

Anahtar kelimeler: Madde tepki kuramına dayalı test eşitleme, madde ayırt ediciliği, *a* parametresi, ortak madde oranı, eşitleme hatası

Giriş

Günümüzde test puanları bireysel ve kurumsal düzeylerde önemli kararlar alırken bilgi sağlamak için sıklıkla kullanılmaktadır (Kolen ve Brennan, 2014). Ülkemizde de geniş ölçekli sınavlarda, okul düzeyindeki sınavlarda ve kurumlara giriş sınavlarında da test puanlarının kullanımı söz konusudur. Bu sınavlardan farklı zamanlarda elde edilen test puanlarının karşılaştırılabilir olması gerekmektedir. Örneğin ülkemizde lisans mezunu öğrenciler lisansüstü eğitimlerine başvurmak için "Yabancı Dil Sınavı"na (YDS) girmektedir. Bu sınav yılda iki kez uygulanmakta olup her sınav farklı formdadır. Yani farklı maddelerden oluşmaktadır. Ayrıca YDS'nin geçerlilik süresi kurumdan kuruma değişmekle birlikte, genel olarak 5 yıldır (ÖSYM, 2014). Aynı lisansüstü programına başvuran ve farklı yıllarda YDS'ye giren öğrencilerin maddelerin farklılığından kaynaklı olarak madde güçlükleri de farklı olacağı için karşılaştırılması yanlış sonuçlara götürmektedir. Bu durum "Akademik Personel ve Lisansüstü Eğitim Sınavı (ALES)" için de geçerlidir. Çünkü aynı yetenek düzeyindeki iki birey farklı sınavlara girerek farklı puanlar alabilir. Bu olumsuzlukları gidermek, farklı bir test formunu alan katılımcının diğer bir katılımcıdan daha zor bir test formu almasının doğuracağı sonucu engellemek ve aynı testin pek çok formu üzerinde karşılaştırılabilir puanlar elde etmek için test formlarının eşitlenmesine ihtiyaç duyulur.

Eşitlemenin "Randum Grup Deseni", "Tek Grup Deseni" , "Dengelenmiş Tek Grup Deseni" , ve "Ortak Maddeli Eşdeğer Olmayan Grup Deseni" olmak üzere dört deseni vardır. Ortak maddeli eşdeğer olmayan grup deseninde ortak maddeler içsel ya da dışsal olmak üzere iki türlü olabilir (Kolen ve Brennan, 2014). Ortak maddeler içsel olduğunda bu maddelerden elde edilen puan test puanına dâhil edilirken, dışsal olduğunda dâhil edilmemektedir. Literatürde ortak madde sayısı tartışmalı bir konu olarak ele alınmış ve araştırmalar yapılmıştır (Hills ve diğ., 1988; Yang ve Houang ,1996; Bastari, 2000; Kim ve Cohen, 2002; Walker ve Kim, 2009; Meng, 2012 ve Uysal, 2014). Bu araştırmaların birçoğunun ortak yanı ortak madde sayısı arttıkça eşitleme hatasının azaldığı sonucunu bulmalarıdır ancak ortak madde sayısının artmasının eşitleme hatasını arttırdığı sonucuna varan araştırmalar da bulunmaktadır.

Bu nedenle eşitleme hatasının düşük olması için en uygun ortak madde oranının ne olduğunu belirlemek amacıyla yapılan araştırmalar devam etmektedir. Bu araştırmada da "Ortak Maddeli Eşdeğer Olmayan Grup Deseni" kullanılmıştır ve araştırmanın amaçlarından biri de ortak madde sayısının eşitleme hatasına etkisini incelemektir.

Eşitlemenin desenlerinin yanı sıra yöntemleri de bulunmaktadır. Bunlar "Ortalama Eşitleme", "Doğrusal Eşitleme", Eşit Yüzdelikli Eşitleme" ve "Madde Tepki Kuramına Dayalı Eşitleme"dir. Madde Tepki Kuramına Dayalı Eşitleme klasik test kuramının sınırlılıkları nedeniyle diğer eşitleme yöntemlerinden üstündür. Literatürde KTK'ya ve MTK'ya dayalı eşitlemelerin eşitleme hatasına etkisini inceleyen araştırmalarda da bu üstünlük kanıtlanmıştır (Caldwell, 1984; Yang, 1997; Chen, 2001, Şahhüseyinoğlu, 2005; Bozdağ, 2007) Bu üstünlük nedeniyle bu araştırmada MTK'ya dayalı eşitleme yöntemi kullanılacaktır.

Son olarak bu araştırma da diğer araştırmalardan farklı olarak madde ayırt ediciliğinin eşitleme hatasına etkisi incelenecektir. Literatürde madde ayırt ediciliğinin eşitleme hatasına etkisini inceleyen bir araştırmaya daha önce rastlanmadığı için bu araştırmanın önemli olduğu düşünülmektedir. Tüm bunlar doğrultusunda araştırmanın problem cümlesi şu şekilde ifade edilebilir:

"Madde Tepki Kuramına dayalı test eşitlemede ortak madde oranı ve madde ayırt ediciliği eşitleme hatasını nasıl etkilemektedir?"

Yöntem

Araştırmada kullanılan veriler WinGen 3.1 (Han, 2007; Han ve Hambleton, 2007) programıyla üretilecektir. Araştırmada 3 farklı madde ayırt edicilik düzeyi (düşük (a=0.00-0.50), orta (a=0.50-1.00), yüksek (a=1,00-1,50))× 3 farklı ortak madde oranı (%10, %20 ve %30) olmak üzere toplam 9 koşul bulunacaktır ve her koşul için 25 replikasyon yapılacaktır. Veri üretimi aşamasında birey parametrelerinin dağılımı normal olacak ve ortalama 0, standart sapma 1 alınacaktır ve veri setleri 3000 bireyden oluşacaktır. Üretilecek test 40 maddeden oluşmaktadır ve veriler 3 parametreli lojistik modele dayalı olarak üretilecektir. Ayrıca her bir koşul için bir X formu (yeni form) bir de Y formu (eski form) üretilecektir. Bu doğrultuda toplam 9 koşul×2 form×25 replikasyon olmak üzere toplam 450 madde parametresi analizi ve 225 eşitleme yapılacaktır.

Toplam 9 koşul için madde parametreleri ve posterior birey dağılımları BILOG-MG programıyla elde edilecektir. Ayrıca her bir koşul için ölçek dönüştürme işlemi ST programıyla yapılacaktır. ST programının çıktılarından elde edilen eğim ve kesme katsayılarından Stocking-Lord'a göre elde edilen katsayılar kullanılacaktır (Stocking ve Lord, 1983). Araştırmada Stocking-Lord'a göre ölçek dönüştürme kullanılacak olmasının nedeni literatürde genel olarak bu yönteme göre yapılan dönüştürmelerin eşitleme hatasının diğer yöntemlerden daha az olmasıdır (French, 1996; Kilmen, 2010; Karkee ve Wright, 2004; Kim ve Kolen, 2006; Meng, 2012; Speron, 2009; Uysal, 2014;). Bu yönteme göre ölçek dönüştürme işlemi yapıldıktan sonra PIE programıyla her bir koşul için MTK'ya dayalı olarak ortak maddeli eş değer olmayan gruplar desenine göre eşitleme yapılacaktır. Her bir koşul için eşitleme yapıldıktan sonra

eşitleme hatası, eşitlemenin standart hatası ve kareler farkının ortalamasının karekökü (RMSD) hesaplanarak incelenecektir.

Sonuçlar

Bu araştırmada madde ayırt ediciliği (a parametrelerinin) düşük, orta ve yüksek olması ile ortak madde oranının %10, %20 ve %30 olmasının eşitleme hatasına etkisi incelenecektir. Araştırmanın sonucunda eşitleme hatası değeri olarak elde edilen RMSD'nin ve eşitlemenin standart hatasının madde ayırt ediciliği bağlamında en düşük değerini hangi düzeydeki madde ayırt ediciliği koşulunda alacağı belirlenmiş olacaktır. Yine madde ayırt ediciliği bağlamında en yüksek değerini ise düşük, orta ve yüksek madde ayırt ediciliği düzeyi koşullarından hangisinde alacağı belirlenecektir. Eşitleme hatasının daha az olduğu madde ayırt edicilik düzeyi belirlenen koşullar altında tespit edilmiş olacağı için araştırmacılara ve uygulayıcılara bu doğrultuda önerilerde bulunulacaktır.

Araştırmada incelenecek bir diğer husus olan ortak madde oranı bağlamında ise eşitleme hatası RMSD ve eşitlemenin standart hatası değerlerinin en yüksek olduğu ortak madde oranı koşulunun belirlenen ortak madde oranlarından (%10, %20 ve %30) hangisi olduğu belirlenmiş olacaktır. Diğer yandan eşitlemenin standart hatasının ve RMSD'nin en düşük olduğu koşulda belirlenecektir. Ortak madde oranı koşulları için genel beklenti, ortak madde oranı arttıkça eşitleme hatasının azalmasıdır. Ek olarak koşulların çaprazlanmasıyla tüm RMSD değerleri birbirleriyle karşılaştırıldığında hatanın en düşük olduğu ve en yüksek olduğu madde ayırt ediciliği × ortak madde oranı koşulu da belirlenmiş olacaktır.

Kaynaklar

Bastari, B. (2000). *Linking multiple choice and constructed response items to a common proficiency scale.* (Unpublished doctoral dissertation) University of Massachusetts, Amherst.

Bozdağ, S. (2007). *Şans başarısının test eşitlemeye etkisi* (Tez No. 209076) [Yüksek lisans tezi, Mersin Üniversitesi], Yükseköğretim Kurulu Tez Merkezi.

Caldwell, L. J. (1984). *A comparision of equating error in lineer and rasch model test equating method.* (Unpublished doctoral dissertation). Florida State University, Tallahassee.

Chen, H. W. (2001). *Calibration of the ITBS test battery to the complete test battery: A comparision five linking methods* (Unpublished doctoral dissertation). University of Iowa, Iowa City.

French, D. J. (1996). *The utility of Stocking & Lord's equating procedure for equating norm-referenced and criterion-referenced tests with both dichotomous and polytomous components* (Unpublished doctoral dissertation). University of Texas, Texas.

Han, K. T. (2007a). WinGen: Windows software that generates irt parameters and item responses. *Applied Psychological Measurement, 31*(5), 457-459.

Han, K. T., and Hambleton, R. K. (2007). User's Manual: WinGen 2 (*Center for Educational Assessment Report No. 642).* Amherst, MA: University of Massachusetts.

Karkee, T. B., & Wright, K. R. (2004, April). *Evaluation of linking methods for placing three parameter logistic item parameter estimates onto a one-parameter scale.* Annual Meeting of the American Educational Research Association'da sunulmuş bildiri, San Diego, California.

Kilmen, S. (2010). *Madde tepki kuramına dayalı test eşitleme yöntemlerinden kestirilen eşitleme hatalarının örneklem büyüklüğü ve yetenek dağılımına göre karşılaştırılması.* (Tez No. 279926) [Doktora tezi, Ankara Üniversitesi], Yükseköğretim Kurulu Tez Merkezi.

Kim, S., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement, 26*(1), 25-41.

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education, 19*(4), 357-381.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scalling, and linking.* (3rd ed.). USA: Springer.

Meng, Y. (2012). Comparison of Kernel equating and item response theory equating methods (Unpublished doctoral dissertation). Massachusetts University, Massachusetts.

ÖSYM. (2014). YDS Kılavuzu. http://www.osym.gov.tr/dosya/1-71452/h/2014-ydskilavuz.pdf sayfasından erişilmiştir.

Speron, E. (2009). *A comparison of metric linking procedures in item response theory* (Unpublished doctoral dissertation). Illinois Institute of Technology, Chicago.

Stocking, M. L., and Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210.

Şahhüseyinoğlu, D. (2005). *İngilizce yeterlik sınavı puanlarının üç farklı eşitleme yöntemine göre karşılaştırılması* (Tez No. 160014) [Doktora tezi, Hacettepe Üniversitesi], Yükseköğretim Kurulu Tez Merkezi.

Uysal, İ. (2014). *Madde tepki kuramına dayalı test eşitleme yöntemlerinin karma modeller üzerinde karşılaştırılması* (Tez No. 300226) [Yüksek lisans tezi, Abant İzzet Baysal Üniversitesi], Yükseköğretim Kurulu Tez Merkezi.

Walker, M. E., & Kim, S. (2009, 13-17 April). *Linking mixed-format tests using multiple choice anchors.* Paper presented at the annual meeting of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME), San Diego.

Yang, W., & Houang, R. T. (1996, 11 April). *The effect of anchor length and equating method on the accuracy of test equating: Comparisons of linear and IRT-based equating using anchor-item design.* Paper presented at the AERA Annual Conference, New York.

Meta analizde ağırlıklandırma ve ağırlıklandırmama durumları genel etki büyüklüğünü nasıl etkilemektedir?

Yıldız Yıldırım ve Melek Gülşah Şahin

Anahtar kelimeler: Meta analizde ağırlıklandırma, Hunter ve Schmidt, Hedges ve Vevea, örneklem büyüklüğü, ters varyans, ağırlıklandırmama

Giriş

Son yıllarda bilimsel bilginin katlanarak çoğalması, bilgiye erişimin kolaylaşması, benzer amaca yönelik farklı koşullarda yapılmış çalışmaların artması gibi durumlar meta analize yönelik ilginin yoğunlaşmasına neden olmuştur. Bunun nedeni araştırmaların artmasıyla tüm sonuçları bir arada değerlendirme isteği ve çok sayıdaki araştırmayı sentezleme ihtiyacıdır. Genel olarak bakıldığında araştırmacıların kendi alanlarındaki konularda meta analiz çalışmaları tasarladığı görülmüştür. Ancak alan eğitimi araştırmacıları kendi alanlarında belirlediği konuda etkiyi sentezlemeyi amaçlarken, ölçme ve değerlendirme uzmanları meta analizde metodolojik açıdan karşılaştırma yapmaya odaklanmaktadır. Örneğin alan eğitiminde tasarlanan bir çalışmada farklı öğretim yöntemlerinin alan akademik başarısına etkisinin sentezlenmesi amaçlanırken tek ağırlıklandırma yöntemi, kestirim yöntemleri, model vb. kullanarak meta analiz yapılmaktadır. Öte yandan bu farklı koşulların da (yöntem, model vb.) etkisinin incelenmesi de meta analizi metodolojik açıdan geliştirmeye katkı sağlayacaktır. Bu çalışmada farklı ağırlıklandırma yöntemlerinin [Hedges ve Vevea (1998)'ın ters varyansla (örneklem hata varyansının tersi) ağırlıklandırma yöntemi, Hunter ve Schmidt (1990)'in örneklem büyüklüğüyle ağırlıklandırma yöntemi kullanılarak meta analiz ve ağırlıklandırma yapmadan meta analiz] meta analiz sonuçlarını nasıl değiştirdiği incelenecektir. Literatür incelendiğinde, farklı ağırlıklandırma yöntemlerinin karşılaştırıldığı çalışmaların bulunduğu görülmektedir (Englund ve diğ., 1999; Marin-Martinez ve Sanchez-Meca, 2009; Scmidt ve diğ., 2009). Englund ve diğ., (1999) ağırlıklandırma yapmamanın güvenilmezliğini test etmek amacıyla ağırlıklandırma ve ters varyansla ağırlıklandırma sonuçlarını gerçek veriler kullanarak karşılaştırmıştır. Marin-Martinez ve Sanchez-Meca (2009) ise yaptıkları araştırmada sabit etkiler ve rastgele etkiler modelinde Hedges ve Vevea'nın ters varyansla ağırlıklandırma yöntemi ve rastgele etkiler modelinde Hunter ve Schmidt'in örneklem büyüklüğü ile ağırlıklandırma yöntemi olmak üzere üç yönteme ilişkin sonuçları farklı koşullar altında simülasyon verisi kullanarak karşılaştırmıştır. Son olarak Scmidt ve diğ. (2009) de Hunter ve Schmidt'in yöntemi ve Hedges ve Vevea'nın yöntemlerini rastgele

etkiler modeli için karşılaştırmışlardır. Bu araştırmada diğer araştırmalardan farklı olarak hem ağırlıklandırma yapılmayarak hem de iki prosedüre (Hunter ve Schmidt'in yöntemi ve Hedges ve Vevea'nın ağırlıklandırma yöntemleri) dayalı ağırlıklandırma yaparak, bu üç koşulu birbiriyle Türk örnekleminde gerçek verilere dayalı olarak karşılaştırmak amaçlanmıştır. Bu durum araştırmanın önemini ortaya koymaktadır. Ayrıca araştırmacılara ağırlıklandırma yöntemi seçiminde ışık tutması açısından da önemli görülmektedir.

Yöntem

Bu çalışmada farklı ağırlıklandırma yönteminin ilişkisini belirlemek amacıyla, Hedges ve Vevea (1998)'ın ters varyansla ve Hunter ve Schmidt (1990)'in örneklem büyüklüğüyle ağırlıklandırma yöntemleri ile ağırlıklandırılmış ve ağırlıklandırılmamış (unweighted) etki büyüklükleri arasında korelasyon katsayısı hesaplanacaktır. Ayrıca bu ağırlıklandırma ve ağırlıklandırmama yöntemlerinden elde edilen genel etki büyüklükleri dahil olmak üzere meta analiz sonuçları birbiriyle karşılaştırılacaktır. Bu yöntemleri karşılaştırmak için araştırma kapsamında gerçek verilerden yararlanılacaktır. Bu doğrultuda meta analiz konusu olarak fen eğitiminde alternatif (tamamlayıcı) ölçme ve değerlendirme teknik ve yöntemlerinin kullanılmasının fen tutumuna etkisi seçilmiştir. Bu konuda çalışmalara erişmek için Google Akademik, Dergipark, ERIC, EBSCO ve Web of Science veri tabanları taranacaktır. Veri tabanlarının taranmasında kullanılan anahtar kelimeler Google Akademik ve EBSCO için ["alternatif ölçme ve değerlendirme" OR "tamamlayıcı ölçme ve değerlendirme" AND "tutum" AND "fen" AND "deneysel"], Dergipark için [alternatif ölçme ve değerlendirme OR tamamlayıcı ölçme ve değerlendirme AND tutum AND fen AND deneysel], ERIC için ["alternative assessment" "authentic assessment" attitude science experimental Turkey] ve Web of Science için ["alternative assessment" OR "authentic assessment" AND attitude AND science AND experimental AND Turkey] şeklinde belirlenmiştir. Bu anahtar kelimeler ile 2015-2021 yılları aralığında erişilen toplam birincil çalışma sayısı 976'dır. Meta analize dahil edilecek birincil çalışmaların belirlenmesi için dahil etme ölçütleri belirlenmiştir. Bu ölçütler i) Birincil çalışmanın 2015-2021 yılları aralığında yayınlanmış olması, ii) Birincil çalışmanın Türk öğrenciler ile yapılmış olması, iii) Fen ve Teknoloji dersinde tamamlayıcı ölçme ve değerlendirme teknik ve yöntemlerinin herhangi birinin kullanılmış olması, iv) Birincil çalışmanın deneysel desende (gerçek, yarı ve zayıf) tasarlanmış olması, v) Birincil çalışmanın 5, 6, 7 ve 8. sınıf düzeylerinden biri için yapılmış olması, vi) Birincil çalışmada bağımlı değişken olarak fen tutumunun ele alınmış olması, vii) Cohen d etki büyüklüğünü hesaplayabilecek istatistiklerin raporlanmış olması ve viii) Örneklem büyüklüklerinin raporlanmış olması şeklindedir. Ayrıca meta analize dahil edilecek çalışmalar sadece makale ya da sadece tez ile sınırlandırılmayacak, araştırma türü fark etmeksizin (bildiri, makale, tez, rapor vs.) ölçütlere uygun tüm çalışmalar meta analize dahil edilecektir. Ölçütlere uygun olmayan araştırmalar ise meta analiz kapsamı dışında bırakılarak dahil edilmeyen çalışmaların sayısı ve neden dahil edilmedikleri raporlanacaktır. Dahil etme kriterlerine uygun birincil çalışmalar için betimsel değişkenler ve etki büyüklüğünün hesaplanması için gerekli istatistikler Microsoft Excel'de kodlandıktan sonra verilerin analizi için CMA 3.0 programı kullanılacaktır. Ağırlıklandırmama ve örneklem büyüklüğü ile

ağırlıklandırma yapmak için varyans ve Cohen d'lerin kodlandığı opsiyon üzerinde varyans yerine sırasıyla 1 ve örneklem büyüklüğünün tersi (1/N) kodlanacaktır. Böylelikle birincil çalışmalardaki etki büyüklükleri ağırlıklandırılmayacak (eşit ağırlığa sahip olacak) ya da örneklem büyüklüğü ile ağırlıklandırılmış olacaktır.

Sonuçlar

Hedges ve Vevea (1998)'ın ters varyans ile ağırlıklandırma yöntemi ile ağırlıklandırmış, Hunter ve Schmidt (1990)'in örneklem büyüklüğü ile ağırlıklandırma yöntemi ile ağırlıklandırılmış ve ağırlıklandırılmamış (unweighted) etki büyüklükleri ile yapılan meta analizin sonuçlarını karşılaştırmayı amaçlayan bu araştırmada ağırlıklandırma yöntemlerinin genel etki büyüklüklerine etkisinin birbirlerinden farklılaşıp farklılaşmadığının ya da ne kadar farklılaştığının belirleneceği düşünülmektedir. Fen eğitiminde alternatif (tamamlayıcı) ölçme ve değerlendirme teknik ve yöntemlerinin kullanılmasının fen tutumuna etkisini meta analizle incelemede, bu yöntem ve teknikler birden fazla yöntem ve tekniği içerdiği (portfolyo, performans görevi, proje, rubrik, kavram haritası, yapılandırılmış grid, tanılayıcı dallanmış ağaç, öz / akran / grup değerlendirme, kelime ilişkilendirme testi vb.) ya da farklı örneklemlerde birincil çalışmalar yürütüldüğü için heterojenliğin söz konusu olacağı düşünülmektedir. Bu nedenle araştırma kapsamında rastgele etkiler modeline dayalı meta analiz yapılacağı öngörülmektedir. Elde edilen sonuçlar ışığında farklı ağırlıklandırma yöntemlerinin birbiriyle ilişkisi ortaya konacaktır. Bu doğrultuda araştırmacılara meta analizde kullanabilecekleri ağırlıklandırma yöntemleri konusunda ışık tutacaktır.

Kaynaklar

Englund, G., Sarnelle, O., & Cooper, S. D. (1999). The importance of data-selection criteria: Meta-analyses of stream predation experiments. *Ecology*, *80*(4), 1132-1141.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. Psychological Methods, 3, 486–504.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Sage Publications, Inc.

Hunter, J. E., & Schmidt, F. L. (2004). Methods of meta-analysis: Correcting error and bias in research findings (2nd ed.). Sage Publications, Inc.

Marín-Martínez, F., & Sánchez-Meca, J. (2010). Weighting by inverse variance or by sample size in random-effects meta-analysis. *Educational and Psychological Measurement*, *70*(1), 56-73.

Schmidt, F. L., Oh, I. S., & Hayes, T. L. (2009). Fixed-versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, *62*(1), 97-128.

Bilişsel tanı modellerine dayalı bireye uyarlanmış testlerde karar ağacı algoritması ile örtük sınıf kestirimi

Hüseyin Yıldız ve Murat Doğan Şahin

Giriş

Ölçme alanında yaygın kullanılan Klasik Test Kuramı (KTK) ve Madde Tepki Kuramı (MTK) bireylerin ölçülen özelliğe sahip oluş düzeylerini (Gerçek Puan) tespit etmeyi hedefleyen kuramlardır. Her iki teori de bireylerin gerçek skorlarını ortaya koymaya çalışsalar da, bireyin belirli becerilere sahip olup olmadığını ortaya koymada yetersiz oldukları söylenebilir.

Bilişsel Tanı Modelleri (BTM) puan ya da yetenek kestirimleri yapmak yerine, bireyleri sahip oldukları ya da sahip olmadıkları özellikler bakımından örtük sınıflara ayırmayı hedeflemektedir (Başokçu, 2014). Burada bahsedilen özellik kavramı; vasıf, görev, alt görev ya da beceri olarak ele alınabilir (Tatsuoka, 1995) Bu sayede herhangi bir bireyin, yerleştiği örtük sınıfa göre önceden tanımlanmış bazı becerilere sahip olup olmadığı ya da hangi olasılıkla sahip olabileceği ortaya konabilir. BTM uygulamaları Bilgisayarlı Bireyselleştirilmiş Test (BBT) formatında da gerçekleştirilebilmektedir (Cheng, 2009). Literatürde BBT temelli BTM uygulamaları için çeşitli başlangıç, ilerleme ve sonlandırma kuralları önerilmiştir (Sorrel ve diğ., 2020).

Son dönemde makine öğrenmesi yöntemlerinin geleneksel BBT uygulamalarına alternatif olabileceği düşünülmektedir. Özellikle, parametrik olmayan bir yaklaşım olması nedeniyle sıkça tercih edilen bir ML yöntemi olan karar ağaçlarının, ölçülen özelliğin boyut sayısının çok olduğu durumlarda efektif çözümler üretebildiğine ilişkin çalışmalara son dönemde sıkça rastlanmaktadır (Zheng ve diğ., 2020; Gonzales, 2020; Gonzalez, 2021).

Bu çalışma bireylerin örtük sınıf kestirimlerini gerçekleştirilecek bir karar ağacı modeli oluşturmayı ve elde edilen sınıflama başarısını ortaya koymayı amaçlamaktadır.

Yöntem

Araştırmada 100 maddelik yapay madde havuzun tamamı kullanılarak GDINA modeli altında yapılan örtük sınıf aidiyeti kestirimlerinin, oluşturulan karar ağacı yardımıyla test uzunlukları bireyin cevaplarına uyumlu olacak şekilde kısaltılarak tekrarlanması amaçlanmıştır. Kullanılan veri üretim koşullları, analiz araçları gibi bilgiler, maddeler halinde şu şekilde sıralanabilir.

1. Araştırmada 100 madde ve 1000 bireyden oluşan veri seti GDINA modeli altında simülatif olarak üretilmiştir.

2. Kullanılan Q matrisi 3 beceri sütunundan oluşmaktadır. Q matrisi "simcdm" paketinin (Balmuta ve diğ., 2019) "sim_q_matrix" fonksiyonu ile türetilmiştir.

3. Her bireyin örtük sınıf aidiyetleri "CDM" (Robitzsch vd, 2020) R paketi kullanılarak kestirilmiştir.

4. Madde havuzundaki 100 madde, aynı Q matris yapısında olan maddeler bir araya getirlecek şekilde 4'er maddelik oluşturacak şekilde madde kümeleri oluşturulmuştur.

5. 4'er maddelik madde kümelerini oluşturan maddelere ait 1-0 puanlar toplanarak, her bir madde kümesi için 0-4 aralığında toplam puanlar elde edilmiştir.

6. 24 madde kümesinin her biri bir değişken olmak üzere, bireyleri 8 (2^3) örtük sınıftan birine yerleştiren karar ağacı modeli WEKA (Frank et.al, 2016) programı kullanılarak oluşturulmuştur.

7. 1000 yapay katılımcıdan 750'si karar ağacını eğitmek için (Training Data), 250'si modeli sınamak için (Testing Data) kullanılmıştır.

8. Modelin testi sonucu elde edilen başarılı sınıflandırma yüzdesi değerleri raporlaştırılmıştır.

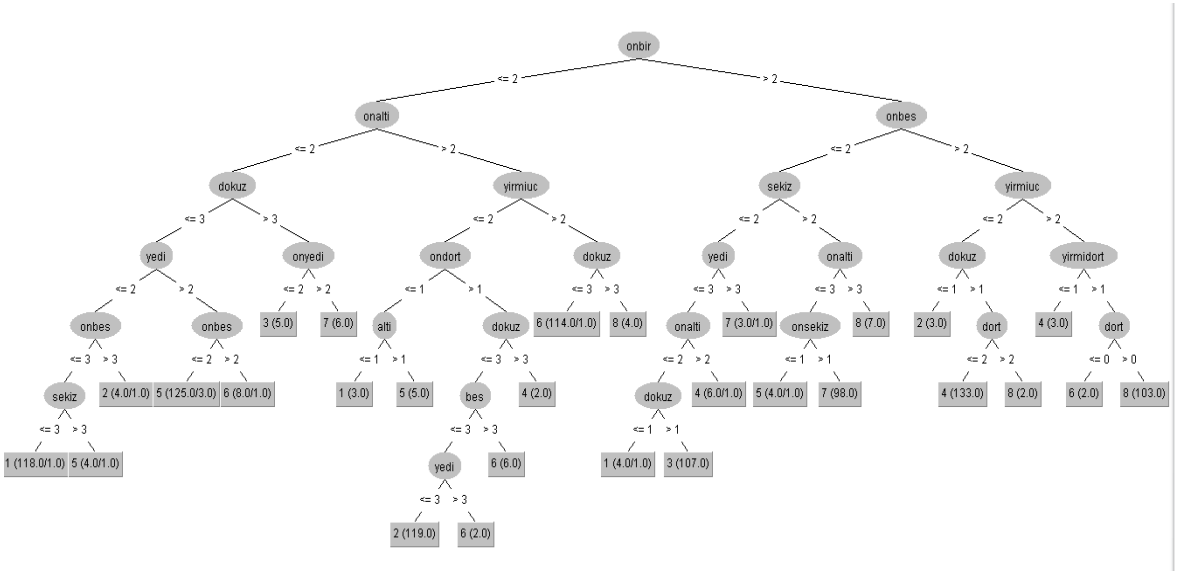
9. Her bireyin örtük sınıfının belirlenmesinde yanıtladıkları madde sayısı (test uzunluğu) belirlenmiştir.

Sonuçlar

Yapılan analizler sonucunda 250 kişilik test grubundaki bireylerin X maddelik kağıt-kalem testi ile belirlenen örtük sınıflar, karar ağacı temelli BBT uygulaması sonucunda %93.82 oranında doğru sınıflanmıştır. Yapılan sınıflandırma tahminleri için oluşturulan confision matrix aşağıda paylaşılmıştır.

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h   <-- classified as
 37  1  0  0  1  0  0  0 |  a = 1
  1 39  0  0  0  2  0  0 |  b = 2
  1  0 39  0  0  0  0  0 |  c = 3
  0  0  1 44  0  0  0  0 |  d = 4
  1  0  0  0 41  1  0  0 |  e = 5
  0  1  0  2  1 48  0  0 |  f = 6
  0  0  3  0  1  0 41  2 |  g = 7
  0  0  0  2  0  0  0 30 |  h = 8
```

250 bireyin karar ağacı temelli BBT uygulaması ile yanıtladıkları madde sayısı 12 ile 28 arasında değişmektedir. Ortalama test uzunluğunun ise 21.33 olduğu görülmüştür. Kullanılan algoritmanın testin ilerleme koşullarını ve sınıflama tahminlerini içeren ağaç görseli aşağıda paylaşılmıştır.



Yukarıdaki görselde her bir düğüm 4 maddelik bir mini testi temsil etmektedir. Her bir yaprak ise yapılan sınıflandırma sonucunu göstermektedir. Her birey teste 11 numaralı madde kümesi ile başlamış ve bu madde kümesinden elde ettiği puana göre farklı madde kümelerine yönlendirilerek sonunda örtük sınıf tahmini yapılmıştır. Bulgular incelendiğinde 96 madde ile kestirilen örtük sınıf aidiyetlerinin 6.18% hata payı ile ortalama 21.33 madde ile gerçekleştirilebildiği sonucuna ulaşılmıştır.

Madde havuzunun genişletilmesi, örneklem büyüklüğünün arttırılması, farklı budama (purification) metodlarının kullanılması, madde kümesi büyüklüklerinin farklılaştırılması ile ideal koşullar sağlandığında daha yüksek başarılı sınıflandırma oranları ya da daha kısa ortalama test uzunlukları elde edilebileceği düşünülmektedir.

Kaynaklar

Balamuta, J. J., Culpepper, S. A., and Hudson A. (2019). *simcdm: Simulate cognitive diagnostic model (CDM) data*. https://cran.r-project.org/package=simcdm

Cheng, Y. When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika,* 74, 619 (2009). https://doi.org/10.1007/s11336-009-9123-2

Frank, E., Hall, M. A., and Witten, I. H. (2016). *the weka workbench: data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann. https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf

Gonzalez, O. (2021). Psychometric and machine learning approaches for diagnostic assessment and tests of individual classification. *Psychological Methods, 26*(2), 236–254. https://doi.org/10.1037/met0000317

Gonzalez, O. (2021). Psychometric and machine learning approaches to reduce the length of scales. *Multivariate Behavioral Research, 56*(6), 903-919. https://doi.org/10.1080/00273171.2020.1781585

Robitzsch, A., Kiefer T., George, A. C., & Ünlü, A. (2020). *CDM: Cognitive diagnosis modeling* (version 7. 5-15) [Computer Software]. https://cran.r-project.org/package=CDM.

Sorrel, M. A., Barrada, J. R., de la Torre, J., & Abad, F. J. (2020) Adapting cognitive diagnosis computerized adaptive testing item selection rules to traditional item response theory. *PLoS One, 15*(1), e0227196. https://doi.org/10.1371/journal.pone.0227196

Tatsuoka, K. K. (1995). *Cognitive assessment: An introduction to the rule space method*. Routledge.

Zheng Y., Cheon H., and Katz C. M. (2020). Using machine learning methods to develop a short tree-based adaptive classification test: case study with a high-dimensional ıtem pool and ımbalanced data. *Applied Psychological Measurement, 44*(7), 499-514. https://doi.org/10.1177/0146621620931198

Measurement invariance testing with many groups: A comparison of BSEM and alignment optimization

Gözde Sırgancı, Gizem Uyumaz and Akihito Kamata

Keywords: Alignment method, bayesian estimation, approximate measurement invariance

Introduction

Measurement invariance is a psychometric property of a scale developed to measure a latent construct. The instrument is measurement invariant when the same construct is measured in the same way across different groups of the target population, such as countries, cultural units, time points, or regions within countries (Horn and McArdle, 1992; Meredith, 1993; Vandenberg and Lance, 2000; Vandenberg, 2002; Millsap, 2011; Davidov et al., 2014). The most widely used method to establish measurement invariance is multigroup confirmatory factor analysis (MGCFA; Bollen, 1989), but mostly for comparing two groups. There are disadvantages when it is used for comparing a large number of groups. The number of pairwise comparisons across groups on any measurement parameters exponentially increases as the number of groups increases and the chances of falsely detecting noninvariance are elevated when such a large number of comparisons are performed (Rutkowski and Svetina, 2014). Furthermore, poor model fit can be an issue when a model of exact invariance (identical measurement parameters across all groups) is specified (Asparouhov and Muthén, 2014). Given these methodological and practical issues of MG CFA in testing MI across many groups, an alternative approach has been called for. We identified two alternative approaches to many-group MI testing, including Bayesian approximate MI testing (Muthén and Asparouhov, 2013), and alignment optimization (Asparouhov and Muthén, 2014).

MG-BSEM relaxes assumptions about exact invariance of the item parameters thus allowing for small cross-group discrepancies (or "wiggle room") in item parameters (Muthén and Asparouhov, 2013; van de Schoot et al., 2013). In other words, whereas item intercepts and loadings are fixed to be equal across all groups ($\tau_{ig} = \tau_{ig}$ *and* $\lambda_{ig} = \lambda_{ig}$) in MG-CFA models, in the approximate invariance approach, applying MG-BSEM models, these constraints are relaxed based on the assumption that item-related parameters are approximately equal ($\tau_{ig} \approx \tau_{ig}$ *and* $\lambda_{ig} \approx \lambda_{ig}$). In MG-BSEM models, non-informative priors are used for all parameters except for the parameters defined for the allowed wiggle room in item

measurement parameters. In practical terms, while most parameters are freely estimated, the size of the expected items' measurement parameter differences must be predefined by the user using a prior. In the implementation in Mplus, the differences between item parameters are expressed in terms of a normal distribution, with a mean of zero and a difference variance that needs to be predefined, usually in the range of 0.001 and 0.1.

Alignment optimization or the alignment method is also a relatively new method developed by Asparouhov and Muthén (2014). Alignment optimization tries to search for an optimal set of measurement parameters from the configural invariance model. The Alignment method can predict factor mean and variance for each group without assuming measurement invariance and by discovering the most suitable measurement invariance pattern. With this aspect, the method gives information about the level of measurement invariance along with intergroup factor mean and variance by calculating approximate measurement invariance. Thereby, which measurement parameters are approximately constant, and which are not specified (Asparouhov & Muthén, 2014; Kim et al., 2017). In other words, the alignment method can estimate factor loadings (λ_g), measurement intercepts (v_g), factor means (α_g) and variances (ψ_g) by predicting the number of variable item parameters and the model that can hold impaired measurement variance at the minimum level (Muthén and Asparouhov, 2018).

The main aim of this article is to discuss their conceptual and implementation similarities and differences that inform decisions in the choice of methods for many-group MI testing.

Methods

Data for this analysis is taken from the European Social Survey (ESS) in 2018. The Questionnaire was developed by Schwartz (2003) to measures achievement and power (two basic values) from self-enhancement (higher-order value). In the real data analysis, we included 11 countries (see table1).

Table 1. *Frequencies and percentages*

Code	Country	f	%	Code	Country	f	%	Code	Country	f	%
1	Bulgaria	1503	8.55	5	Ireland	2082	11.84	9	Slovenia	1226	6.97
2	Cyprus	757	4.31	6	Norway	1353	7.70	10	Switzerland	1423	8.09
3	France	1804	10.26	7	Poland	1296	7.37	11	United	2122	12.07
4	Germany	2246	12.78	8	Serbia	1768	10.06		kingdom		
	Total									17580	100.00

The posterior predictive p value (PPP), deviance information criterion (DIC) and 95% confidence interval (CI) are used for model evaluation for BSEM. A good fit is achived if the PPP is around 0.50. In approximate MI testing, when the magnitude of noninvariance is large compared to the specified prior variance (model misspecification), the PPP is expected to be low, indicating model misfit (Muthén and Asparouhov, 2012). In the context of model comparison in Bayesian analysis, Cain and Zhang (2018) suggested that ΔPPP > .10 or 0.15 and ΔDIC > 7, 5, or 3 imply a considerable change in model fit

depending on the sample size and complexity of the model. Similarly, when the 95% confidence interval (CI) for the difference between observed and replicated chi-square values does not include zero, it indicates model misfit.

In Alignment methods, each measurement parameter the noninvariant groups are shown within parentheses. The fit function contribution is calculated to measure the level of noninvariance of a particular indicator. The R^2 measure reflects the variation of a measurement parameter across groups in the configural model that can be explained by the variation of factor means and factor variances. It ranges from 0 to 1, and the closer to 1 the R^2 is, the more invariant the parameter is.

The analyzes were performed in the Mplus 8.5 program.

Results

Table 2. *Maximum Likelihood CFA Model Fit*

Model	χ^2	df	p	RMSEA	CFI	SRMR
Configural	42.802	22	.005	.024	.998	.009
Metric	179.644	52	.000	.039	.988	.034
Scalar	1945.766	82	.000	.119	.820	.082

The results based on maximum likelihood CFA suggest that at least exact metric MI is supported (Table 2). The χ-difference metric to scalar model is very large. Further, Chen (2007) developed criteria for other fit statistics to indicate when MI is not given. According to these criteria, change in RMSEA should be less than .015, change in SRMR should be less than .010, and change in CFI should be less than -.010. The scalar model does not meet these criteria. However, the Bayesian exact scalar MI model by far has the worst fit.

Table 3. *Bayesian CFA Model Fit*

	Prior	DIC	ppp	95%CI
Exact MI				
Configural		223003.874	0.198	[-26.140, 67.890]
Metric		224304.461	0.198	[-26.182, 68.570]
Scalar		223278.985	0.000	[1108.976, 1214.548]
Approximate MI				
Scalar	N(0,0.001)	225702.043	0.000	[465.849, 609.438]
Scalar	N(0,0.005)	225298.519	0.000	[58.145, 173.523]
Scalar	N(0,0.010)	194136.303	0.018	[3.965, 105.582]
Scalar	N(0,0.025)	211380.694	0.135	[-20.947, 74.825]
Scalar	N(0,0.050)	212304.183	0.181	[-25.191, 69.168]
Scalar	N(0,0.100)	213624.802	0.192	[-26.247, 67.906]

DIC=deviance information criterion; PPP=posterior predictive p-value; CI=credibility interval; MI=measurement invariance.

We began the Bayesian analyses with a very small prior variance and successively increased it while we monitored DIC, PPP, and 95% credibility intervals for the difference between observed and

replicated chi-square value. Although this strategy is not a test of the adequacy of the prior variances (Hoijtink and Van de Schoot, 2017), it is possible to decide if model fit substantially improves with larger prior variance. The results show that the smallest prior variance υ = 0.001 does not fit the data, i.e. υ is too close to zero. By gradually adjusting υ we were able to achieve good model fit while maintaining convergence and identifiability. The less strict prior variance υ = 0.100 reveal no major improvement of model fit compared to υ = 0.025. Further, the PPP, and limits of 95% credibility interval for the difference between observed and replicated chi-square values for υ = 0.025 do not substantially differ from the exact metric MI model. Thus, υ = 0.025 is deemed sufficient to consider minor deviations from exact intercept equivalence and we assume that approximate scalar MI holds. Thus, it is reasonable to compare the latent means of self enhancement across groups.

Table 4. *Approximate Measurement (Non)Invariance for Intercepts and Loadings Over Countries*

Items	Intercepts	Loadings
ipshabt	1 2 3 **(4)** **(5)** 6 **(7)** 8 **(9)** 10 **(11)**	1 2 **(3)** 4 5 6 **(7)** 8 9 10 11
ipsuces	1 **(2)** **(3)** **(4)** 5 **(6)** **(7)** 8 9 **(10)** 11	1 2 3 4 **(5)** 6 7 8 9 10 11
imprich	1 2 3 **(4)** **(5)** **(6)** **(7)** **(8)** 9 10 11	1 2 3 4 5 6 7 8 9 10 11
iprspot	1 **(2)** 3 **(4)** 5 6 7 8 9 **(10)** **(11)**	1 2 3 4 5 6 7 8 9 10 11

Table 4 shows the (non)invariance results for the measurement intercepts and factor loadings. The countries that are deemed to have a significantly noninvariant measurement parameter are shown as bolded within parentheses. As seen in Table 4, most of the items show a large degree of measurement noninvariance for the measurement intercepts and, to a lesser extent, the loadings. The large degree of noninvariance is in line with the findings of the traditional approach using the scalar model. However, Table 4 also shows that items imprich and iprspot have no significant loadings, and therefore, factor variances and structural relationships can be compared across groups for these items.

Tablo 5

Model Fit Alignment Invariance Analysis

Items	Approximate measurement ınvariance holds for groups	Weighted average value across invariant groups	R^2 explained variance/ ınvariance index
1	1 2 3 6 8 10	4.498	0.559
2	1 5 8 9 11	3.106	0.567
3	1 2 3 9 10 11	3.361	0.790
4	1 3 5 6 7 8 9	3.242	0.801
	Approximate Invariance (Noninvariance) Holds For Groups		
1	1 2 4 5 6 8 9 10 11	0.617	0.174
2	1 2 3 4 6 7 8 9 10 11	1.047	0.317
3	1 2 3 4 5 6 7 8 9 10 11	1.122	0.771
4	1 2 3 4 5 6 7 8 9 10 11	0.731	0.633
Average Invariance index: 0.576			

The contribution of each item's interceps and factor loading to the optimized simplicity function is calculated by the measure of R^2. The R^2 is a useful descriptive statistic that gives the degree of noninvariance that can be absorbed by group-varying factor means and variances (Muthén and Asparouhov, 2014; 2018). In the configural model, it shows how much of the parameter variation across groups for each measurement parameter can be explained by factor means and variation in factor variances. R^2 value close to "1" implies a high degree of invariance and to "0" a low degree of invariance. The R^2 values of the item 1 indicate that the item contributed the least to the simplicity function. In other words, this item has most degree of noninvariant across the groups. Table 6 shows the factor means estimated for all groups by the alignment method and groups that have factor means significantly different on the .05 level.

Table 6

Comparison of Factor Means Between Countries

Ranking	Group	Factor Means	Groups with Significantly Smaller Factor Mean
1	3	0.652	11 6 4 5 2 10 8 1 7 9
2	11	0.155	5 2 10 8 1 7 9
3	6	0.127	5 2 10 8 1 7 9
4	4	0.075	2 10 8 1 7 9
5	5	0.000	2 10 8 1 7 9
6	2	-0.181	1 7 9
7	10	-0.217	1 7 9
8	8	-0.263	9
9	1	-0.352	9
10	7	-0.584	
11	9	-0.738	

For convenience of the presentation, the groups are ordered from high to low according to factor means and the groups that have factor means that differ on the 0.05 significance level are determined. For example, as seen in Table 7, the factor means of the 3th country estimated by the alignment method is 0.652 and this value is significantly higher than the countries whose codes written in the last column.

Discussion

While the Bayesian MG-CFA is a scale-level MI testing, but an individual item MI can be evaluated with modification indices, allignment is an item-level MI testing. The items of noninvariance are listed with the information of noninvariant groups. Bayesian CFA does not enforce exact MI that can be realistically challenging to achieve with my groups. Any benefits of Bayesian estimation (e.g., posterior distribution of parameter estimates) can be applied. Researchers can incorporate their knowledge on the noninvariance distribution in MI testing. Alignment allows factor mean comparisons under a certain degree of noninvariance. Alignment provides the most detailed information of noninvariance and factor mean differences. While specifying proper priors can be challenging in Bayesian

MG-CFA, the number of groups tested for MI can be limited in alignment methods. In terms of model specification, the Bayesian approach has greater flexibility than alignment optimization because the Bayesian MI testing is basically equivalent to MG-CFA using the Bayesian estimation method (Kim, Cao, Wang and Nguyen, 2017).

References

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(4), 495–508. https://doi.org/10.1080/10705511.2014.919210

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Cain, M. K., & Zhang, Z. (2018). Fit for a Bayesian: An evaluation of PPP and DIC for structural equation modeling. *Structural Equation Modeling, 26*, 39–50. https://doi.org/10.1080/10705511.2018.1490648

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, 40, 55–75. https://doi.org/10.1146/annurev-soc-071913-043137

Hoijtink, H., & van de Schoot, R. (2017). Testing small variance priors using prior-posterior predictive p values. *Psychological Methods, 23*, 561–569. https://doi.org/10.1037/met0000131

Horn, J. L. & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*(3), 117–144. https://doi.org/10.1080/03610739208253916

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(4), 524-544. https://doi.org/10.1080/10705511.2017.1304822

Millsap, Roger E. (2011). *Statistical Approaches to measurement ınvariance*. Routledge.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525–543. https://doi.org/10.1007/BF02294825

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological methods*, *17*(3), 313. https://doi.org/10.1037/a0026802

Muthén, B., & Asparouhov, T. (2013). BSEM measurement invariance analysis. *Mplus web notes*, *17*, 1-48.

Muthén, B., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups: alignment and random effects. *Sociological Methods & Research, 47*(4) 637-664. https://doi.org/10.1177/0049124117701488

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*, 31–57. https://doi.org/10.1177/0013164413498257

Schwartz, S. H. (2003). A proposal for measuring value orientations across nations. *Questionnaire Package of the European Social Survey*, *259*(290), 261.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the MI literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-70. https://doi.org/10.1177/109442810031002

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational research methods*, *5*(2), 139-158. https://doi.org/10.1177/1094428102005002001

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in psychology,* 4, 770. https://doi.org/10.3389/fpsyg.2013.00770

Tepki stillerinin ölçme değişmezliği üzerindeki etkisi: TIMSS 2019 örneği

Zafer Ertürk ve Oya Erdinç Akan

Giriş

Uluslararası yapılan geniş ölçekli sınavlar, kültürel olarak çeşitli ülkelerin eğitime yönelik tutumları ve değer yapılarındaki farklılıklarına dair önemli bilgiler sağlayabilir. Eğitim alanında, Uluslararası Öğrenci Değerlendirme Programı (PISA) ve Uluslararası Matematik ve Bilim Çalışmasında Eğilimler (TIMSS), uluslararası çapta deneysel araştırmalar için fırsatlar sunan büyük ölçekli çalışmalardan ikisidir. Bu çalışmalardan elde edilen anketler ve yayınların sayısının artması, tutum ve değerler arasındaki uluslararası farklılıklara olan ilginin artmasına kanıt niteliğindedir (Lu, 2012). TIMSS, çalışmaya katılan ülkelerdeki öğrencilerin matematik ve fen alanlarındaki başarı düzeyleri, öğretim programları, öğrenci, öğretmen ve okul niteliklerine ilişkin çeşitli bilgiler sağlamaktadır. Ülkelere ilişkin bu bilgilerden faydalanılarak matematik ve fen eğitiminin kalitesini arttırmak üzere ulusal ve uluslararası düzeyde karşılaştırmalı olarak öğretim programlarının ve yöntemlerinin değerlendirilmesi mümkün olmaktadır (Martin ve diğ., 2016).

Kültürlerarası karşılaştırma çalışmalarına büyük bir ilgi olmasına rağmen hala bu tür çalışmalara özgü genel kabul görmüş bir yaklaşım bulunmamaktadır. Yapılan çalışmalarda ilgili tüm gruplar için aynı araçlar (yani anketler, ölçekler, testler, vb.) kullanıldığı için, elde edilen sonuçların gruplar arasında karşılaştırılabilir olduğu varsayılmaktadır. Geçerlilik için, sonuçların karşılaştırılabilirliği varsayımı kritik olmasına rağmen, çoğu kez bu test edilmemektedir. Araştırmacılar genellikle yalnızca iki ya da daha fazla kültür grubunun ortalama puanlarındaki farklılığa odaklanmaktadır. Bununla birlikte her bir kültürel yapı birçok etken, süreç ve niteliği yansıttığından aynı sorular veya sınavlar seti farklı kültürlerden gelen insanlar için farklı anlamlara sahip olabilir. Yani her kültürde farklı yapılar ölçülüyor olabilir. Böyle durumlarda karşılaştırmalı araştırmanın sonuçlarının geçerliliği söz konusu olur. Bu nedenle, herhangi bir kültürlerarası araştırmayla ilgili temel konu, kültürler arası farklılıkları test ederken eşdeğerliği (yani karşılaştırılabilirliği) temin etmektir (Hui ve Triandis, 1989).

Yanlılık, farklı kültürlerde uygulanan araçların geçerliliğini tehlikeye atan olumsuz faktörleri ifade etmektedir. Bunun yanında eşdeğerlik, ölçme araçlarının içsel bir özelliği değil, kültürler arası karşılaştırmaların bir özelliğidir (He ve van de Vijver, 2013). Çünkü her bir kültürde uygulanan ölçme

aracının, o kültüre göre içeriği düzenlenmiştir. Böylece uygulanan ölçme aracından elde edilen sonuçlar, o kültür özelliklerini yansıtacaktır. Fakat uluslararası çalışmalarda kullanılan ölçme araçlarında her bir kültür için aynı anlamı yansıtacak içerik olmalıdır.

Kültürler arası karşılaştırma çalışmalarında, yanıtlayıcıların anket verileri yanıtlama stilleri ve hangi koşullar altında bu yanıt stillerinin meydana geldiği istatistiksel modellerle açıklanmaya çalışılmıştır. Uluslararası sınavlardaki öğrencilerin karşılaştırılması için kullanılan ölçeklerin belirtilen gruplar arasında ölçme değişmezliğinin sağlanması gerekmektedir (Borsboom, 2006). TIMSS' de matematik başarı puanları ile tutum ölçümleri arasındaki ilişki ülkeler arasındaki tepki stili farklılıklarından kaynaklanabilmektedir. Bu araştırmada da ülkelerin matematiğe yönelik tutum ölçümlerindeki ölçme eşdeğerliğinin sağlanamaması ve hatalı sonuçların elde edilmesi durumunda tepki stillerinin etkisinin incelenmesi amaçlanmaktadır. Bu amaç doğrultusunda tepki stili yanlılığı bulunan verilerin düzeltilmesinde ve kümelenmesinde kullanılan TSYDK (Tepki Stili Yanlılığının Düzeltilmesi ve Kümelenmesi) yönteminden yararlanılmıştır. TSYDK yöntemi ile tepki stili yanlılığı bulunan veriler düzeltilebilmektedir.

Geliştirilen ölçme araçlarında ilk olarak "araç uygulandığı her grupta aynı yapıyı ölçer" (Başusta ve Gelbal, 2015) varsayımı ile oluşturulmaktadır. Bu doğrultuda, farklı grup ve kültürlere uygulanan ölçme araçlarının ölçme değişmezliğinin incelenmesi önemli hale gelmektedir. Ölçme değişmezliği üzerinde olumsuz olduğu yapılan çalışmalar ile kanıtlanmış olan (He ve Fons, 2016; Vlimmeren ve diğ., 2017). Tepki stillerinin yanlılığının düzeltilmesi ölçme değişmezliğinin sağlanmasında önemli görülmektedir. Bu doğrultuda aşağıdaki araştırma sorulara yanıt aranmıştır:

1. TIMSS 2019 "Matematiği Sevme" ölçeği ülke düzeyinde ölçme değişmezliğini sağlamakta mıdır?
2. TSYDK yöntemi ile tepki stili yanlılığından arındırılmış TIMSS 2019 "Matematiği Sevme" ölçeği ülke düzeyinde ölçme değişmezliğini sağlamakta mıdır?

Yöntem

Bu çalışmanın amacı TIMSS 2019 sınavında yer alan "Matematiği Sevme Ölçeği" nin ülke düzeyindeki ölçme değişmezliği üzerindeki tepki stillerinin etkisinin incelenmesidir. Bu amaç doğrultusunda araştırmanın deseni TIMSS 2019 "Matematiği Sevme Ölçeği"nin kültürler arası geçerlik düzeyini saptamaya yönelik olduğundan betimsel bir araştırmadır (Karasar, 2013).

Araştırmanın evrenini TIMSS 2019 uygulamasına katılan ülkelerdeki 8. sınıf öğrencileri oluşturmaktadır. Ülkelerdeki öğrenci sayılarının fazla olması ve kayıp verilere çoklu atamanın yapılmasının tepki stillerini etkileyeceği düşünülerek liste bazında veri silme işlemi yapılmıştır. Araştırma kapsamında ülkeler belirlenirken tutum-başarı paradoksu sergileme durumları göz önüne alınmıştır. Tutum-Başarı paradoksunda, bir konu alanına yönelik olumlu tutum sergileyen öğrencilerin beklenilenin aksine akademik başarıları düşük olurken, tutumları olumsuz olan öğrencilerin ise

başarılarının yüksek olmasıdır. Ülkelerin tutum-başarı paradoksu sergileme durumları, ülkeler arası tutum-başarı korelasyon sonuçlarının negatif, ülke içi tutum-başarı korelasyonlarının pozitif olması ile ortaya çıkmaktadır.

Tutum-başarı paradoksunu belirlemek için ülkelerin matematiği sevme ölçeğinin örneklem ağırlıklandırması ile ortalaması alınmıştır. Bunun yanında, ülkelerin matematik başarılarını göstermek için TIMSS 2019'daki 5 olası (plausible) değerin örneklem ağırlıklandırması ile ortalaması alınmıştır. Ülke ortalamalarına göre matematik başarısı ile matematiği sevme ölçeği arasında negatif korelasyon elde edilirken, ülkelerin kendi içinde matematik başarısı ile matematiği sevme arasında pozitif korelasyon bulunmuştur. Araştırmaya dâhil edilecek ülkeleri belirlemek amacıyla ülkelerin matematik başarı puanları ile matematiği sevme ölçeğinin ağırlıklandırılmış örneklem ortalamaları alınarak saçılım diyagramı elde edilmiştir. Elde edilen saçılım diyagramı ile ülkeler üç gruba ayrılmıştır. Bu üç gruptan üçer ülke seçilmiş ve araştırmada dokuz ülke yer almıştır.

Bu doğrultuda her bir grup için seçilen ülkeler şu şekildedir:

1.Grup (Matematiğe Yönelik Tutumu Olumlu-Matematik Başarısı Düşük)= 6-Mısır, 20-Fas, 30-Güney Afrika.

2.Grup (Matematiğe Yönelik Tutumu Olumsuz-Matematik Başarısı Yüksek)= 4-Çin, 12-Japonya, 15-Kore Cumhuriyeti.

3.Grup (Matematiğe Yönelik Tutumu Orta-Matematik Başarısı Orta)= 2-Bahreyn, 32-Türkiye, 33-Birleşik Arap Emirlikleri.

Veriler TIMSS 2019 matematiği sevme ölçeği ve matematik başarı testlerinden elde edilmiştir. Verilerin analizlerinde kolaylık sağlaması açısından IDB Analyzer 4.0 programı kullanılarak 5 olası (plausible) değerin ağırlıklandırılmış örneklem ortalaması alınmıştır. Ayrıca matematik başarısı ile matematiği sevme ölçeği arasındaki saçılım grafiğinin oluşturulmasında da SPSS 21.0 programından yararlanılmıştır.

Araştırmada tepki stillerinin ölçme değişmezliği üzerindeki etkisinin incelenmesi amaçlandığı için ilk olarak tepki stillerinin etkisinin düzeltilmediği ham veriler üzerinden ülke düzeyinde ölçme değişmezliğine bakılmıştır. Burada ölçme değişmezliğini incelemek amacıyla Çoklu Grup Doğrulayıcı Faktör Analizi (ÇG-DFA) kullanılmıştır. İkinci aşamada ise TSYDK yöntemi ile tepki stillerinin etkisi düzeltildikten sonra elde edilen veri üzerinden ÇG-DFA aracılığıyla ölçme değişmezliği test edilmiştir.

ÇG-DFA yönteminde, ölçme değişmezliğinin sağlanıp sağlanmadığı belirlemek için iki modelin karşılaştırılmasında CFI ve TLI değerleri arasındaki farklar hesaplanması yöntemi kullanılmıştır. ΔCFI ve ΔTLI değerlerinin -0.01 değerinden düşük ya da 0.01 değerinden yüksek olması ölçme değişmezliğinin sağlanmadığını göstermektedir (Byrne ve diğ., 1989; Wu ve diğ., 2007). ÇG-DFA için LISREL 8.7 programı, TSYDK yöntemi için R programındaki "ccrs" (Takagishi, 2019) paketi kullanılacaktır.

Sonuçlar

Araştırmada verilerin analiz süreci devam etmektedir. Araştırmada kapsamında, tepki stili yanlılığının düzeltildiği veriler üzerinden gerçekleştirilen ölçme değişmezliğinin, tepki stili yanlılığının bulunduğu veriler üzerinden gerçekleştirilen ölçme değişmezliği testlerinden daha iyi sonuçlar vermesi beklenmektedir.

Kaynaklar

Başusta, N. B. ve Gelbal, S. (2015). Gruplar arası karşılaştırmalarda ölçme değişmezliğinin test edilmesi: PISA öğrenci anketi örneği. *Hacetepe Üniversitesi Eğitim Fakültesi Dergisi, 30*(4), 80-90.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425-440.

Büyüköztürk, Ş. (2005). *Sosyal Bilimler için veri analizi el kitabı* (5. baskı). Pagem Akademi.

Byrne, B. M., Shavelson, R. J., and Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466. https://doi.org/10.1037/0033-2909.105.3.456

He, J., & van de Vijver, F. J. (2013). A general response style factor: Evidence from a multi-ethnic study in the Netherlands. *Personality and Individual Differences, 55*(7), 794-800. https://doi.org/10.1016/j.paid.2013.06.017

He, J., & Van de Vijver, F. (2016). Correcting for scale usage differences among Latin American Countries, Portugal, and Spain in PISA. *Electron. J. Educ. Res. Assess. Eval*, *22*(1), 1-12.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20*(3), 296-309. https://doi.org/10.1177/0022022189203004

Karasar, N. (2013). *Bilimsel araştırma yöntemi*. Nobel Akademi.

Lu, Y. (2012). *A multilevel multidimensional ıtem response theory model to address the role of response style on measurement of attitudes in PISA 2006* (Unpuablished Doctoral Dissertation). University of Wisconsin-Madison.

Martin, M. O., Mullis, I. V. S., Hooper, M., Yin, L., Foy, P., & Palazzo, L. (2016). Creating and interpreting the TIMSS 2015 context questionnaire scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 15.1-15.312). http://timss.bc.edu/publications/timss/2015-methods/chapter-15.html

Takagishi, M. (2019). *Ccrs: Correct and cluster response style biased data* (version 0.1.0) [Computer Software]. https://rdrr.io/cran/ccrs/man/correct.rs.html

Van Vlimmeren, E., Moors, G. B., and Gelissen, J. P. (2017). Clusters of cultures: Diversity in meaning of family value and gender role items across Europe. *Quality & Quantity*, *51*(6), 2737-2760.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation, 12*(3), 1-26. https://doi.org/10.7275/mhqa-cd89.

Çok boyutlu BOBUT uygulamaları için parametre kestirim yöntemlerinin karşılaştırılması

F. Gül İnce Aracı, Yıldız Yıldırım ve Tuba Gündüz

Anahtar kelimeler: Bilgisayar ortamında bireye uyarlanmış test, çok boyutlu madde tepki kuramı, parametre kestirim yöntemleri

Giriş

Kâğıt kalem testleri alışılageldik bir metot olarak yaygın şekilde uygulansa da, beraberinde getirdiği bazı dezavantajlar test geliştiricileri farklı arayışlara yöneltmiştir. Son yıllarda uygulanan bazı geniş ölçekli testlerin çevrimiçi (online) şekilde uygulanması bu duruma bir örnektir. Çevrimiçi sınavlar zaman, maliyet, yanıtların internet ortamına kaydedilmesi ve geribildirimin kısa sürede yapılması açısından avantajlıdır (Angus ve Watson 2009; Jordan ve Mitchell, 2009). Ancak kâğıt kalem ve çevrimiçi uygulaması olarak tasarlanan sınavlarda bireylere uygulanan madde sayısı ve sırası aynıdır. Bireylerin yetenek düzeyleri göz önünde bulundurulmamaktadır. Bilgisayar ortamında bireye uyarlanmış test (BOBUT) uygulamaları ise hem çevrimiçi sınavların avantajlarını sağlamakta, hem de bireylerin yetenek düzeylerine göre madde yanıtlamasını sağlamaktadır. Böylece bireyler daha az sayıda madde ve daha az süre kullanarak test uygulamasını tamamlayabilmektedir (Segall, 2005; Weiss ve Gibbons, 2007).

Alan yazına bakıldığında ilk tasarlanan BOBUT uygulamalarının tek boyutlu olarak tasarlandığı, ancak Segall (1996)'in çalışmasında çok boyutlu BOBUT uygulamasının, tek boyutlu uygulamaya göre üç kat daha az madde ile aynı ölçme kesinliğine ulaştığının tespit edilmesinden sonra çok boyutlu BOBUT uygulamalarının ivme kazandığı görülmektedir. Birçok psikolojik yapının çok boyutlu olmasından dolayı, araştırmacılar farklı boyutluluk modellerine göre özelleştirilmiş testler tasarlayabilmektedir.

Araştırmacılar geliştirmek istedikleri BOBUT uygulamalarında farklı kalibrasyon yöntemleri, farklı başlama ve sonlandırma kuralları, farklı madde seçme yöntemleri ve parametre kestirim yöntemi kullanabilmektedir. Asıl uygulamada kullanılacak olan yöntemlerin belirlenmesi için genellikle ön uygulama (pre-study) olarak simülasyon çalışmaları gerçekleştirilmektedir. Böylece BOBUT uygulaması için en güvenilir ve en uygun yöntemler karşılaştırmalı olarak belirlenebilmektedir.

BOBUT uygulamalarında kullanılacak olan yetenek parametresi kestirim yönteminin belirlenmesi oldukça önem taşımaktadır. Bu çalışmanın amacı, çok boyutlu BOBUT uygulamalarında kullanılan parametre kestirim yöntemlerini farklı koşullar altında karşılaştırmaktır. Amaç doğrultusunda farklı boyutluluk modelleri ile farklı korelasyon değerleri, farklı boyut sayısı ve kestirim yöntemi kullanarak çok boyutlu BOBUT uygulamalarından elde edilen istatistikler ve uygulama süreleri karşılaştırılacaktır.

Yöntem

Bu araştırmada, maddeler arası boyutluluk ve madde-içi boyutluluk modellerine göre, Çok Boyutlu 3 Parametreli Lojistik Model ile kalibre edilen simülasyon verisi üretilecektir. 2, 3 ve 5 boyutlu olan farklı yapılara uygun olarak üretilecek veri için, boyutlar arası korelasyonlar, .2, .5 ve .8 olarak belirlenecektir. Başlama kuralı olarak ilk maddeden başlama kuralı; sonlandırma kuralı olarak ise .4 standart hata kuralı uygulanacaktır. Madde seçme kuralı, ön deneme uygulamaları ile belirlenecektir. Araştırmada üç parametre kestirim yöntemine odaklanılmaktadır. Bunlar (1) EAP (Expected a Posteriori), (2) MAP (Maximum a Posteriori, (3) ML (Maksimum Likelihood) yöntemleridir. Çok boyutlu BOBUT uygulamalarında boyutlar arası korelasyonun, testin madde düzeyindeki boyutluluk yapısının ve boyut sayısının 3, 3'ten az ya da 3'ten fazla olmasının uygulama sonuçlarına etkisini irdeleme amacıyla, yürütülen simülasyonlara ait RMSE, bias, üretilen yetenek parametreleri ile simülasyon sonucu elde edilen yetenek parametreleri arasındaki korelasyon değeri ve uygulama süreleri sunulacaktır. Böylece farklı yetenek kestirim yöntemlerine göre test verimlilikleri karşılaştırılacaktır.

Çalışma kapsamında Monte Carlo Simülasyonları gerçekleştirilecek olup, araştırma bir simülasyon çalışması olarak tasarlanmıştır. Çalışma kapsamında üretilecek olan veriler ve gerçekleştirilecek olan analizler için açık kaynak kodlu R (versiyon: 4.0.2) platformunda, mirtCAT (Chalmers, 2016) paketi kullanılacaktır. Belirlenen paket aracılığıyla çok boyutlu yapılara ilişkin veri üretimi ve çok boyutlu BOBUT simülasyon uygulamaları Monte Carlo simülasyonu olarak tasarlanabilmektedir.

Sonuçlar

Bu çalışmada, çok boyutlu yapılar ile gerçekleştirilen BOBUT uygulamaları için üretilen veri setleri ile gerçekleştirilecek olan simülasyonlar aracılığı ile madde içi ve maddeler arası boyutluluk modeline göre, değişen korelasyonlar ve boyut sayısına göre hangi parametre kestirim yönteminin daha verimli ve kullanışlı olduğu irdelenecektir. Madde içi ve maddeler arası boyutluluk modeline göre gerçekleştirilecek simülasyonlarda, boyutlar arası korelasyonun artmasının ML, EAP ve MAP yöntemleri ile gerçekleştirilecek olan çok boyutlu BOBUT uygulamalarından elde edilen RMSE, bias ve $r(\theta_i,\theta_j)$ değerlerini hangi yönde etkilediği sonucuna ulaşılmış olacaktır. Boyut sayısının artması durumunda tüm simülasyonların tamamlanma süresinin artması beklenmekle birlikte, boyutlar arası korelasyon değerinin artması durumunda, sürenin nasıl değiştiği gözlenecektir. EAP, MAP ve ML yöntemleri ile gerçekleştirilecek olan analizlerden elde edilecek sonuçlar doğrultusunda, hangi yöntem ile ölçme kesinliğinin daha yüksek olduğu, yanlılığın daha az olduğu, üretilen parametreler ile hangi yöntemle

kestirilen parametrelerin daha yüksek korelasyon gösterdiği ve süre açısından daha kullanışlı olacağı tespit edilmiş olacaktır.

Kaynaklar

Angus, S. D., & Watson, J. (2009). Does regular online testing enhance student learning in the numerical sciences? Robust evidence from a large data set. *British Journal of Educational Technology*, *40*(2), 255-272. https://doi.org/10.1111/j.14678535.2008.00916.x

Chalmers, R. P. (2015). mirtCAT: Computerized adaptive testing with multidimensional item response theory (version 1.12) [Computer Software]. https://cran.r-project.org/web/packages/mirtCAT/index.html

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software*, *71*(5), 139. https://doi.org/10.18637/jss.v071.i05

Jordan, S., & Mitchell, T. (2009). e-Assessment for learning? The potential of short-answer free-text questions with tailored feedback. *British Journal of Educational Technology*, *40*(2), 371-385. https://doi.org/10.1111/j.1467-8535.2008.00928.x

R Core Team (2020). *R: A language and environment for statistical computing* (version 3.0.1) [Computer Software]. http://www.R-project.org/

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika,* 61, 331-354.

Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement.* Academic Press.

Weiss, D. J., & Gibbons, R. D. (2007). *Computerized adaptive testing with the bifactor model.* Paper presented at the New CAT models session at the 2007 GMAC conference on computerized adaptive testing. https://mail.iacat.org/sites/default/files/biblio/cat07weiss%26gibbons.pdf

Öğretmenlerin çoktan seçmeli soru yazma farkındalık düzeylerine göre üst düzey soru yazma becerilerinin incelenmesi

Sami Sezer Arbağ ve Gül Güler

Giriş

Eğitim-öğretim süreci bir döngü olarak düşünüldüğünde, öğrencinin merkezde olduğu; eğitim-öğretimin hedeflerinin, faaliyetlerinin ve ölçme ve değerlendirme etkinliklerinin niteliği çıktının da niteliğini belirleyen önemli unsurlardandır. Bu bağlamda ölçme ve değerlendirme, eğitim öğretim sürecinin ayrılmaz bir parçasıdır (Popham, 2002).

Günümüz yaşam koşulları bireylerden okuduğunu anlama, problem çözme, eleştirel düşünme, işbirlikli çalışma gibi birçok üst düzey beceriyi beklemektedir. Bu kapsamda okullarda eğitim öğretimin en önemli amaçlarından biri de öğrencileri bu becerileri kazandırıp günlük yaşam koşullarına hazır hale getirmektir. Bu nedenle okul ölçme ve değerlendirme uygulamalarında üst düzey becerilerin ölçülmesi de önemli hale gelmiştir. Alanyazında bu becerilerin ölçülmesinde çoktan seçmeli, açık uçlu maddeler, performans testleri gibi farklı yaklaşımlar önerilmektedir (Scully, 2017; Soland ve diğ., 2013). Sanılanın aksine üst düzey düşünme becerilerinin ölçülmesinde çoktan seçmeli madde formatları sınırlı bir özelliğe sahip değildir (Scully, 2017). Nitelikli hazırlanmış çoktan seçmeli soru formatıyla üst düzey düşünme becerilerini ölçen sorular hazırlanabilir. Alanyazında bu becerilerin ölçülmesinde Bloom başta olmak üzere Haladyna, Marzona Solo gibi çeşitli taksonomilerin kullanımı önerilmektedir (Barnett ve Francis, 2012; Haladyna, 2006; Yüksel, 2007). Taksonomiye dayalı soru sormak bununla birlikte soruları sadece anımsama düzeyinde değil, anlama, eleştirel düşünme yorumlama düzeyinde sormak, sorunun niteliğini önemli ölçüde arttıracaktır. Okul ölçme ve değerlendirme uygulamalarında kullanılan soruların üst düzey düşünme becerilerine dayalı hazırlanmasının yanı sıra kazanımların da dikkate alınması öğrenmelerin niteliği hakkında karar verirken daha geçerli ve güvenilir sonuçlara dayanak oluşturacaktır.

Bu araştırmanın amacı, farklı branşlardaki öğretmenlerin öğretimin farklı kademelerindeki öğrenci başarısını ölçmek için hazırladığı soruları çoktan seçmeli soru yazma ölçütlerine göre incelemektir. Bununla birlikte öğretmenlerin test geliştirme ve soru yazma deneyimleri ve farkındalık düzeylerine göre ilgili dersin kazanımını dikkate alıp çoktan seçmeli soru yazma becerilerinin incelenmesi amaçlanmıştır. Bu amaç doğrultusunda aşağıdaki sorulara cevap aranmıştır:

1. Öğretmenlerin çoktan seçmeli soru yazma eğitimi alma durumuna göre soru yazma farkındalık puanları farklılık göstermekte midir?
2. Öğretmenlerin aldıkları çoktan seçmeli soru yazma eğitimi sayısına göre soru yazma farkındalık puan ortalamaları nasıldır?
3. Öğretmenler aldıkları çoktan seçmeli soru yazma eğitimi sayısına göre kategorilere ayrıldığında düşük, orta ve yüksek farkındalık düzeyine sahip öğretmenlerin soru yazma becerileri nasıldır?

Yöntem

Öğretmenlerin çoktan seçmeli soru yazma farkındalık düzeylerine göre çoktan seçmeli soru yazma eğitimi alma durumu değişkeni açısından ve öğretmenlerin soru yazma becerilerinin incelendiği bu araştırmada karma araştırma yöntemi kullanılmıştır. Karma yöntem araştırması, araştırma probleminin kapsamlı ve çok boyutlu incelenmesi amacıyla nitel ve nicel yöntemlerin birlikte kullanıldığı araştırma türüdür (Yıldırım ve Şimşek, 2013).

Karma yöntem araştırması modeli olarak, bu çalışmanın amacına en uygun olduğu belirlenen açıklayıcı ardışık desen tercih edilmiştir. Açıklayıcı ardışık desende araştırma, nicel aşama ile başlayıp ikinci aşamada nicel sonuçları açıklamak için nitel çalışma ile devam etmektedir (Creswell ve Plano Clark, 2007).

Açıklayıcı ardışık desen modelinin ilk aşaması olan nicel kısmında öğretmenlerin soru yazma farkındalık düzeylerinin çoktan seçmeli soru yazma eğitimi alma durumu değişkeni açısından karşılaştırması yapıldığı için ilişkisel tarama modeli kullanılmıştır. Nicel araştırma süreci sonunda öğretmenlerin soru yazma farkındalık düzeyleri belirlendikten sonra öğretmenler soru yazma eğitim sayılarına göre 3 kategoriye ayrılmıştır. Bu kategorideki görev yapan 97 öğretmene "Kişisel Bilgi Formu" ve "Çoktan Seçmeli Soru Yazma Farkındalık Testi" uygulanmış, elde edilen verilen sonucunda örnekleme yöntemlerinden maksimum çeşitlilik örnekleme yoluyla belirlenen 18 öğretmen ile 2'şer adet soru yazma çalışması yapılmıştır. Yazılan sorular araştırmacılar tarafından geliştirilen "Soru Yazma Değerlendirme Formu" ile değerlendirilmiştir. Bu formun, çalışmanın amacına uygun olup olmadığını ortaya koymak amacıyla uzman görüşüne başvurulmuştur. Uzmanlardan alınan düzeltmeler ve araştırmacıların soru inceleme deneyimleri doğrultusunda forma son hali verilmiştir. Öğretmenler tarafından yazılan 36 soru iki araştırmacı tarafından form doğrultusunda incelenmiş olup puanlayıcılar arası tutarlık çalışması yapılmıştır. "Çoktan Seçmeli Soru Yazma Farkındalık Testi" çoktan seçmeli 12 sorudan oluşmaktadır. Bu testten alınabilecek en yüksek puan 12, en düşük puan ise 0'dır. Bu testten elde edilen puanların güvenirlik katsayısı 0.84 olarak hesaplanmıştır.

Nicel verilerin çözümlenmesinde frekans, yüzde ve aritmetik ortalama değerlerinden, araştırma kapsamındaki değişkenin öğretmenlerin soru yazma farkındalık puanlarını nasıl farklılaştığını belirlemek için bağımsız örneklemler t-testi analizinden yararlanılmıştır. Bu analiz .05 anlamlılık düzeyinde gerçekleştirilmiştir. Nitel veriler ise betimsel analiz yöntemi ile analiz edilmiştir.

Sonuçlar

Öğretmenlerin %57'si soru yazma ya da redaksiyon eğitimlerinin hiçbirine, %26'sı bir ya da ikisine, %17'si ise tamamına katılmıştır. Hiçbirine katılım göstermeyen öğretmenlerin (7.23) çoktan seçmeli soru yazma farkındalık puan ortalamaları tamamına (8.23) ve bir ya da ikisine (7.80) katılanlara göre daha düşük olduğu belirlenmiştir.

Öğretmenlerin soru yazma farkındalık puanları çoktan seçmeli soru yazma eğitimi alma durumuna göre anlamlı farklılık gösterdiği görülmüştür (t= -2.22, p< .05). Soru yazma eğitimi alan öğretmenlerin soru yazma farkındalık puan ortalamaları (8.32) almayanlara (7.29) göre daha yüksek olduğu belirlenmiştir.

Eğitimlerin tamamını alan öğretmenler, eğitimlerin bir ya da ikisini ve hiçbirini almayan öğretmenlere göre sorunun gündelik yaşamla ilişkilendirme ve üst düzey düşünme becerilerini ölçme, Haladyna taksonomisine ve MEB kazanımlarıyla eşleştirme, ölçme değerlendirme tekniğine ve Türkçe dilbilgisi kurallarına uygunluğu bakımından daha başarılı olduğu tespit edilmiştir.

Eğitimlerin tamamını alan öğretmenlerin daha çok problem çözme ve eleştirel düşünme becerilerini ölçen sorular yazma eğiliminde iken diğer öğretmenler daha çok anımsama, anlama ve problem çözme becerilerini ölçen sorular yazma eğiliminde oldukları tespit edilmiştir.

En yüksek soru yazma farkındalık puanına sahip olup eğitim alan ve almayan öğretmenlerin, orta ve en düşük soru yazma farkındalık puanına sahip eğitim alan ve almayan öğretmenlere göre soru yazma süreçlerinde daha başarılı oldukları tespit edilmiştir.

Kaynaklar

Barnett, J. E., and Francis, A.L. (2012). Using higher order thinking questions to foster critical thinking: A classroom study. *Educational Psychology, 32*(2) 201-211. https://doi.org/10.1080/01443410.2011.638619

Creswell, J. W., and Plano Clark, V. L. (2007). *Designing and conducting mixed methods research* (1st ed.). Sage Publications Ltd.

Popham, W. J. (2002). *Classroom assessment: What teachers need to know?* Allyn and Bacon.

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research, and Evaluation*, Vol. 22, Article 4. https://doi.org/10.7275/swgt-rj52

Soland, J., Hamilton, L. S., and Stecher, B. M. (2013). *Measuring 21st century competencies guidance for educators.* RAND Corporation.

Yıldırım, A. ve Şimşek, H. (2013). *Sosyal bilimlerde nitel araştırma yöntemleri.* Seçkin Yayıncılık.

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Derneği
www.epodder.org
CMEEP 2020

Fleiss kappa ve Krippendorf alfa katsayılarının örneklem büyüklüğü, örneklemden seçim oranı, uzlaşma oranı ve puanlayıcı sayısı koşulları altında incelenmesi: Bir simülasyon çalışması

Sibel Ada

Anahtar kelimeler: Fleiss kappa, Krippendorff alfa, puanlayıcılar arası güvenirlik

Giriş

Güvenirlik tesadüf hataların sebeplerine göre duyarlılık, kararlılık ve tutarlılık anlamlarında kullanılır (Baykul, 2010). Performansa dayalı ölçümlerde gözlemler her bir birey için iki ya da daha fazla puanlayıcı tarafından değerlendirilir ve böyle durumlarda elde edilen gözlemlerin tutarlılığı ile ilgilenilir (Crocker ve Algina, 2006). Sınıf içi değerlendirmelerde ve dereceli puanlama anahtarı (rubrik) kullanılan değerlendirmelerde genel olarak dikkate alınan güvenirlik puanlayıcı güvenirliğidir. Puanlayıcılar arası güvenirlik öğrenci puanının puanlayıcıdan puanlayıcıya değişip değişmediğine odaklanır ve puanlarda değerlendiricinin subjektif yargıları olabileceği düşünülür (Moskal ve Leydens, 2000). Puanlayıcılar arası güvenirlik, iki veya daha fazla puanlayıcının belirli bir ölçüm ile ilgili uzlaşmasının ya da tutarlılığının derecesini gösterir (Cohen ve Swerdlik, 2013). Puanlayıcılar arası güvenirlik terimine alternatif terimler olarak gözlemciler arası (interobserver/inter-observer), yargılayıcılar arası (interjudge/inter-judge), kodlayıcılar arası (intercoder/inter-coder) ifadeleri kullanılır ve hepsi aynı nitelikte ölçümleri ifade eder (Gwet, 2001). Puanlayıcılar arası güvenirliğin kabul edilebilir bir düzeyine ulaşması puanlama planının temel doğasını oluşturma ve çoklu puanlayıcı kullanımının tasarlanması yönüyle önemlidir (Neundorf 2002). Alan yazın incelendiğinde puanlayıcılar arası güvenirlik katsayılarının farklı örneklem büyüklüğü (gözlem sayısı), puanlanan ölçek kategorisi, kayıp veri oranı, puanlayıcı sayısı, uzlaşma oranı koşulları altında inceleyen çalışmalar bulunmaktadır (örneğin Cicchetti ve diğ., 1985; Nying, 2004; Sullivan, 2014; Temel ve diğ., 2016; Zapf ve diğ., 2016). Bu çalışmanın amacı örneklem büyüklüğü, örneklemden seçim oranı, uzlaşma oranı ve puanlayıcı sayısı koşulları altında puanlayıcılar arası güvenirliği belirlemede yaygın olarak kullanılan Fleiss Kappa ve Krippendorf alfa katsayıları incelemektir. Örneklem seçim oranı bir çalışmada puanlayıcılar arası güvenirlik hesaplamak için örneklemin (yapılan gözlem veya puanlanan olayların) tamamının iki ya da daha fazla puanlayıcı tarafından puanlanmasın mümkün olmadığı durumda örneklemin ne kadarlık kısmının güvenirlik çalışması için seçilmesi gerektiği ifadesi için kullanılmaktadır. Araştırmacılar çoğu zaman bu durumla bir sorun olarak karşılaşmaktadır. Fan ve Chen,

(2000) araştırmalarda zaman, maddi olanaklar vb. kaynakların eksikliği nedeniyle araştırmacıların örneklemin yalnızca küçük bir bölümünden puanlayıcılar arası güvenirlik tahmini elde etmenin yaygın olduğunu ve toplam örneklemin veya gözlem oturumlarının yalnızca %10-%15'ini kullanarak iki bağımsız değerlendirici tarafından puanlanması ile puanlayıcılar arası güvenirlik tahmini yapan çalışmalar ile alan yazında karşılaşıldığını belirtmektedir. Bu araştırma puanlayıcılar arası güvenirliğin kabul edilebilir bir düzeyine ulaşmada çalışma planını oluşturmada araştırmacılara yol göstermesi yönüyle önemli görülmektedir.

Yöntem

Araştırma farklı koşullar altında Fleiss Kappa ve Krippendorf alfa katsayılarının incelenmesine yönelik bir simülasyon çalışmasıdır. Simülasyon koşulları olarak dört farklı örneklem büyüklüğü/gözlem sayısı (30: 50: 100: 200), dört farklı örneklemden seçim oranı (%10: %20: %25: %50), üç farklı uzlaşma oranı (düşük (0,3): orta(0,6): yüksek (0,8)) ve iki farklı puanlayıcı sayısı (2:3) alınmıştır. Replikasyon sayısı ise her bir koşul için 100 olarak alınmış ve çalışma kapsamında toplam 9600 (4x4x3x2x100) analiz yapılmıştır. Farklı koşullar altında Fleiss Kappa ve Krippendorf alfa katsayılarının karşılaştırılması iki yönlü karma ANOVA ile değerlendirilmiştir. Simülasyon çalışmasında ve verilerin analizinde r programı RStudio ara yüzü kullanılmıştır. Çalışmada "jbryer/IRRsim" ve "irr" paketleri kullanılmıştır.

Sonuçlar

Yapılan analizler sonucunda puanlayıcı sayısı ile kullanılan puanlayıcılar arası güvenirlik katsayısı (Fleiss Kappa ve Krippendorf alfa), örneklem büyüklüğü ile kullanılan puanlayıcılar arası güvenirlik katsayısı ve örneklem seçim oranı ile kullanılan puanlayıcılar arası güvenirlik katsayısı ortak etkilerindeki farklılaşmanın düşük etki büyüklüğü ile manidar bir farklılık gösterdiği belirlenmiştir. İki puanlayıcının olduğu durumda Fleiss Kappa ve Krippendorf alfa katsayıları arasındaki farkın üç puanlayıcı olduğu duruma göre bir miktar daha fazla olduğu tespit edilmiştir. Örneklem büyüklüğü 30 olduğunda Krippendorf alfa katsayısının Fleiss Kappa katsayısından bir miktar fazla iken örneklem artıkça aradaki farkın azaldığı ve örneklem 200 olduğunda Fleiss kappa ve Krippendorf alfa katsayılarının neredeyse aynı değere sahip olduğu belirlenmiştir. Benzer şekilde örneklem seçim oranı %10 olduğunda Krippendorf alfa katsayısının Fleiss kappa katsayısından bir miktar fazla olduğu, örneklem seçim oranı artıkça aradaki farkın azaldığı ve örneklemin tamamında Fleiss kappa ve Krippendorf alfa katsayılarının neredeyse aynı değere sahip olduğu ortaya konulmuştur. Ayrıca Fleiss kappa ve Krippendorf alfa için ayrı ayrı seçilen örneklem oranına göre örneklemin tamamı ile farklılaşma durumu incelenmiştir. Krippendorff alfa katsayısı için örneklemin (gözlem biriminin) %10, %20, %25, %50 seçilerek hesaplanan değer ile örneklemin tamımı için hesaplanan değerler arasında manidar farklılık olmadığı belirlenmiştir. Fleiss kappa katsayısı için ise örneklemin (gözlem biriminin) %10, %20 seçilerek hesaplanan değer ile örneklemin tamamı için hesaplanan değerin manidar farklılık gösterdiği belirlenmiştir. Örneklemin (gözlem biriminin) %10, %20 olduğu durumlarda Fleiss kappa değeri örneklemin tamamı için hesaplanan değerden daha düşük kestirilmiştir.

Kaynaklar

Baykul, Y. (2010). *Eğitimde ve psikolojide ölçme: klasik test teorisi ve uygulaması* (2. Baskı). Pegem Akademi.

Cicchetti, D. V., Shoinralter, D., and Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of interrater reliability: A Monte Carlo investigation. *Applied Psychological Measurement*, *9*(1), 31-36. https://doi.org/10.1177/014662168500900103

Cohen, R. J., and Swerdlik, M. E. (2013). Güvenirlik (G. Gözen, Çev). E. Tavşancıl (Çev. Ed.), *Psikolojik test ve değerleme testlere ve ölçmeye giriş* içinde (ss. 139-171). Nobel Akademik Yayıncılık. (2010).

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. United States of America.

Fan, X., and Chen, M. (2000). Published studies of interrater reliability often overestimate reliability: Computing the correct coefficient. *Educational and Psychological Measurement*, *60*(4), 532-542. https://doi.org/10.1177/00131640021970709

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development, validity and reliability. *Pratical Assessment, Research and Evaluation, 7*(10), 1-6. https://doi.org/10.7275/q7rm-gg74

Neuendorf, K. A. (2002). *Content analysis guidebook*. Thousand Oaks, CA: Sage.

Nying, E. (2004). *A comparative study of interrater reliability coefficients obtained from different statistical procedures using monte carlo simulation techniques* (Publication No. 3138768) [Doctoral Dissertation, Western Michigan University] Proquest Dissertations and Theses database.

Sullivan, A. D. (2014). Determining an inter-rater agreement metric for researchers evaluating student pathways in problem solving (Unpublished Master Thesis) (Publication No. 1560361) [Master Dissertation, Iowa State University] Proquest Dissertations and Theses database.

Temel, G. O., Erdogan, S., Selvi, H., & Kaya, I. E. (2016). Investigation of coefficient of individual agreement in terms of sample size, random and monotone missing ratio, and number of repeated measures. *Educational Sciences: Theory and Practice, 16*(4), 1381-1395. https://doi.org/10.12738/estp.2016.4.0080

Zapf, A., Castell, S., Morawietz, L., & Karch, A. (2016). Measuring inter-rater reliability for nominal data–which coefficients and confidence intervals are appropriate? *BMC medical research methodology*, *16*(1), 1-10. https://doi.org/10.1186/s12874-016-0200-9

Araştırma etiği konusunda lisansüstü öğrencilerin görüşlerinin incelenmesi

Sibel Ada

Anahtar kelimeler: Etik, araştırma etiği, etik dışı davranışlar

Giriş

Bilim insanında bulunması gereken genel özellikler bilme arzusu ve merak, gerçekleri sezme ve algılama gücü ve yaratıcılık olarak sıralanabilir. Bilimde kuşkuculuk ise çok önemli ve elden bırakılmaması gereken bir özeliktir. Bu sayede hem yeni bilgilerin açılımı sağlanır hem de başkaları tarafından yapılmış araştırmalardaki yanlışlıklar, ihmaller ve etik dışı davranışlar belirlenebilir (Ertekin ve diğ., 2002). Etik, Türk Dil Kurumu (2015) tarafından 'çeşitli meslek kolları arasında tarafların uyması ya da kaçınması gereken davranışlar bütünü' olarak tanımlanmaktadır. Etik genel olarak, neyin doğru neyin yanlış olduğunu ayırmaya yarayan davranış kuralları olarak tanımlanabilir (Aypay, 2014). Etiğin farklı tanımları bulunmak ile birlikte genel olarak dört farklı başlık altında toplanabilir: (a) Etik, doğru-yanlış, iyi-kötü vb. durumlar arasındaki farkı ayırt eden davranış standartlarıdır. (b) Etik, davranış standartlarının çalışıldığı akademik bir disiplindir. (c) Etik, karar verme sürecinde bir yaklaşımdır. (d) Etik, karakterin özel bir halidir. Etik; görev, onur, doğruluk, erdem, adalet ve iyi yaşam konuları hakkında çok eski sorunları cevaplamak ile ilgilenen bir disiplindir (Shamoo & Resnick, 2009; s.14). Bu çalışmanın amacı lisansüstü öğrencilerin araştırma etiği konusundaki görüşlerini belirlemektir. Akademik alana adım atmış olan ve akademik yayınlar yapması beklenen lisansüstü öğrencilerinin araştırma etiği konusundaki görüşleri daha güvenilir ve geçerli çalışmalar elde edilmesi açısından önemlidir. Ayrıca araştırma sürecinde etik davranışlara uygun davranmak başkalarının haklarına saygı gösterildiğini, doğruluk ve dürüstlük ilkeleri çerçevesinde çalışıldığını gösterir. Çalışma lisansüstü öğrencilerin araştırma etiği konusundaki görüşlerinin belirlenmesi, mevcut durum hakkında bilgi vermesi, bireylere nasıl ve ne zaman araştırma etiği eğitiminin verilmesi gerektiği gibi konularda alan yazına ışık tutması yönüyle önemlidir.

Yöntem

Araştırma hem nitel hem nicel verilerin kullanılması yönüyle bir karma yöntem çalışmasıdır. Kullanılan karma yöntem, nitel ve nicel verilerin eş zamanlı olarak kullanıldığı zenginleştirilmiş karma

yöntem şeklindedir. Araştırmada öncelikle araştırma görevlilerin araştırma etiği konusuna yönelik düşüncelerini ortaya çıkarmaya yönelik açık uçlu soruların bulunduğu görüş formu uygulanmıştır. Bu yönüyle araştırmanın nitel kısmı olgubilim çalışması niteliğindedir. Araştırma görevlilerin araştırma etiğine yönelik görüşlerini belirtmesinin ardından yarı-yapılandırılmış sorulardan oluşan anket formu uygulanmıştır. Bu yönüyle araştırmanın nicel kısmı tarama türündedir. Çalışma grubu 55'i kadın, 40'ı erkek olmak üzere toplam 95 lisansüstü öğrenciden oluşmaktadır. Katılımcıların 20'si hem görüş formunu hem anketi 75'si ise sadece anketi doldurmuştur. Veriler araştırmacı tarafından oluşturulan "Araştırma Etiği Görüş Formu" ve "Araştırma Etiği Anketi" yardımıyla lisansüstü öğrencilerden toplanmıştır. Ölçme araçları uzman görüşü alınarak araştırmacı tarafından geliştirilmiştir. Beş farklı uzmandan görüş alınmış ve hesaplanan kapsam geçerlik indeksi Araştırma Etiği Görüş Formu için 0,94 ve Araştırma Etiği Anketi için 1,00 olarak elde edilmiştir. Görüş formu gönderilecek katılımcılar ile önceden görüşülmüş, araştırmanın amacından bahsedilmiştir. Bireylere görüş formu gönderildikten sonra anket formu gönderilmiştir. Her iki formun eş zamanlı olarak gönderilmemesinin nedeni katılımcıların anket formundaki yapılandırılmış maddelerden etkilenerek görüş formundaki soruları cevaplarken görüşlerini doğru bir şekilde yansıtamayacaklarının düşünülmesidir. Hem görüş formu hem anket mail yoluyla katılımcılara ulaştırılmıştır. Sadece anket formu uygulanan katılımcılara ise sanal olarak anket formu oluşturulup mail yoluyla bağlantı gönderilerek ulaşılmıştır. Nitel veriler içerik analizi ile incelenmiştir. Her bir katılımcının cevapları soru bazında tablolaştırılmış, birimlere ayrılmış ve birimlere uygun kategoriler belirlenmiştir. Nicel verilerin analizinde ise frekans ve yüzde kullanılmıştır. Yapılan içerik analizinden elde edilen verilerin güvenirliği için kodlayıcılar arası güvenirliğe başvurulmuştur. Seçilen bir soru üzerinden farklı bir kodlayıcıya birimler ve kategori isimleri verilerek her birimin hangi kategoriye uygun olduğunu belirlemesi istenmiştir. Kodlayıcı ve araştırmacı arasındaki uzlaşma yüksek düzeyde (Krippendorff alfa= .83; Uzlaşma yüzdesi= 0.85) bulunmuştur ve bu durum yapılan kodlama işleminin güvenilir olduğunu gösterdiği kabul edilmiştir.

Sonuçlar

Lisansüstü öğrencileri etik ve araştırma etiği kavramını daha çok ahlakî/toplumsal değerler, özgünlük ve evrensel değerler olarak görmektedir. Aynı zamanda araştırma etiği insan haklarını koruma ve güvenilir-geçerli sonuçlar üretme yönüyle önemli olarak belirtilmiştir. Lisansüstü öğrenciler etik dışı davranışlara örnek olarak daha çok intihal, sahtecilik, gizliliğin ihlali, kopyalama ve deneklere zarar verme durumlarını belirtmişlerdir. Etik dışı davranış göstermenin nedeni olarak ise unvan sahibi olabilme hırsı, hızlı yükselme hırsı, prestij elde etme, mali kazanç elde etme hırsı ve yetersiz araştırma eğitimi gösterilmiştir. Etik dışı davranışların önlenmesi için ise eğitim verilmesi ve yasal düzenlemelerin artırılması ve öğretilmesi önerilmektedir. Lisansüstü öğrenciler, sadece insanlık için çok faydalı olacağı düşünülen ve yapılmasının çok fayda sağlayacağı durumlar için etik dışı davranışların gösterilmesinin göz ardı edilebileceğini söylemekle birlikte, mümkün olduğu kadar etik değerler çerçevesinde hareket edilmesi gerektiğini düşünmektedir. Lisansüstü öğrenciler eğitim yaşamlarını değerlendirdiklerin de çok yeterli bir araştırma etiği eğitimi almadıklarını düşünmektedir. Araştırma etiği konusunda kendilerini

yetersiz bulan lisansüstü öğrencileri araştırma etiği konusunun ayrı bir ders olarak verilmesi gerektiğini belirtmişlerdir. Ayrıca etik konusunun bir karakter özelliği olması gerektiğini düşünmekle birlikte araştırma süreci ve uyulması gereken kurallar konusunda eğitimin ilkokul yıllarından itibaren verilmesi uygun bulunmaktadır.

Kaynaklar

Ertekin, C., Berker, N., Tolun, A. ve Ülkü, D. (2002). *Bilimsel araştırmada etik ve sorunlar*. TÜBA Yayınları.

Türk Dil Kurumu (2015). http://www.tdk.gov.tr/index.php?option=com_bts&arama=kelime&guid=TDK.GTS.56489c8e6ca514.99718376 adresinden 15.11.2015 tarihinde alınmıştır.

Aypay, A. (2014). Bilimsel etik. A. Tanrıöğen (Ed*.), Bilimsel araştırma yöntemleri* içinde (4. baskı). Anı Yayıncılık.

Shamoo, A., and Resnick, D. (2009). *Responsible conduct of research* (2nd ed.). Oxford University Press.

Farklı kayıp veri mekanizmalarının üç adımlı en çok olabilirlik örtük sınıf analizine olan etkilerinin incelenmesi

Ömer Emre Can Alagöz

Anahtar kelimeler: Örtük sınıf analizi, kayıp veri, dışsal değişken, üç adımlı yaklaşım, iki adımlı yaklaşım, tek adımlı yaklaşım

Giriş

Örtük sınıf analizi (ÖSA) kullanılarak, bireyler, belirli sayıdaki gözlenen kategorik değişkene (gösterge değişken) verdikleri cevaplara göre küçük ve anlamlı örtük sınıflara ayrılabilirler (Lazarsfield ve Henry, 1968; Goodman, 1974). Her ne kadar bu sınıfları bulmak yeterli olabilse de kimi araştırmacılar bu ortaya çıkan sınıfları açıklayan yapıları da incelemek istemektedir. Bu durumda, dışsal değişkenlerin örtük sınıflar üzerindeki etkisini incelemek için tek adımlı, iki adımlı (Bakk ve Kuha, 2018) veya üç adımlı en çok olabilirlik (3A-EO; three-step Maximum Likelihood; Vermunt, 2010) ÖSA kullanılabilmektedir. Tek adımlı ÖSA'da dışsal değişkenin sınıflara olan etkisi tek bir ÖSA ile modellenir ve bu model tam bilgi en çok olabilirlik (TBEO) yöntemi ile tahminlenir.

Vermunt (2010)'un ortaya koyduğu 3A-EO ÖSA'daki adımlar ise şu şekildedir: (1) Bireylerin gösterge değişkenlere verdikleri cevaplara göre örtük sınıf parametreleri bulunur; (2) Birinci adımda hesaplanan bireylerin sonsal sınıf üyelik olasılıklarına göre her birey bir sınıfa atanır. Sonsal sınıf üyelik olasılıkları genellikle 1 olmadığı için bu adımda yapılan sınıflama hataları hesaplanır; (3) Dışsal değişkenlerin örtük sınıflara olan etkisi, yapılan sınıflama hatalarını hesaba katarak tahminlenir. Eğer ikinci adımdaki sınıflama hataları hesaba katılmazsa (standart 3A ÖSA) üçüncü adımdaki dışsal değişken etkileri gerçekte olduğundan daha düşük bir şekilde tahminlenmektedir (Bakk ve diğ., 2013; Bolck ve diğ., 2004; Vermunt, 2010).

Görgül çalışmalarda bazı katılımcılar, gözlenen değişkenlerde kayıp veriye sahip olabilir. Bu kayıp verilerin altında yatan mekanizma şunlar olabilir: tümüyle seçkisiz kayıp (missing completely at random; MCAR), seçkisiz kayıp (missing at random; MAR), seçkisiz olmayan kayıp (missing not at random; MNAR). MCAR mekanizması, kayıp verilerin sistematik olmayan ve rastgele bir şekilde meydana geldiğini ifade eder (örn. korna sesinden dikkatin dağılıp maaş sorusunun atlanması). MAR mekanizması, kayıp verinin bir örüntüye sahip olduğunu ve bu örüntünün veri setindeki diğer değişkenler tarafından açıklanabildiğini ifade eder (örn. yaşı 50'den büyük olanların maaş sorusunu

cevapsız bırakması). MNAR mekanizması ise kayıp verinin sebebinin o kayıp verinin gözlendiği değişkenin kendisi olduğu durumları ifade eder (örn. maaşı 2000 TL'den daha az olanların maaş değişkenine cevap vermemesi). Mekanizması MCAR veya MAR olan kayıp veriler söz konusu olduğunda, bir en çok olabilirlik yöntemi olan tek adımlı TBEO ÖSA kullanılarak örtük sınıflar ve onları açıklayan dışsal değişkenlere ait parametreler sorunsuzca tahminlenebilir, ancak NMAR mekanizmasında yanlı parametre tahminlemeleri ortaya çıkmaktadır (Buuren, 2008; Dong ve Peng, 2013; Schafer, 1997).

Her ne kadar 3A-EO ÖSA'nın tahminlenmesi hem birinci hem üçüncü adımda en çok olabilirlik tahminlemesiyle yapılsa da bu tahminleme yönteminin kayıp verilerle olan başarılı çözümü üç adımlı ÖSA'ya genellenemez. Bunun sebebi, (daha çok) kayıp veriye sahip olan bireylerin (daha yüksek) sınıflama hatasına sahip olması ancak ikinci adımda yapılan hata hesaplamasının birey bazında ve kayıp veriler dikkate alınarak yapılmamasıdır. Bu durumda, üçüncü adımdaki tahminlemede bu hatalar hesaba katılırken kayıp veriye sahip olan kişiler için yapılan hata düzeltmesi gerçekte olduğundan daha az olabilir. Bu da dışsal değişkenlerin örtük sınıflara olan etkisinin yanlı bir şekilde tahminlenmesine sebep olabilir.

Bu düzeltme yönteminin kayıp veriler olduğunda uygulanmasının doğruluğu, dolayısıyla her bir kayıp veri yanıt örüntüsü için ayrı bir sınıflama hatası hesaplanması gerekliliği incelenmemiştir. Bu araştırmada, gösterge değişkenlerdeki kayıp verilerin (her üç kayıp veri mekanizması altında) üçüncü adımdaki tahminlenen parametrelere bir etki edip etmediği bir dizi koşul altında yapılan simülasyon çalışmasıyla incelenmiştir. Ayrıca, sınıflama adımına sahip olmayan, dolayısıyla sınıflama hatasının hesaplanmasının gerekmediği yöntemlerle (tek adımlı ÖSA, iki adımlı ÖSA, 3A-EO yöntemi de karşılaştırılmıştır.

Yöntem

Yapılan simülasyon çalışmasında, kayıp verilerin tek, iki ve üç adımlı ÖSA'da dışsal değişkenlerin parametre tahminlemelerine olan etkisi incelenmiştir. Eğer bir yöntemde tahminlemeler düşük yanlılık içeriyorsa ve replikasyonlar arasında düşük varyasyona sahipse o yöntem "iyi çalışan" bir yöntem olarak değerlendirilmektedir.

Yukardaki yöntemlerin yeterliliği dört faktör manipüle edilerek incelenmiştir: örneklem büyüklüğü (500, 1000, 10.000), sınıf ayrımı (düşük, orta, yüksek), kayıp veri mekanizması (MCAR, MAR, MNAR) ve kayıp verilerin analize dahil edilmesi (dahil, liste-bazı-silinik). Örneklem büyüklüğü arttıkça eldeki bilgi miktarı artacaktır ve hem birinci adımda sınıf parametlerinin hem de üçüncü adımda dışsal değişken etkilerinin tahminlenmesi iyileşecektir. Sınıf ayrımı, örtük sınıfların birbirinden ayırt edilebilmesi anlamına gelmektedir ve bu ayrım arttıkça bireyler daha yüksek bir eminlikle sınıflara atanacak, sınıflama hatası azalacak ve dışsal değişken etkilerinin tahminlemeleri iyileşecektir. MCAR ve MAR en çok olabilirlik yöntemlerinde "göz ardı edilebilir" (Little ve Rubin, 1987) olarak değerlendirilmiştir ve parametre tahminlemesine bir etkisi beklenmemektedir. Ancak, sınıf ayrımı örneklem sayısı azaldıkça yöntemlerin kayıp verilere toleransı azalabilir. NMAR mekanizmaları ise "göz

ardı edilemez" olarak değerlendirilmektedir ve yanlılık katmaları beklenmektedir. Kayıp verilerin üçüncü adımda liste-bazı silinmesi hem örneklem büyüklüğünü azaltacağı hem de MAR ve MNAR koşullarında bir grup kişiyi sistematik olarak analiz dışı bırakacağı için dışsal değişken etkilerinin tahminlemelerine yanlılık katacaktır.

Simüle edilen popülasyon modelinde üç sınıf ve altı kategorik gösterge değişken vardır. Bu göstergelerin ikisinde kayıp veriler belirli mekanizmalara göre üretilmiştir. Modelde bulunan üç kategorik dışsal değişkense bireylerin sınıf üyeliklerini açıklamaktadır. Kayıp veriler iki gösterge değişkende üretilmiştir ve bir göstergenin %24'ü diğerinin %16'sı silinmiştir. MCAR mekanizmasında silinecek gözlemler üniform dağılımla rastgele seçilmiştir. MAR mekanizmasında silinecek gözlemler diğer göstergelerdeki yanıtlara göre seçilmiştir. İncelenen iki MNAR koşulunun ilkinde, bir verinin "kayıp" olmasını bireyin bulunduğu örtük sınıf etkilerken diğerinde dışsal değişkenlere verdiği yanıtlar etkilemektedir. Her koşulda 100 replikasyon olmak üzere toplamda 5400 veri seti üretilmiş ve Latent Gold 5.1 (Vermunt ve Magidson, 2016) ile analiz edilmiştir.

Sonuçlar

Bulgulara göre 3A-EO ÖSA, kayıp veri mekanizması MCAR ve MAR olduğunda dışsal değişkenin örtük sınıflar üzerindeki etkileri yansız ve isabetli standart hata ile tahminlenebilmektedir. 3A-EO'nun tek ve iki adımlı ÖSA'dan farkları incelenmiştir. Bulgular, her kayıp veri mekanizması altında 3A-EO ve iki adımlı ÖSA'daki yanlılığın ve standart hatanın tek adımlı ÖSA'dan marjinal düzeyde daha yüksek olduğu gözlenmiştir, ancak bu kayıp verilerden bağımsız olarak literatürde gösterilmiş olan bir farklılıktır (Vermunt, 2010). Sonuç olarak, halihazırda kullanılan sınıflama hatalarını düzeltme yöntemi (bkz., EO), kayıp verilerin sınıflama hatalarını arttırdığı durumlarda dahi etkili bir şekilde çalışmaktadır ve üçüncü adımda yansız ve tutarlı tahminlemeler elde edilmektedir.

Kaynaklar

Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological methodology*, *43*(1), 272-311. https://doi.org/10.1177/0081175012470644

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, *2*(1), 3-27. https://www.jstor.org/stable/25791751

Dong, Y., & Peng, C. Y. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1), 222. https://doi.org/10.1186/2193-1801-2-222

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*(2), 215-231. https://www.jstor.org/stable/2334349

Lazarsfeld, P. F., & Henry, N. W. (1968) *Latent structure analysis.* Houghton Mifflin.

Little, R. J. A., and Rubin, D. B. (1987). *Statistical analysis with missing data.* John Wiley & Sons.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* Chapman and Hall/CRC.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political analysis*, *18*(4), 450-469. https://doi.org/10.1093/pan/mpq025

Vermunt, J. K., & Magidson, J. (2016). *Technical guide for Latent GOLD 5.1: Basic, advanced, and syntax*. Belmont: Statistical Innovations Inc.

Kanada eğitim kalite ve hesapverebilirlik ofisi uygulamalı matematik değerlendirmesi delil geçerliliğine bir örnek: Lord'un ki kare yöntemi ile değişen madde fonksiyonu bulgusu hesaplaması[1]

Nazlı Uygun Emil

Anahtar kelimeler: Matematik başarı testi, delil geçerliği, değişen madde fonksiyonu, Lord ki kare yöntemi, cinsiyet eşitliği

Giriş

Bu araştırmanın amacı Eğitim Kalite ve Hesapverebilirlik Ofisi (EKHO)'nce uygulanan dokuzuncu sınıf matematik başarı testinin delil geçerliği tespitinin öğrenci gruplarının kıyaslanması yöntemi ile bulgulanmasıdır. Messick (1989) test geçerliği çalışmalarını delilsel geçerlik ve sonuçsal geçerlik olarak sınıflandırmıştır (Reckase, 1998). Gruplarası farkların hesaplanması ve gruplarası kıyaslamaların istatistiksel veri analizi bulgularının yordanması delil geçerliği hesaplamaları için kullanılabilecek bir yöntem olarak Messick (1989) tarafından tanımlanmakdadır. Bu araştırmada cinsiyet gruplarının kıyaslanması ile geçerlilik bulgusu elde edilmesi amaçlanmıştır.

Bu nedenle geçerlilik tespiti için istatiksel delil toplamak amacı ile cinsiyet grupları kıyaslanarak diferansiyel madde fonskiyonu (değişen oğe işlevi) yöntemi hesaplamalarından faydalanılmıştır. Araştırmada diferansiyel madde fonsksiyonu hesaplamalarında uygulanacak metodoloji olarak Lord Ki Kare yöntemi (Lord, 1980'in Seybert, Stark, Chernyshenko; 2014'te referans edildiği üzere) seçilmiştir.

Sonuç olarak, araştırmada 2015 kış akademik yarıyılında dokuzuncu sınıf öğrencilerine uygulanan *"uygulamalı matematik"* alan başarı testinin erkek ve kız öğrenci grupları arasındaki madde değişkenliğinin hesaplanması gruplarası kıyaslama metodu ile delil geçerliğine istatistik bulgu sağlaması amaçlanmıştır. Araştırma sorgulaması yapmak amacı ile aşağıdaki nicel araştırma soruları cevaplanmıştır:

1) EKHO uygulamali matematik başarı testinin cinsiyet gruplari kıyas yöntemi ile delil geçerliği sonuçları nelerdir?
2) Dokuzuncu sınıf kız ve erkek öğrencilerinin Lord Ki Kare yöntemi kullanılarak değişen madde fonsksiyonu hesaplamaları sonuçları nelerdir?

[1] Bu çalışmanın verisini sağlayan Kanada Eğitim Kalite ve Hesapverebilirlik Ofisi (Education Quality and Accouantability Office)'ne ve doktora tez danışmanım Dr. Mark Reckase'e teşekkürlerimi sunarım.

Özetle, bu çalışmada Kanada da dokuzuncu sınıf öğrencilerinin uygulamalı matematik alanındaki başarı tespitinin yapılabilmesi için öğrencilere verilen değerlendirme testindeki maddeler değişen madde fonsksiyonu analizi yöntemi ile incelenmiş ve sonuçlar delil geçerliğinin yordanması amacı ile tartışılmıştır. Ayrıca araştırma soruları ve istatistiksel sonuçları Messick (1989)'un geçerlik yaklaşımı temel alınarak yapılandırılmış Kane (1992)'nin geçerlik yaklaşımı olan argüman tabanlı geçerlik yaklaşımı da göz önünde bulundurularak incelenmiştir.

Bu araştırmada yapılan istatistik hesaplamaların ve bulguların matematik başarı testlerinin geçerlilik çalışmaları açısından ölçme ve değerlendirmedeki uygulamalı araştırmalar alanına bir örnek teşkil edeceği ve katkıda bulunacağı düşünülmektedir. Öğrencilerin matematik başarıları ve matematik okur yazarlıkları tüm dünyaca ve Ekonomik İşbirliği ve Kalkınma Organizasyonu (OECD, 2002) tarafinca mesleki hayat, ailevi ve sosyal hayat, yapılandırmacı ve ifadeci vatandaşlıkları bağlamında matematiği tanımlama, anlama, ve ilişki kurabilmeleri olarak tanımlanmıştır. Bu nedenle öğrencilerin matematik alanındaki akademik başarısı tüm diğer Fen Teknoloji Mühendislik Matematik (FTMM) alanlarındaki gibi topluma ve insanlığa katkıda bulunmakta (Ulusal Araştırma Konseyi, 2011); hem günümüz ekonomisine hem gelecekteki ekonomik güçlenmeye önemli ölçüde katkı sağlamaktadır (Schmidt, 2011). Bu nedenle matematik başarı testlerinin psikometrik açıdan incelenmesi ve geçerlik delillerinin toplanarak sunulması açısından bu araştırmanın eğitimde ölçme ve değerlendirme alanındaki uluslararası literatüre anlamlı bir katkı sunacağına inanılmaktadır.

Araştırma verisi 2015 akademik yılında toplanmış olup EKHO tarafından araştırmacıya doktora tez çalışmaları için 2016'da kullanıma açılmıştır, tüm verilerin analizi Uygun-Emil (2020) tarafından farklı bir yaklaşımla test puanlarının madde seviyesindeki eşitlik analizi yöntemi olan Raju'nun madde tepki kuramı tabanlı alan hesaplamaları metodu ile hesaplanmıştır. Bu çalışmadan farklı olarak kış yarıyılındaki *uygulamalı matematik* test puanlarının farklı bir diferansiyel madde fonsksiyonu hesaplama metodu olan Lord'un ki kare yöntemi kullanılarak delil geçerliğinin farklı bir boyutunun incelenmesi amaçlanmıştır.

Yöntem

Araştırmanın amacı ve araştırma soruları bir önceki problem durumunda betimlendikten sonra, bu bölümde araştırma sorularını cevaplamak amacı güdülerek kullanılan sayısal yöntem tanımlanmaktadır.

Öncelikle öğrencilerin matematik başarısını ölçmek için uygulanan testin tanımı yapılıp testi alan öğrencilerden oluşan örneklem tarif edilmiştir. Daha sonrasında istatistik bulgu toplamak amacı ile kullanılan nicel yöntem tanımlanmaktadır.

Ölçme aracı: 2015 senesinde öğrencilerin matematik alanındaki başarısını ölçmek amacı ile kış ve bahar yarıyıllarında iki farklı matematik alan testi uygulanmıştır. Bu testler uygulamalı matematik ve akademik matematik olarak ve herbir testte 24 çoktan seçmeli soru olacak şekilde dokuzuncu sınıf öğrencilerine uygulanmıştır. Bu araştırmada delil geçerliği çalışmasına bir örnek olarak kış döneminde

uygulanmış olan uygulamalı matematik alan testi seçilmiştir. Uygulamalı matematik alan testinde uygulanan 24 çoktan seçmeli madde: Sayılar ve Cebir, Lineer İlişkiler ve Ölçme ve Geometri olmak üzere üç farklı matematik alanını kapsamaktadır. Bu alanların madde sayısı 7 adet Sayılar ve Cebir alan sorusu, 11 adet Lineer İlişkiler alan sorusu ve 6 adet Ölçme ve Geometri alan bilgi sorusu şeklinde dağılım göstermektedir. Bu maddelerde ölçülen bilişsel süreç bererilerine örnek olarak:

1- Sayısal ve cebir hesaplamaları
2- Matematiksel ilişkilendirme yöntemleri
3- Matematik kuralları ve uygulamarı
4- Veri yönetimi ve geometri teknikleri

öğrenme alanları Kanada ülkesi Ontario eyaleti müfredat kazanımları (Ontario Eğitim Bakanlığı, 2005) ve EKHO test yapıçatısı (Eğitim Kalite ve Hesapverebilirlik Ofisi, 2009)'ndan örnek olarak verilebilir.

Örneklem, 2015 kış sömestırı uygulamalı matematik alanında 15,994 öğrenci çoktan seçmeli 24 soru ile alan başarısı bakımından değerlendirilmiştir. Bu örneklemdeki öğrencilerden 8995 (%56)'i erkek cinsiyet grubuna 6999 (%44)'u ise kız öğrenciler cinsiyet grubuna dahildir. Çalışmada delil geçerliğine yönelik istatistiksel bulgu elde edebilmek amacı ile Lord'un Ki Kare yöntemi (Lord, 1980) ile diferansiyel madde fonksiyonu hesaplaması yapılmıştır. Veri analizinde R programlama dili yazılımı 4.1.0 versiyonu ve R Studio arayüzünden faydalanılmıştır. Lord metodu uygulamasının kodu ise 13.05.2020 yayım tarihli 'difR' paketi kullanılarak yazılmıştır.

Sonuçlar

Aşağıdaki tabloda Lord Ki Kare yöntemi ile diferansiyel madde fonsiyonu incelemesi amacıyla yapılan istatistik hesaplamalarının sonuçları sergilenmektedir.

Tablo 2

EKHO 2015 Kış Dönemi Uygulamalı Matematik Lord Ki Kare DMF Bulguları

Madde	Lord Ki-Kare değeri	p	mF-mR	Delta Lord etki büyüklüğü	Etki büyüklüğü kodu
MC01r	2.21	.137	0.0695	-0.1633	A
MC02r	8.00	.0047**	-0.1079	0.2536	A
MC03r	0.04	.8612	0.0067	-0.0157	A
MC04r	1.04	.3074	-0.0407	0.0956	A
MC05r	1.35	.2446	0.0449	-0.1055	A
MC06r	42.82	0***	0.2603	-0.6117	A
MC07r	23.84	0***	-0.1952	0.4587	A
MC08r	35.12	0***	-0.2284	0.5367	A
MC09r	0.20	.6538	0.0178	-0.0418	A
MC10r	0.68	.4081	-0.0333	0.0783	A
MC11r	18.72	0***	-0.1671	0.3927	A
MC12r	106.32	0***	0.4139	-0.9727	A

(devam ediyor)

Tablo 2 *(devam)*

Madde	Lord Ki-Kare değeri	p	mF-mR	Delta Lord etki büyüklüğü	Etki büyüklüğü kodu
MC13r	20.34	0***	0.1749	-0.411	A
MC14r	0.03	.8572	-0.0073	0.0172	A
MC15r	0.22	.6398	-0.0183	0.043	A
MC16r	22.93	0***	-0.1943	0.4566	A
MC17r	71.19	0***	-0.3243	0.7621	A
MC18r	5.11	.0238*	-0.0929	0.2183	A
MC19r	4.04	.0445*	0.0811	-0.1906	A
MC20r	9.83	.0017**	-0.1372	0.3224	A
MC21r	0.37	.5457	-0.0231	0.0543	A
MC22r	0.21	.6495	-0.0174	0.0409	A
MC23r	23.42	0***	0.185	-0.4348	A
MC24r	55.18	0***	0.3334	-0.7835	A

Anlam değeri kodları: 0= ***; .001=**; .01= *. Etki büyüklüğü kodları: 0='A' 1='B' 1.5='C'

Yukarıdaki tabloya göre 18 ve 19'uncu maddeler "gözardı edilebilir" düzeyde, 2 ve 20'inci maddeler "orta/makul" düzeyde, 6, 7, 8, 11, 12, 13, 16, 17, 23 ve 24 maddeleri ise "yüksek" düzeyde diferansiyel madde fonskiyonu özelliği bulgusu sergilemektedir.

6, 12, 13, 23 ve 24 maddeleri, kız öğrencilerinin avantajına çalışırken 7, 8, 11, 16 ve 17'inci maddeler erkek cinsiyet grubundaki öğrencilerin avantajina işlev göstermektedir.

Kız öğrenciler lehine işleyen maddelerin kapsam bakımından dağılımı 1 Cebir maddesi, 1 Lineer İlişkiler maddesi, 3 Ölçme ve Geometri maddesi şeklindedir. Erkek cinsiyet grubundaki öğrencilerin avantajına işlev gösteren maddelerin konu dağılımı ise 1 Cebir, 3 Lineer İlişkiler, 1 Ölçme ve Geometri maddesi şeklindedir.

Tartışma

Bu araştırma sonuçlarınca kız ve erkek öğrencilerin lehine işlev gösteren maddeler cebir alanında sayıca eşittir. Matematik eğitimi literatüründe kız öğrencilerin aritmetik ve cebir alanında daha başarılı olabileceğine dair bulgu sunan çalışmalar vardır. Kanada'daki bu bulgu EKHO testinin cebir alanında cinsiyet eşitşiği bakımından geçerlik delili sunmaktadır yordaması yapılabilir. Ayrica matematik alanındaki araştırmalar erkek öğrencilerin bu alandaki akademik başarısını desteklerken (Maccoby ve Jacklin, 1974; Fennema ve Carpenter, 1981; Benbow, 1988; Halpern, 2000'nin Zhu, 2007'de referans edildiği üzere) Ontario eyaletinden farklı bir bulgu olarak dokuzuncu sınıf kız öğrencilerinin geometri alanında lehine işlev gösteren maddelerin erkek öğrencilerininkinden sayıca fazla oluşu ise literatüre geometri ve aritmetik alanlarındaki bulguların daha detaylı incelenmesi konusunda akademik merak unsuru oluşturmaktadır.

Kaynaklar

Cohen, A. S., & Kim, S. H. (1993). A comparison of Lord's χ2 and Raju's area measures in detection of DIF. *Applied Psychological Measurement, 17*(1), 39-52.

Education Quality and Accountability Office. (2009). *Framework Grade 9 Assessment of Mathematics*. EQAO.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.), (pp. 13-103). American Council on Education and Macmillan.

Ontario Ministry of Education. (2005). *The Ontario curriculum grades 9 and 10: Mathematics*. (ISBN 0-7794-7940-80). Queen's Printer for Ontario.

Organization of Economic Co-operation and Development. (2002). *Education at a glance 2002*. Retrieved from http://www.oecd.org/education/skills-beyond-school/educationataglance2002-home.htm

Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice, 17*(2), 13–16. https://doi.org/10.1111/j.1745-3992.1998.tb00827.x

Schmidt, W. H. (2011). *STEM reform: Which way to go?* Paper presented at the National Research Council Workshop on Successful STEM Education in K-12 Schools. Available at: http://www7.nationalacademies.org/bose/STEM_Schools_Workshop_Paper_Schmidt.pdf

Seybert, J., Stark, S., & Chernyshenko, O. S. (2014). Detecting DIF with ideal point models: A comparison of area and parameter difference methods. *Applied Psychological Measurement, 38*(2), 151-165. https://doi.org/10.1177/0146621613508306

Zhu, Z. (2007). Gender differences in mathematical problem solving patterns: A review of literature. *International Education Journal, 8*(2), 187-203. https://files.eric.ed.gov/fulltext/EJ834219.pdf

Puanlayıcılar arası yüksek uyumun puanlayıcı güvenirlik katsayı üzerindeki etkisi

Sümeyra Soysal

Giriş

Puanlayıcı güvenirliği, iki bağımsız puanlayıcı tarafından (inter-rater) veya farklı zamanlarda aynı puanlayıcı tarafından (intra-rater) atanan puanların tutarlılığıdır (Reddy ve Andrade, 2010). Bir başka ifade ile değerlendiriciler tarafından verilen derecelendirmelerdeki anlaşma veya fikir birliği derecesidir ve eğitim, iletişim, hesaplamalı dilbilimi, pazarlama, psikoloji, sosyoloji, tıp bilimi gibi birçok disiplin alanında, ancak genellikle farklı terminolojilerle birlikte ele alınan öneli bir konudur (Feng, 2015). Puanlayıcı güvenirliği, genellikle eğitim alanında, performans değerlendirme, ölçek ve ölçme aracı geliştirme çalışmalarının bir parçası olarak karşımıza çıkmaktadır. Değerlendiricilerin üzerinde anlaştıkları durumların sayısının, analiz edilen toplam durum sayısına bölünmesiyle elde edilen uyum yüzdesi en eski ve sıklıkla kullanılan bir uzlaşma katsayısıdır (Krippendorff, 2004). Ancak bu istatistik zor derecelendirme görevlerinde, puanlayıcıların tamamen tesadüfi uzlaşmaların etkisi ortadan kaldırılmaz, bu da puanlayıcı güvenilirliğinin olduğundan fazla kestirilmesi ile sonuçlanır (Cohen, 1960; Zhao ve diğ., 2013). Uyum yüzdesinin bu sınırlılığı üzerine, tesadüfi uzlaşmayı hesaba katan Cohen Kappa, Fleiss Kappa, Scott'ın Pi katsayısı gibi çeşitli puanlayıcılar arası güvenirlik katsayıları geliştirilmiştir. Ancak böyle bir düzeltmenin fazla tutucu olduğu için eleştirilir (Grant ve diğ., 2017). Bir diğer ifade ile puanlayıcılar arasında tesadüfi olarak ortaya çıkmayan yüksek bir uzlaşma oranını tesadüfi olan uzlaşmadan ayıramadığı için güvenirliği olduğunda daha düşük kestirmektedir (Feinstein ve Cicchetti, 1990; Gwet, 2012). Alanyazında, Cohen'nin Kappası, Scott'in Pi katsayısı gibi iki puanlayıcı arasındaki güvenirliği hesaplayan indeksler için bu sorunsalı ele alan araştırmalar mevcuttur (Gwet, 2002a; Gwet, 2002b; Feuerman ve Miller, 2008; Kanik ve diğ., 2012). İkiden fazla puanlayıcılar için bu sorunsalı ele alan araştırma sayısı sınırlıdır (Gwet, 2008; Zepeda ve Jimenez, 2019; Tong ve ark, 2020). Bu nedenle eğitim alanında ikiden fazla puanlayıcılar için sıklıkla kullanılan Krippendorff'un alfası, sınıf içi korelasyon katsayısı, Fleiss'in kappası, genellebilirlik kuramının g katsayısı ile Gwet'in AC1 katsayısının puanlayıcılar arası yüksek uyum varlığında performansı araştırılacaktır. Böylece, bu araştırma ile araştırmacılara puanlayıcı güvenirlik katsayısının dikkatli kullanımı noktasında yol gösterilmesi beklenmektedir.

Yöntem

Deneysel desenler değişkenler arasındaki neden sonuç ilişkilerini test etmeyi amaçlayan araştırma desenleridir. Bu amacı gerçekleştirebilmek için Fraenkel, Wallen ve Hyun'a (2011) göre deneysel desenler, bağımsız değişken/lerin bağımlı değişken/ler üzerindeki etkisini incelemek için en az iki koşulun karşılaştırılmasını ve bağımsız değişkenin araştırmacı tarafından doğrudan değişimlenmesini (manipüle edilmesini) gerektirir. Ayrıca, araştırmacılar iç geçerliği korumak için dışsal değişkeni (ilgilenilmeyen ya da istenmeyen değişken) kontrol altına alarak bağımlı değişken üzerinde ölçme yapmalıdır (Gall ve diğ., 2003). Simülasyon çalışmaları doğası gereği araştırmacılara bağımsız değişkenleri değişimleme ve dışsal değişkenleri kontrol altına alma imkânı sağlar. Puanlayıcı sayısı 2 ve 5, madde sayısı sabit olup 10, puanlayıcıların kullanacağı dereceleme ölçeği 2, 3 ve 5 olmak üzere 25 tekrarlı veri setleri üretilecektir. Puanlayıcılar arası uyum, 0.00-1.00 arasında 0.10 artışlarla manipüle edilecektir. Bu durumda puanlayıcı sayısı (2), dereceleme ölçeği (3), puanlayıcılar arası uyum (11) olarak 2x3x11= 66 koşul incelenecektir. Verilerin üretilmesi R programında yapılacaktır Verilerin analizlerinde ise Krippendorff'un alfası, sınıf içi korelasyon katsayısı ve Fleiss'in kappası katsayılarının kestiriminde "irr" paketi (Gamer ve diğ., 2012), AC1 katsayısının kestiriminde "irrCAC" paketi (Gwet, 2019) ve genellebilirlik kuramının g katsayısının kestirimleri için EduG (version 6.1 e) programı kullanılacaktır. Bir katsayısı için 25 tekrarın ortalaması alınarak uyum yüzdelerine göre katsayıların performansı karşılaştırılacaktır.

Sonuçlar

Cohen kappa, Scott pi, Kendal w gibi güvenirlik katsayılarında ortaya çıkarılan yüksek uzlaşı düşük güvenirlik paradoksunun Fleiss kappa ve Krippendorff'un alfası, sınıf içi korelasyon katsayısında da ortaya çıkması beklenmektedir. Uyum katsayısına göre bu katsayıların performansının değişmesinin yanış sıra puanlayıcı ve kategori sayısından da olumsuz yönde etkilenmesi beklenmektedir. Genellebilirliğin g katsayısının uyum yüzdesi ile tutarlı sonuçlar vermesi ve puanlayıcı ve kategori sayısından daha az etkilenmesi beklenmektedir. Benzer sonuçlar Gwet'in AC1 katsayısı için de beklenmektedir. Çünkü Gwet (2008), kappanın paradoksunu ele almak için tesadüfi uzlaşmaya göre düzeltilmiş bir endeks olan AC1 katsayısını önerdi ve Gwet (2012) simülasyon çalışmasında AC1'in kappa paradoksuna dirençli olduğunu göstermiştir.

Kaynaklar

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Feuerman, M., & Miller, A. R. (2008). Relationships between statistical measures of agreement: sensitivity, specificity and kappa. *Journal of evaluation in clinical practice, 14*(5), 930-933.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*(6), 543–549.

Feng, G. C. (2015). Mistakes and how to avoid mistakes in using intercoder reliability indices. *Methodology, 11*(1), 13-22.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2011). *How to design and evaluate research in education.* (8th ed.). McGraw – Hill.

Gall M. D., Gall, J. P., & Borg, W., R. (2003). *Educational research: An introduction.* (7th ed.). Pearson Education, Inc.

Gamer, M., Lemon, J., Gamer, M. M., Robinson, A., & Kendall's, W. (2012). *Irr: Various coefficients of interrater reliability and agreement* (version 0.84.1) [Computer Software]. http://cran.cc.uoc.gr/mirrors/cran/web/packages/irr/irr.pdf

Grant, M. J., Button, C. M., & Snook, B. (2017). An evaluation of interrater reliability measures on binary tasks using d-prime. *Applied Psychological Measurement, 41*(4), 264–276.

Gwet, K. (2002a). Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical Methods for Inter-rater Reliability Assessment, 1*(6), 1-6.

Gwet, K. (2002b). Inter-rater reliability: dependency on trait prevalence and marginal homogeneity. *Statistical Methods for Inter-Rater Reliability Assessment Series*, *2*(1), 9.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29-48.

Gwet, K. L. (2012). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among multiple raters* (3rd ed.). Gaithersburg, MD: Advanced Analytics.

Gwet K. L. (2019). irrCAC: Computing chance-corrected agreement coefficients (version 1.0) https://cran.csiro.au/web/packages/irrCAC/irrCAC.pdf

Kanik, E. A., Erdoğan, S. ve Temel, G. O. (2012). İki sonuçlu tanı testlerinde iki hekim arasındaki uyum istatistiklerinin prevalanstan etkilenme durumları, *İnönü Üniversitesi Tıp Fakültesi Dergisi, 19*(3), 153-158.

Krippendorff, K. (2004). Reliability in content analysis. *Human Communication Research, 30*(3), 411–433.

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education,* 35(4), 435-448.

Tong, F., Tang, S., Irby, B. J., Lara-Alecio, R., & Guerrero, C. (2020). The determination of appropriate coefficient indices for inter-rater reliability: Using classroom observation instruments as fidelity measures in large-scale randomized research. *International Journal of Educational Research*, *99*, 101514.

Zepeda, S. J., & Jimenez, A. M. (2019). Teacher evaluation and reliability: Additional insights gathered from inter-rater reliability analyses. *Journal of Educational Supervision*, *2*(2), 11-26.

Zhao, X., Liu, J. S., & Deng, K. (2013). Assumptions behind intercoder reliability indices. *Annals of the International Communication Association, 36*(1), 419–480.

Klasik istatistiksel yöntemler ile veri madenciliği yöntemlerinin yordayıcı değişken belirleme ve sınıflama etkinliği bakımından karşılaştırılması

Gürkan Cüvitoğlu ve Tuncay Öğretmen

Giriş

Eğitim, genel anlamda kasıtlı kültürleme sürecidir. Buradaki kasıt kavramı, beklentileri tesadüflere bırakmayan, dolayısıyla istenmedik durumları dışarda tutan bir programa vurgu yapar. Bu program ne kadar doğru ve etkili bir şekilde düzenlenmiş olursa olsun, süreç sonunda yine de ölçme ve değerlendirme işlemleriyle kalite kontrolü yapılır. Eğitimde gerçekleşen ölçme ve değerlendirme işlemleri, hem öğrencinin, hem de programın etkililiği açısından dönütler sağlar (Hotaman, 2020) ve sürecin başarısı test edilmiş olur. Eğitimde başarı kavramıyla genellikle okulda okutulan derslerde geliştirilen ve öğretmenlerce takdir edilen notlarla, test puanlarıyla ya da her ikisiyle birlikte belirlenen beceriler veya kazanılan bilgilerin ifadesi olan "akademik başarı" kastedilmektedir. Akademik başarı öğrencilerin mesleki ve toplumsal yaşama donanımlı şekilde hazırlanmalarını sağladığı ve geleceklerini şekillendirdiği için, aileleri ve çevreleri açısından da oldukça önemli görülmektedir (Sarıer, 2016). Öğrenciler eğitim-öğretim sürecinin sonunda bir üst eğitim-öğretim seviyesine geçebilmek için genelde seçme ve yerleştirme sınavlarına katılmaktadır. Seçme ve yerleştirme sınavlarında amaç adaylar arasından ölçülen özellikleri taşıyan üst grubu belirlemek, yeterli olanları birbirinden ayırt edebilmek ve kendi aralarında sıralayabilmektir (Turgut ve Baykul, 2013).

Aynı okullardan, aynı sınıflardan hatta ikiz kardeşler olmalarına rağmen öğrencilerin akademik başarı bakımından birbirlerinden çok farklı seviyelerde olmaları gözlenen bir durumdur. Bu durumda başarı bakımından öğrencilerin üst grupta bulunmalarını sağlayan faktörler neler olduğu düşünülmelidir. Sarıer, 2016'da yayınlanan meta-analiz çalışmasında akademik başarının öğrenme hızı, zeka gibi zihinsel etmenlerle, benlik saygısı, kişilik yapısı, öz-yeterlik, motivasyon ve ders çalışma alışkanlıkları gibi duyuşsal etmenlerle, anne-baba tutumu, ailenin sosyo-ekonomik durumu, okul yöneticilerinin ve öğretmenlerin yeterliliği ve tutumu gibi çevresel etmenlerle ilişkili olduğunun farklı araştırmalar tarafından belirtildiğini aktarmıştır. Ayrıca Türkiye'de gerçekleştirilen araştırmalarda, akademik başarıyı, öğrenciden, okuldan ve aileden kaynaklı bazı faktörlerin etkilediği belirlenmiştir. Bu noktada akademik başarıyı etkileyen öğrenci kaynaklı faktörler, benlik saygısı, özyeterlik, motivasyon ve ders çalışma alışkanlığı olarak bulunmuştur (akt. Sarıer, 2016).

Genel anlamda ve matematik dersi özelinde Miñano ve Castejón (2011) başarı bakımından duyuşsal değişkenlerin, zekâdan daha açıklayıcı olduğunu bulmuşlardır. Garcia ve diğ., (2016) matematiğe dair motivasyonun ve matematik dersinden alınan zevkin matematik başarısının güçlü yordayıcıları olduğu sonucuna varmıştır. Benzer şekilde, Lipnevich ve diğ., (2016), zekânın matematik başarısının önemli bir yordayıcısı olduğunu fakat matematiğe yönelik tutumun öğrencilerin başarısını açıklamada daha önemli olduğunu belirtmişlerdir.

Yukarıda bahsedilen çalışmalarda ve bunlardan başka birçok çalışmada öğrencilerin matematik başarılarının açıklanmasına yönelik modeller önerilmiştir ve matematik başarısına etki eden yordayıcılar ve bu yordayıcıların başarıyı açıklama bakımından önemleri belirlenmeye çalışılmıştır. İstatistiksel modeller (Yapısal Eşitlik Modelleri, Boylamsal Modeller, Örtük Sınıf Modelleri gibi.) kavramsal bir altyapıya sahiptir yani model, modeli oluşturan değişkenler ve örneklem kısaca araştırma deseni sürecin başında belirlenir ve kullanılan yönteme göre sağlanması gereken varsayımlar mevcuttur. Fakat son yıllarda bilişim teknolojilerinde yaşanan gelişmeler ve zaman içerisinde birçok alanda veri birikmesi büyük veri tabanlarının oluşmasını da beraberinde getirmiştir. Ancak bu veriler içinden anlamlı ve yararlı olanları ortaya çıkarmada güçlük yaşanmaktadır (Alan, 2012). Geleneksel istatistiksel yöntemlerde büyük miktarda veriyi çözümlemek ve anlamlandırmak kolay olamamaktadır. Veri madenciliği (VM) yöntemleri yaşanan bu güçlükleri ortadan kaldırma gereksinimi üzerine ortaya çıkmıştır (Özkan, 2008). Veri tabanlarındaki veriler üzerinde farklı disiplinler, farklı amaçlarla istatistiksel ya da matematiksel analizler yapmaktadırlar (Alan, 2014). Birçok alanda olduğu gibi eğitim alanında da uzaktan eğitim, web tabanlı eğitim, bilgisayar tabanlı öğretim ve bunlar gibi birçok etkileşimli eğitim ortamı da göz önüne alındığında öğrenme ortamlarında zaman içerisinde büyük veri birikmektedir. Eğitsel veri madenciliği (EVM) bu bağlamda istatistik, bilgisayar bilimleri ve eğitim alanlarının kesişiminde bulunan bir çalışma alanı olarak görülmektedir. (Romero ve Ventura, 2013). Aydoğdu (2020) Türkiye'de EVM ile ilgili yayınlanan 51 çalışmayı incelemiştir. Bu çalışmanın sonuçlarına göre Türkiye'de EVM kapsamında yapılan çalışmaların 20'sinde başarı/performans tahmini ve 6'sında başarıya göre sınıflama çalışılmıştır. Veri madenciliğinde sınıflandırma, tahminleme, kümeleme ve sınıflandırma ve biriktelik kuralları olmak üzere dört yöntem olsa da Baker (2010) sınıflandırma yöntemini tahminleme yönteminin kapsamında değerlendirmiştir. Bu durum göz önünde bulundurulacak olursa Aydoğdu (2020) incelediği çalışmaların yaklaşık %70'inde sınıflandırma (tahminleme) yöntem ve algoritmaları kullanılmıştır.

Veri madenciliği, bilinmeyen desenlerin ortaya konması amacıyla veri tabanındaki bilgiyi keşfetmeyi amaçlayan bir analiz yöntemidir (Larose, 2005). Veri madenciliği, büyük veri setleri içinde saklı kalmış önemli bilginin açığa çıkarılmasında kullanılan bir yöntem ve bu verideki örüntülerin geçerli, özgün, kullanışlı ve oldukça anlaşılır biçimde tanımlanması sürecidir (Fayyad, 1998; Fayyad ve diğ., 1996). Sürecinin önemli aşamalarından biri veri boyutunun azaltılması işlemidir. Veri boyutunun azaltılması veri kümesinden ilgisiz veya gereksiz değişkenlerin çıkartılması olarak tanımlanmaktadır. Veri boyutunun azaltılması için kullanılan yöntemlerin başında özellik/öznitelik seçimi (yordayıcı değişken seçimi) gelmektedir. Yordayıcı değişken seçimi, orijinal veri setini temsil edebilecek en iyi altkümenin

seçimi olarak tanımlanmaktadır. Bu işlem, ilgilenilen problem için en faydalı ve en önemli özellikleri seçerek veri setindeki değişken sayısını azaltmayı yani veri boyutunu düşürmeyi amaçlamaktadır (Budak, 2018). Benzer işlemler klasik istatistiksel yöntemlerle de yapılabilmekle beraber verinin normal dağılımı başta olmak üzere uygulanacak istatistiksel yönteme göre sağlanması gereken başka varsayımlar (hataların/artıkların normal dağılması ve ilişkisiz olması, bağımsız değişkenlerin bağımlı değişken ile lineer ilişkiye sahip olması, çoklu bağlantılılık problemi gibi) mevcuttur.

Bu çalışmanın amacı klasik istatistiksel yöntemler olan lojistik regresyon analizi ve diskriminant analizi ile veri madenciliği kapsamında kullanılan farklı yordayıcı değişken belirleme yöntemleri (feature selection methods) yardımıyla belirlenen yordayıcı değişken kümelerinin öğrenci matematik başarısını yordama etkinliğinin test edilmesi ve karşılaştırılmasıdır.

Yöntem

Bu çalışmada uluslararası düzeyde, eğitsel girişimler ve politikalar geliştirmek amacıyla grup performansını belirlemek için uygulanan Uluslararası Matematik ve Fen Eğilimleri Çalışmasına (TIMSS) ait 2019 yılı 8. Sınıf düzeyi Türkiye örneklemine ait veri seti kullanılacaktır. TIMMS gibi geniş ölçekli uygulamalarda öğrenci performansını ölçen bilişsel testlerin yanında, öğrenci, öğretmen ve okul düzeyinde bilgi toplanmasını sağlayan ölçekler de kullanılır. Ayrıca öğrenci yeteneğine ilişkin kestirimler olarak makul değerler (plausible value) hesaplanır. (Tat ve arkadaşları, 2019). Bu çalışmada Türkiye örneklemine ait 4077 öğrenciye ait değişkenlerin bulunduğu veri seti kullanılmıştır. Matematik için 1. makul değer ile öğrenci anketiyle toplanan ve öğrencilerin başarısına etki edeceği düşünülen türetilen sürekli değişkenler (derived scale variables) kullanılmıştır. Veri setinde makul değerler uluslararası kesme puanlarına (uluslararası yeterlilik düzeyleri/international benchmarks) göre 5 kategoriye (<400, 400-475, 475-550, 550-625 ve >625) ayrılmıştır. Veri setinde makul değerler uluslararası kesme puanlarına (international benchmarks) göre 5 kategoriye ayrılmıştır. Çalışmada en alt (507 öğrenci) ve en üst (441 öğrenci) kategorilerdeki öğrenciler alt ve üst gruplar olarak kodlanmıştır. Çalışmada ev ortamı, kaynak olanakları, matematik dersine karşı tutum, okul ve sınıf ortamı gibi yordayıcı değişkenler arasından lojistik regresyon analizi, diskriminant analizi ve veri madenciliği yöntemleri (filter, wrapper ve embedded) ile öğrencinin matematik başarısı açısından bulunduğu grubu (alt-üst grup) en etkili yordayan değişken alt kümeleri (yeni veri setleri) elde edilecektir. Ardından elde tüm yeni veri setleri gene diskriminant analizi, lojistik regresyon analizi ve farklı veri madenciliği sınıflandırma yöntemleri (Yapay Sinir Ağları, Karar Ağaçları, Naif Bayes, k-en Yakın Komşu gibi) ile analiz edilecektir. Burada ilk amaç en iyi sınıflandırmayı sağlayan veri setinin hangi yöntemle elde edildiğini bulmak yani başka bir deyişle klasik istatistiksel yöntemler ile VM yöntemlerinin yordayıcı değişken belirleme etkinliğinin test edilmesidir. İkinci amaç ise elde edilen ve başarıyı yordamada etkinliği test edilmiş bir veri seti ile klasik istatistiksel yöntemler ile VM yöntemlerinin sınıflandırma başarılarının test edilip karşılaştırılmasıdır.

Sonuçlar

Bu çalışmada veri madenciliği prosedürleri olan öznitelik seçme yöntemleri ile belirlenen yordayıcıların sınıflandırma etkinliğinin klasik istatistiksel yöntemlerinden daha yüksek olacağı düşünülmektedir. Gerekli varsayımların karşılanması durumunda klasik istatistiksel yöntemlerin sınıflandırma başarılarının büyük etki büyüklüğüne ulaşacağı düşünülmektedir. Öte yandan VM madenciliği sınıflandırma algoritmalarının optimize edilebiliyor olasından dolayı (örneğin; yapay sinir ağlarındaki katman sayısı ya da en yakın komşuluk sayısı gibi) sınıflandırma etkinliklerinin klasik istatistiksel yöntemlerden yüksek olacağı beklenmektedir. Eğitsel veri madenciliği kapsamında öğrencilerin başarı yönünden sınıflandırılmasında en etkili olan değişkenlerin belirlenmesi yanında olabildiğince az değişken (parsimony) ile daha etkili bir sınıflama yapılmasının önemli olduğu düşünülmektedir.

Kaynaklar

Alan, M. A. (2012). Veri madenciliği ve lisansüstü öğrenci verileri üzerine bir uygulama. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, 33*, 165-174. https://dergipark.org.tr/tr/pub/dpusbe/issue/4775/65775

Alan, M. A. (2014). Karar ağaçlarıyla öğrenci verilerinin sınıflandırılması. *Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi, 28*(4), 101-112. https://dergipark.org.tr/tr/pub/atauniiibd/issue/2715/35968

Aydoğdu, Ş. (2020). Educational data mining studies in Turkey: A systematıc review. *Turkish Online Journal of Distance Education, 21*(3) , 170-185. https://doi.org/10.17718/tojde.762046

Baker R. S. J. (2010). Mining Data for Student Models. In: R. Nkambou, J. Bourdeau, R. Mizoguchi (Eds.) *Advances in ıntelligent tutoring systems* (vol 308, pp. 323-337). Springer. https://doi.org/10.1007/978-3-642-14363-2_16

Budak, H. (2018). Özellik seçim yöntemleri ve yeni bir yaklaşım. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 22*(Özel sayı), 21-31. https://doi.org/10. 10.19113/sdufbed.01653.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM, 39*, 27-34. https://doi.org/10.1145/240455.240464.

Fayyad, U. (1998). Mining databases: Towards algorithms for knowledge discovery. *DE Bulletin, 21*(1), 41-48.

García, T., Rodríguez, C., González-Castro, P., Torrance, M., & Gonzalez-Pienda, J. A. (2016). Elementary students' metacognitive processes and post-performance calibration on mathematical problem-solving tasks. *Metacogn. Learn,* 11, 139–170. https://doi.org/10.1007/s11409-015-9139-1

Hotaman, D. (2020). Öğrenci başarısının değerlendirilmesinde eğitsel veri madenciliğinin kullanımı. *Ulakbilge, 48*, 577–587. https://doi.org/10.7816/ulakbilge-08-48-08

Larose, D. T. (2005). *Discovering knowledge in data.* Wiley Publication.

Lipnevich, A. A., Preckel, F., & Krumm, S. (2016). Mathematics attitudes and their unique contribution to achievement: Going over and above cognitive ability and personality. *Learning and Individual Differences, 47*, 70–79. https://doi.org/10.1016/j.lindif.2015.12.027

Miñano, P., & Castejón, J. L. (2011). Cognitive and motivational variables in the academic achievement in language and mathematics subjects: A structural model. *Rev. Psicodidáct.* 16, 203–230.

Özkan, Y. (2008). *Veri madenciliği yöntemleri*. Papatya Yayınları.

Romero, C., & Ventura, S. (2013). Data mining in education. *WIREs Data Mining Knowledge Discovery, 3*(1), 12–27. https://doi.org/10.1002/widm.1075

Sarıer, Y. (2016). Türkiye'de öğrencilerin akademik başarısını etkileyen faktörler: Bir meta-analiz çalışması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *31*(3), 609-627.

Tat, O., Koyuncu, İ., & Gelbal, S. (2019). The influence of using plausible values and survey weights on multiple regression and hierarchical linear model parameters. *Journal of Measurement and Evaluation in Education and Psychology, 10*(3), 235-248. https://doi.org/10.21031/epod.486999

Turgut, M. F. ve Baykul, Y. (2013). *Eğitimde ölçme ve değerlendirme*. Pegem Akademi.

Bulanık mantık yaklaşımımın madde seçiminde kullanılması

Kübra Çetiner Koç ve Fatih Kezer

Anahtar kelimeler: Bulanık TOPSIS, bulanık VIKOR, test geliştirme, madde seçimi.

Giriş

Testlerden elde edilen puanların geçerlik ve güvenirlik gibi temel psikometrik özelliklerinin istenilen düzeyde olması için test geliştirme işlemlerinin belirli basamaklar dikkate alınarak yürütülmesi gerekir. Bu basamaklar takip edildiğinde, maddeler yazılıp redaksiyon işlemlerinden geçirilerek uygun bulunan maddelerle deneme formu hazırlanır ve esas testin uygulanacağı gruba benzer nitelikler taşıyan bir grup üzerinden deneme uygulaması yapıldıktan sonra madde analizleri gerçekleştirilerek madde seçim işlemi gerçekleştirilir (Crocker ve Algina, 1986; Çetin, 2019; Downing ve Haladayna, 2006; Tekin, 2019). Ayrıca madde analizlerinin değerlendirilmesinde, maddelerin, yanıt formatlarının uygunluğunun gözden geçirilmesinde uzman yargılarına başvurulması gerektiği ve uzmanların nitelikleri, ilgili alanda deneyimleri ve demografik özelliklerinin ve maddelerin seçildiği süreç ve madde güçlüğü, ayırt ediciliği gibi madde seçimi için kullanılan verilerin belgelendirilmesi gereklidir (APA, 1999). Madde analizlerinin hesaplanması; maddelerin testin amacına uygun olarak hazırlanmasının ve beklenenden çok güç ya da kolay bir test geliştirilmesinin önlenmesi açısından önemlidir.

Kişilerin meslek edinmesi gibi hayati kararların verilmesinde etkili olan bazı sınavlarda güvenlik problemleri nedeniyle maddelerin deneme uygulamasının yapılamadığı durumlar mevcuttur. Bu gibi durumlarda madde seçimi, madde analizlerine bakılmadan doğrudan uzman görüşü ile yapılmaktadır. Uzmanların bir madde hakkında değerlendirme yapması sırasında maddenin tahmini madde güçlüğü, ayırt ediciliği ve ölçülmek istenilen davranışı yoklama durumu gibi konularda fikir yürütmesi ve maddenin dikkate alınan özellikler bakımından iyi ya da kötü olduğu hakkında karar vermesi beklenir. Ancak insan düşünme yapısı iyi/kötü şeklinde ikili mantık sistemi ile düşünmeye uygun değildir. Bu sebeple uzmanlar değerlendirme işlemlerinde kararsız kalabilmekte ve maddeyi "biraz iyi", "fena değil", "oldukça iyi" gibi dilsel ifadelerle değelendirilebilmektedir. Bu ifadeler; uzmanların niteliklerinin, ilgili alanda deneyimlerinin ve demografik özelliklerinin aynı olmaması sebebiyle her uzman için aynı anlamı ifade etmemekte ve bu durum da belirsizlik probleminin temelini oluşturmaktadır. Sözel değişkenlerin herkes için aynı anlama gelmesi ve matematiksel olarak ifade edilebilmesi için bulanık kümelerin

kullanılması gerekmektedir (Özkan, 2003). Ayrıca madde seçim işleminin bir uzman grubu ile birlikte yapıldığı durumlarda farklı problemler de ortaya çıkmaktadır. Kararın oy çokluğuna göre alınması durumunda madde hakkında olumsuz yargıda bulunan uzmanlarıın değerlendirmesi dikkate alınmamaktadır. Uzman değerlendirmelerinin uyum yüzdesine bakılarak madde seçim işlemi yapılmasında ise uzmanların herhangi bir madde hakkında uyumlu karar veremedikleri durumlar olacaktır. Bu sebeple değerlendirmelerin tümünün eş zamanlı işe koşulamaması objektiflik problemini oluşturmaktadır. Bu çalışmada test geliştirme adımlarından biri olan madde havuzundan madde seçim işlemi, çok kriterli belirsizlik içeren karar problemi olarak ele alınıp Bulanık TOPSIS ve Bulanık VIKOR yöntemleri ile gerçekleştirilmiş ve bu iki yöntemin benzerlik ve farklılıkları üzerinde durulmuştur.

Yöntem

Bu çalışmada bulanık çok kriterli karar verme yöntemlerinden bulanık TOPSIS ve bulanık VIKOR yöntemlerinin test geliştirmede madde havuzundan madde seçiminde kullanılması bakımından temel araştırma modelindedir. Temel araştırmalar, kuramlara dayanarak geliştirdiği varsayımları test ederek, test sonuçlarını bilimsel olarak yorumlayarak ortaya çıkaran araştırmalardır (Karasar, 2010). Bulanık TOPSIS yöntemi karar vericilerin, karar kriterlerini ve alternatifleri dilsel değişkenleri kullanarak değerlendirirmesine bu değerlendirmelerin üçgen ve yamuk bulanık yamuk sayılara dönüştürülmesine ve her bir alternatif için alternatifler arasından belirlenen pozitif ve negatif ideal çözüm noktalarına göre yakınlık katsayısı hesaplanmasına dayanır (Paksoy ve diğ., 2013). Alternatiflerin yakınlık katsayı değerleri 0 ile 1 arasında çıkar ve yakınlık katsayısının büyük olan alternatifin karar vericiler tarafından tercih edilmesi beklenir (Ateş ve diğ., 2006; Chen, 2000). Bulanık VIKOR yönteminde ise benzer basamaklar takip edilir ve yöntemin amacı alternatifleri sıralamada ve seçimde uzlaştırıcı çözümü bulabilmektir (Akyüz, 2012).

Yöntemlerin uygulanması ve nihai testte madde seçim işleminin gerçekleştirilebilmesi için uzmanlardan oluşan değerlendirme komitesi, madde havuzu ve maddelerin değerlendirmesinde dikkate alınacak değerledirme krtiterlerine ihtiyaç vardır. Gerekli olan verilerin hepsi Microsoft Excel programı üzerinde sümilatif veriler oluşturularak elde edilmiştir. Değerlendirme komitesinde yer alan beş uzman adına havuzdaki maddeler, kazanıma uygunluk, madde ayırt ediciliği ve madde güçlüğü kriterleri altında Chen (2000) tarafından belirlenen dilsel değişkenler kullanılarak değerlendirilmiştir. Dilsel değişkenler üçgen bulanık sayılara dönüştürülerek bulanıklaştırılmıştır. Değerlendiricilerin kriterlere ilişkin değerlendirmeleri, kriterlerin önem ağırlıkları ve her uzmanın alternatifler hakkındaki görüşleri birleştirilerek bulanık karar matrisi oluşturulmuştur. Elde edilen matris üzerinden bulanık TOPSIS yöntemi ile yakınlık katsayıları ve bulanık VIKOR yöntemi ile karar indeksi hesaplanarak alternatif maddeler sıralanmıştır. Yöntemlerin madde havuzundan madde seçiminde kullanılmasında benzerlikleri, farklılıkları ve üstünlükleri araştırılmıştır.

Sonuçlar

Bulanık TOPSIS yönteminde alternatif maddelerin hem Chen (2000) tarafından önerilen standartlaştırılmış hem de bulanık karar matrisinden elde edilen standartlaştırılmamış pozitif ve negatif ideal kümelerine göre yakınlık katsayıları hesaplandığında iki yöntemle de seçilen maddelerin aynı olduğu tespit edilmiştir. Standartlaştırılmamış çözüm kümelerine göre yapılan analizde maddelerin indeks değerleri arasındaki fark standartlaştırılmış çözüm noktaları dikkate alınan analize göre daha büyük bulunduğu için bu yöntemle maddeler arasında seçim yapma işleminin daha kolay yapılabileceği düşünülmektedir.

Bulanık VIKOR yöntemiyle alternatif maddelerin sıralanması için maksimum grup faydası stratejisinin ağırlığını gösteren v değeri sırasıyla "0.00, 0.50 ve 1.00 alınarak hesaplanan $\tilde{Q}_j$ indeksinin indeks değeri sorununca her üç sıralamada ilk sırada aynı maddenin yer aldığı ve bu sonucun Bulanık TOPSIS yöntemi ile tutarlı olduğu tespit edilmiştir. Alanyazında uzlaşmacı çoğunluk için önerilen v değerinin 0.50 olarak alındığı $\tilde{Q}_j$ indeksi sıralamasının Bulanık TOPSIS yöntemi ile yapılan sıralamayla büyük oranda benzerlik gösterdiği ancak sıralamada yer alan dört maddenin farklı bulunduğu tespit edilmiştir.

İki yöntem sonucunda belirlenen en iyi maddenin aynı olduğu ve sıralamada yer alan diğer maddeler kontrol edildiğinde seçilen maddelerin büyük oranda benzerlik gösterdiği tespit edilmiştir. Yöntemlerin madde seçim problemine kolayca uygulanabilir olduğu, karar alma sürecinin içerdiği belirsizlik ve objektiflik probleminin çözümünde bilimsel temellere dayalı çözüm arayışı öne sürdüğü sonucuna ulaşılmıştır.

Kaynaklar

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Akyüz, G. (2012). Bulanık VIKOR yöntemi ile tedarikçi seçimi. *Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi, 26*(1), 197-214.

Ateş, N. Y. Çevik S., Kahraman C., Gülbay M., & Erdoğan S. (2006). Multi attribute performance evaluation using a hierarchical fuzzy TOPSIS method. *Studies in Fuzziness and Soft Computing, 20(1), 537-572.* https://doi.org/10.1007/3-540-33517-X_22

Chen, C. T. (2000). Extensions of the TOPSIS for group decision-making under fuzzy environment. *Fuzzy Sets and Systems, 114*(1), 1-9.

Crocker, L., Algina J. (1986). *Introduction classical and modern test theory.* Harcourt Brace Javonovich College Publishers.

Çetin, B. (2019). Test geliştirme. Bayram Çetin (Ed.), *Eğitimde ölçme ve değerlendirme* (1. baskı) içinde. Anı Yayıncılık.

Downing S. M., & Haladayna T. M. (2006). *Handbook of test development.* Lawrence Erlbaum Associates Publıshers.

Karasar, N. (2010). *Bilimsel araştırma yöntemi.* Nobel Akademik Yayıncılık.

Paksoy, T., Yapıcı Pehlivan N. ve Özceylan E. (2013). *Bulanık matematiksel programlamaya giriş: Bulanık küme teorisi.* Nobel Akademik Yayıncılık.

Tekin, H. (2018). *Eğitimde ölçme ve değerlendirme* (27. baskı). Yargı Yayınevi.

Özkan, M. M. 2003. *Bulanık hedef programlama.* Ekin Kitabevi Yayınları.

PISA 2018'de okuduğunu anlama başarısını yordayan değişkenlerin eğitsel veri madenciliği sınıflama ve regresyon ağacı ile belirlenmesi

Yusuf Kasap ve Nuri Doğan

Anahtar kelimeler: Okuduğunu anlama, veri madenciliği, sınıflama, tahmin, PISA

Giriş

Eğitimde de son yıllarda hem uluslararası düzeyde hem de ulusal düzeyde yapılan çalışmalarda çok değişkenli veri toplama eğilimi yaygınlaşmıştır. Bu sınavlardan birisi de Uluslararası Öğrenci Değerlendirme Programıdır (PISA). PISA verileri kullanılarak, çok değişkenli VM modelleri ile başarıyı yordayan veya başarı ile ilişkili bağımsız değişkenlerin hangileri olduğu saptanabilmektedir.

Alanyazında, VM yöntemlerini kullanarak başarıyı yordayan veya başarıyla ilişkili olan değişkenlerin belirlenmesine yönelik birçok çalışma bulunmaktadır. Bu çalışmalardan bazıları PISA başarı düzeyini yordayan değişkenleri tespit etmeye yöneliktir (Abad ve Lopez, 2016; Aksu ve Güzeller, 2016; Gamazo ve Abad, 2020; Kıray ve diğ., 2015; Yung ve diğ., 2012). Bununla birlikte Bezek Güre ve diğ. (2020) yaptıkları çalışmada, SPSS Modeler ve Matlab programlarını kullanarak, başarıyı en yüksek düzeyde yordayabilen önemli değişkenleri belirlemişlerdir. Bezek Güre ve arkadaşlarının çalışması (2020) alanyazındaki diğer çalışmalardan farklıdır. Alanyazındaki çalışmaların neredeyse tamamında başarıyı etkileyen faktörleri belirlemede, veri madenciliği algoritmalarından elde edilen çıktılara göre karar verilirken, Bezek Güre ve diğ. (2020)'nın yaklaşımında model performansını en yüksek yapan değişkenler en iyi yordayıcılar olarak görülmektedir. Ulusal alanda, PISA verilerini kullanılarak başarıyı yordayabilen önemli yordayıcıları belirlemeyi amaçlayan VM performansına dayalı olarak Bezek Güre ve diğ., (2020) çalışmasının dışında bir çalışmaya taranan kaynaklarda rastlanmamıştır. Bezek Güre ve diğ., (2020) yaklaşımlarının alanyazındaki çalışmaların çoğunluğunda kullanılan VM modellerinden elde edilen çıktılara göre önemli değişkenlerin belirlenmesi yaklaşımından daha pratik olduğu söylenebilir. Çünkü; önemli değişkenlerin kullanılan modellerden elde edilen çıktılar yorumlanarak belirlenmesinde, VM tahmin modelinin performansının ne olduğu dikkate alınmamaktadır. Böylece; performansı yeterince yüksek olmayan modeller kullanılarak önemli değişkenler belirlenebilmektedir. Oysaki Bezek-Güre ve diğ., (2020) önemli değişkenlerin saptanması yaklaşımında performansı yüksek olan bir VM tahmin modeline göre karar verilmektedir. Diğer yandan alanyazında yapılan araştırmalar incelendiğinde daha çok tek bir ülke ile sınırlı olan homojen gruplarla yapılan çalışmalar yoğunluktadır.

Farklı başarı gruplarında ve bu gruplarının birleştirilmesiyle oluşan heterojen gruplarda VM yöntemlerinden elde edilen sonuçların değişip değişmediğini, yani yordayıcı değişkenlerin ve sayısının değişip değişmediğine yönelik çalışmaya rastlanmamıştır. Aslında veriden öğrenen (akıllı) bir yöntem olduğu düşünülen VM yöntemlerinin veri (örneklem) değiştikçe farklı sonuçlar verebileceğinin araştırılması önem teşkil etmektedir. Alanyazında PISA başarı puanını yordayan önemli değişkenlerin belirlenmesi çalışmalarında kullanılan örneklemlerin de dar kapsamlı olduğu görülmektedir. Örneğin; Aksu ve Güzeller (2016) ve Bezek Güre ve diğ., (2020) çalışmalarında sadece PISA Türkiye örneklemini kullanmışlardır. Yapılan çalışmalar başarı veya performansı yordayan değişkenler üzerinde yoğunlaşmakta ancak bu çalışmalar birbirine benzer şekilde fen, matematik veya okuma puanını bağımlı değişken kabul ederek sözü edilen üç alandaki başarıyı etkileyen faktörlerin açığa çıkarılmasına yöneliktir. Ancak bu çalışmalar farklı başarı düzeyleri dikkate alınarak yapılan çalışmalar değildir. Oysa farklı alt başarı gruplarında farklı değişkenlerin etkili olabileceğinden dolayı bu iddiayı farklı başarı gruplarında çalışılarak doğrulamak önemlidir. Dolayısıyla bu araştırmada, PISA 2018 okuduğunu anlama başarısına yönelik olarak olarak, her biri 3'er ülkeden oluşan alt, orta ve üst başarı grubuna sahip ülkeler ve bu ülkelerin birleştirilmesiyle oluşturulan çalışma örneklemi, dikkate alınmıştır. Böylece, bu çalışmada seçilen örneklemlerin çeşitli olması, farklı başarı gruplarını dikkate alınması, homojen ve heterojen gruplara ilişkin sonuçlar elde edilmesi ve örneklemin yeterince büyük olması nedeniyle evrenin temsil edilme ve sonuçların genelenebilirlik düzeyinin artması beklenebilir.

Bu araştırmanın temel amacı, PISA öğrenci anketinde bulunan 34 bağımsız değişkene ilişkin puanları kullanarak farklı başarı düzeyine sahip (alt, orta ve üst başarı grubundaki) ülkelerin okuduğunu anlama başarısını etkileyen önemli yordayıcıları belirlemektir. Çalışmanın diğer amacı ise; belirlenen az sayıda daha önemli bağımsız değişkenleri kullanılarak performansı yüksek sınıflama modellerinin elde edilip edilmeyeceğini göstermektir. Böylece; çalışmanın, VM performansını en yüksek yapan daha önemli bağımsız değişkenleri seçerek değişken sayısını azaltmaya yönelik az sayıda çalışmadan biri olması hedeflenmektedir.

Yöntem

Araştırmada öğrencilerin duyuşsal özelliklerini ölçen ölçekler ve sosyodemografik özelliklerini ölçen anketlerle alt, orta, üst başarı grubundan seçilen ülkeler, seçilen ülkelerin birleştirilmesiyle oluşturulan PISA çalışma örneklemi ve bağımsız olarak Türkiye örnekleminin başarı düzeylerinin PISA okuduğunu anlama puanlarına göre (başarılı-başarısız) tahmin yapılması açısından etkili olan değişkenlerin belirlenmesi amaçlandığından araştırmanın türü ilişkisel nicel araştırma tasarımı olarak ifade edilebilir (Büyüköztürk ve diğ., 2018).

Çalışmanın amacı için ülkeler; PISA okuma puanı sıraları dikkate alınarak alt (zayıf), orta, üst (yüksek) başarı grubuna göre belirlendiği için amaçlı örnekleme yöntemi kullanılmıştır (Büyüköztürk ve diğ., 2018). Ayrıca, Çalışma örneklemine ülkeler seçilirken kayıp veri oranının düşük olmasına ve farklı yüzdelik dilimlere dağılacak şekilde örüntü oluşturmasına dikkat edilmiştir.

Araştırma sürecinde kullanılan veriler 2020 yılında paylaşıma açılan veri tabanından http://www.oecd.org/pisa/data/2018database/ adresi kullanılarak elde edilmiştir. Araştırma kapsamına 2018 PISA verilerine ait toplamda 34 bağımsız ve okuduğunu anlama başarı düzeyine karşılık gelen on makul değerin ortalaması olan bir bağımlı değişken alınmıştır.

Araştırmada veriler 2018 yılında uygulanan PISA okuduğunu anlama sınavındaki test ve öğrenci anketinden elde edilmiştir. Kullanacağımız bağımlı değişkeni elde etmek için öncelikle okuduğunu anlama başarı puanı sürekli nicel değişken olarak elde edilmiş ve PVREAD olarak adlandırılmıştır. PVREAD değişkeni PISA 2018 okuduğunu anlama başarı testinden her bir öğrencinin aldığı 10 makul puan değerinin (PV1, PV2, …, PV10) ortalamasıdır. Daha sonra, PVREAD puanının sınıflandırılması sürecinde PISA 2018 Türkiye ön raporundaki, okuma becerileri yeterlilik düzeyi tablosu kullanılmıştır (MEB, 2019). PVREAD değeri 0-552,999 arası düşük ve 553-1000 arası yüksek olacak şekilde sınıflanarak elde edilmiştir.

Verilerin analizinde; öncelikle PISA'ya katılan 79 ülke başarı yüzdelik sırasına konulmuştur. Ardından eşit dilimlerle ülkeler alt, orta ve üst dilimler olarak ayrılmıştır. Yüzdelik dilimlerine ve kayıp veri azlığına göre Türkiye dahil 9 ülke seçilmiştir. Veriler analiz edilirken önce bağımsız olarak Türkiye verileri, sonra alt, orta ve üst başarı grupları ardından da seçilen ülkelerin birleştirilmiş verileri analiz edilmiştir. Araştırma kapsamında 34 değişken bağımsız değişken olarak ve iki düzeyli başarı değişkeni (0=başarısız, 1=başarılı) bağımlı değişken olarak alınmıştır. Bağımsız değişkenlerin bağımlı değişkeni sınıflama performansını belirlemek için CART (Sınıflama ve regresyon ağaçları) veri madenciliği sınıflama modeli kullanılmıştır. Öncelikle; SPSS Modeler programı kullanılarak 2 düzeyli başarı değişkenini yordayan önemli bağımsız değişkenler saptanmıştır. Elde edilen önemli bağımsız değişkenler kullanılarak Weka programında; eğitim verisi, 10 katlı çapraz geçerlik verisi ve test verisi ile geçerliğe ve güvenirliğe dayalı performans kriterleri hesaplanmıştır.

Sonuçlar

Türkiye, Alt, orta, üst grup ülkeler örneklemi ve çalışma örneklemi ile analiz yapılarak hem başarıyı yordayan önemli değişkenler hemde kullanılan modelin sınıflama performansları incelenmiştir. Yordayıcı değişkenlerin en önemliden önemsize doğru önem sırası Türkiye için okuma keyfi, öğrenciden beklenen mesleki statü ve sosyo-ekonomik düzey, alt grup ülkeler için PISA testinin zorluk algısı, yaşamın anlamı ve okulun değeri, orta grup ülkeler için sosyo-ekonomik düzey, PISA testinin zorluk algısı, okuma keyfi, ebeveynlerin en yüksek mesleki statüsü ve başarısızlık korkusu, üst grup ülkeler için PISA testinin zorluk algısı ve okuma keyfi, Çalışma örneklemi için PISA testinin zorluk algısı, sosyo-ekonomik düzey, yaşamın anlamı, okuma keyfi ve başarısızlık korkusu biçimindedir. Elde edilen sonuçlara göre grupların tamamında başarıyı yordama da önemli ortak yordayıcıların; PISA testinin zorluk algısı, okuma keyfi ve sosyo-ekonomik düzey olduğu belirlenmiştir. Diğer yandan modele alınan tüm değişkenler ve önemli değişkenlerle eğitim verisi, geçerlik verisi ve test verisi ile modelin grupların tamamında hesaplanan sınıflama performansları yüksek düzeyde çıkarken değişken sayısının azalmasının sınıflama performansını

önemli düzeyde etkilemediği gözlenmiştir. Buna ek olarak, Kappa istatistiği alt grupta düşük diğer gruplarda kabul edilebilir düzeyde çıkmıştır. Sonuç olarak modelin eğitim verisi, geçerlik verisi ve test verisi için her bir grupta önemli bağımsız değişkenlerle hesaplanan sınıflama performansı ile modele alınan tüm bağımsız değişkenlerle hesaplanan sınıflama performansı çok yakın çıkmıştır.

Kaynaklar

Abad, F. M., and Lopez, A. C. (2016). Data-mining techniques in detecting factors linked to academic achievement. *School Effectiveness and School Improvement, 28*(1), 39-55.

Aksu, G., and Güzeller, C.O. (2016).Classification of PISA 2012 mathematical literacy scores using decision-tree method: Turkey sampling. *Education and Science, 41*(185), 101-122.

Bezek Güze, Ö., Kayri, M., and Erdoğan, F. (2020). Analysis of factors effecting PISA 2015 mathematics literacy via educational data mining. *Education and Science 45*(202), 393-415.

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş. ve Demirel, F. (2018). *Bilimsel araştırma yöntemleri*. Pegem Yayıncılık.

Gamazo, A., & Abad, F. M.(2020). An exploration of factors linked to academic performance in PISA 2018 through data mining techniques. *Journal Frontiers in Psycholog, 11,* 575167. https://doi.org/10.3389/fpsyg.2020.575167

Kiray, S.A., Gok, B., & Bozkır, A. S. (2015). Identifying the factors affecting science and mathematics achievement using data mining methods. *Journal of Education in Science, Environment and Health (JESEH), 1*(1), 28-48.

MEB (2019). *PISA 2015 ulusal ön raporu.* MEB.

Yung, J. L., Hsu, Y. C., & Rice, K. (2012). Integrating data mining in program evaluation of k-12 online education. *Journal of Educational Technology & Society, 15*(3), 27-41. http://www.jstor.org/stable/jeductechsoci.15.3.27

Açık uçlu maddelerin puanlanmasında bulanık mantık yaklaşımının kullanımı: Bulanık TOPSIS yöntemi örneği

Aykut Çitci ve Fatih Kezer

Anahtar kelimeler: Bulanık mantık, açık uçlu madde puanlama, bulanık TOPSIS

Giriş

Bir eğitim sisteminin başarılı yönlerinin ve başarısızlık kaynaklarının bilinmesi sistem hakkında önlem alınmasını ve gelecek eğitim etkinlikleri için daha doğru planlamalar yapılmasını kolaylaştırır. Bu anlamda ölçme ve değerlendirme; eğitim sisteminin izlenmesi, kontrol edilmesi ve geliştirilmesi açısından büyük bir öneme sahiptir (Turgut ve Baykul, 2010). Rowntree (1987), eğitim sistemleri hakkında gerçeklerin öğrenilebilmesi için öncelikle ölçme değerlendirme süreçlerinin incelenmesinin gerekli olduğunu ifade ederek bu ögenin önemine değinmiştir.

Eğitimde akademik başarının düzeyini belirlemede yazılı yoklamalar, testler, sözlü yoklamalar vb.; duyuşsal alan davranışlarını ölçmek için ilgi envanterleri, tutum ölçekleri, likert tipi araçlar vb.; devinişsel alan davranışlarını ölçmede de dereceleme araçları, çeklist vb. araçlar kullanılmaktadır (Tekindal, 2017). Okullarda en çok kullanılan ölçme araçlarından biri çoktan seçmeli testlerden sonra açık uçlu maddelerden oluşan testlerdir. Bu maddelerin hazırlanmasının çoktan seçmeli maddelere göre nispeten kolay olması, çoğu öğretmenin bu tarzda madde hazırlamaya alışık olması ve öğrencilerin bilgi birikimleriyle kendini daha iyi ifade etmelerine olanak sağlaması, açık uçlu maddelerin kullanılmasına gerekçe olarak sunulabilir. Açık uçlu maddelerden oluşan sınavlar genellikle az soru sayısından oluşmakta bu da bu tür sınavlardan elde edilen güvenirlik ve geçerlik katsayılarının çoktan seçmeli testlere göre düşük kalmasına sebep olmaktadır. İki ölçme aracı sonucunda elde edilen bu verilerin farklı olmasının bir diğer sebebi de çoktan seçmeli testlerin objektif olarak puanlaması ama açık uçlu maddelerin dereceli puanlama anahtarı kullanılmasına rağmen subjektiflikten tam olarak kurtulamaması gösterilebilir.

Doğru–yanlış, çoktan seçmeli, kısa cevap gerektiren yani objektif olarak puanlanan maddelerin daha çok klasik puanlamaya uygun olduğu düşünülürken öğrenci cevaplarının öğretmenler tarafından subjektif olarak değerlendirildiği açık uçlu maddelerin bulanık mantıkla puanlanmaya daha uygun olduğu düşünülmektedir. Bulanık mantık yönteminde açık uçlu madde puanlamak için oluşturulmuş olan kriterlerin ağırlıkları ve öğrenci cevapları bulanık kümeler yani kesinlik içermeyen değerler

kullanılarak hesaplanmaktadır. Oysa klasik yöntemlerde kriterlerin ağırlıkları ve öğrenci cevapları kesin sayılar verilerek hesaplanmakta bu da esnekliği yitirerek puanları kesin formlara dönüştürmektedir.

Eğitim sistemimizde açık uçlu maddeler halihazırda klasik yöntemler kullanılarak puanlanmaktadır. Bu çalışmada açık uçlu maddelerin bulanık mantık yöntemi kullanılarak puanlanmasının klasik puanlamaya göre gerçek hayata daha yakın sonuçlar elde edilip edilemeyeceği araştırılmak istenmiştir.

Yöntem

Çalışmada, öğrencilerin açık uçlu matematik maddelerine verdikleri cevapların, puanlayıcıların değerlendirmelerine göre sıralanma durumları nicel araştırma yöntemi kullanılarak araştırılmıştır. Çalışma açık uçlu matematik maddelerinin bulanık mantık yöntemiyle puanlanmasına yönelik bir model önerisi olması yönünden temel araştırma kapsamında değerlendirilebilir. "Temel araştırmalarda amaç, salt bilgi üretmektir. Bu tür araştırma modelleri anlama, açıklama ve kuram geliştirme düzeylerinde bilgi üretilebilir" (Tutar, 2013, s. 518).

Çalışmanın Covid-19 sebebi ile gerçek uygulaması okullarda yapılamamış olunup öğrenci cevaplarının bilgisayarda simülatif olarak oluşturulmasının bir problem oluşturmayacağına karar verilmiştir. Simülatif veri için ortalama sınıf mevcudu olarak öğrenci sayısı 25 olarak belirlenmiştir (MEB, 2020, s. 24).

Açık uçlu matematik maddelerinin puanlanması için kriterlerin belirlenmesinde öncelikle alanyazın çalışması yapılmış ve bazı alt kriterler oluşturulmuştur (Altun, 2002; Van De Walle ve diğ., 2019; Karadeniz, 2016; Damlar-Demirci, 2019). Daha sonra bu kriterler üç matematik alan uzmanıyla ayrı ayrı görüşülerek araştırma kapsamında değerlendirilerek yedi kriter olarak belirlenmiştir. Alt kriterlerin ağırlıklarının belirlenmesi için, oluşturulmuş olan kriterler uzmanlara gönderilmiştir. Uzmanlar bu kriterlere kendi bireysel görüşleri doğrultusunda; çok düşük, düşük, biraz düşük, orta, biraz yüksek, yüksek ve çok yüksek değerlerini atamışlardır.

Çalışmada birinci yöntem olarak okullarda kullanılan klasik yöntemle öğrenci sıralamaları elde edilmiştir. Bu yöntemde alt kriterlerin ağırlıkları kullanılmamış olunup puanlayıcıların vermiş oldukları toplam puanların ortalamaları alınarak sıralamalar yapılmıştır. İkinci yöntemde ise Opricovic ve Tzeng (2004) tarafından önerilen işlem adımları ile çok kriterli karar verme yöntemi olarak klasik TOPSIS kullanılmıştır. Çalışmada üçüncü yöntem olarak kullanılan Bulanık TOPSIS yönteminde ise Chen (2000) tarafından önerilen işlem adımları kullanılarak öğrenci sıralamaları elde edilmiştir. Açık uçlu matematik maddelerinin klasik, TOPSIS ve bulanık TOPSIS yöntemleri ile öğrenci sıralamaları elde edilmiştir. Sıralamalar arasındaki korelasyon değerlerinin incelenmesi için Spearman sıra farkları korelasyon katsayısı hesaplanmıştır.

Sonuçlar

Çalışmada kullanılan çok kriterli karar verme yöntemlerinde öğrenci puanları birbiriyle eşit olmazken okullarda kullanılan klasik yöntemin kullanıldığı sıralamada bazı öğrencilerin aynı puanları aldıkları görülmektedir. Bu da klasik yöntemin kullanım kolaylığının yanında daha az duyarlı ölçümler yaptığını göstermektedir.

Öğrencilerin sıralamaları değerlendirildiğinde çoğu öğrencinin kullanılan yöntemlere göre sıralamalarının değiştiği görülürken, Öğrenci 16 ve Öğrenci 11'in kullanılan tüm yöntemlerde sırasıyla ilk ve son sırada yer aldığı görülmektedir. Bunun nedeni olarak Öğrenci 16'nın almış olduğu ortalama puanların açık uçlu matematik değerlendirme kriterlerinde diğer öğrencilere göre daha yüksek olduğu, Öğrenci 11'in ise görece daha düşük olduğu söylenebilir. Öğrenci sıralamaları incelendiğinde en fazla değişimin beş sıra ile Öğrenci 21'de olduğu görülmüştür.

Araştırma kapsamında ele alınan Bulanık TOPSIS yöntemi kullanılarak öğrencilerin açık uçlu matematik sorularından aldıkları puanlar belirlenmiş ve araştırmanın amacı kapsamında klasik puanlarla nasıl bir ilişki gösterdiği belirlenmiştir. Bunun için Spearman sıra farkları korelasyon katsayısı kullanılmış ve bulanık TOPSIS ile klasik yöntemler arasındaki benzerlik katsayıları klasik yöntem ile r= .984, TOPSIS ile r= .975 olarak bulunmuştur (p< .01, n=25). Bu durum yöntemler arasında öğrenci sıralamalarının güçlü ve pozitif bir ilişki olduğunu göstermektedir. Elde edilen bu sonuç açık uçlu maddelerle öğrenci sıralama yöntemi olarak klasik, TOPSIS ve bulanık TOPSIS yöntemlerinin birbiri yerine kullanılabileceğini göstermektedir.

Kaynaklar

Altun, M. (2002). *İlköğretim ikinci kademede matematik öğretimi* (2. baskı). Alfa.

Chen, T. C. (2000). Extensions of the TOPSIS for group decision - making under fuzzy environment. *Fuzzy Sets And Systems, 114*(1), 1-9. https://doi.org/10.1016/S0165-0114(97)00377-1

Damlar-Demirci, P. (2019). *Açık uçlu soruların puanlama yöntemlerinin genellenebilirlik kuramına göre incelenmesi* (Tez No. 588976) [Yüksek lisans tezi, Anadolu Üniversitesi] Yükseköğretim Kurulu Tez Merkezi.

Karadeniz, A. (2016). *Kitlesel açık ve uzaktan öğrenmede başarının açık uçlu sorularla ölçülmesine yönelik bir sistemin tasarımı, uygulanması ve değerlendirilmesi* (Tez No. 449995) [Doktora tezi, Anadolu Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

MEB. (2020). *Milli eğitim istatistikleri: Örgün eğitim 2019-20*. Millî Eğitim Bakanlığı Strateji Geliştirme Başkanlığı.

Opricovic, Serafim, Gwo – Hshiung Tzeng (2004). Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS. *European Journal of Operational Research, 156*(2), 445 – 455.

Rowntree, D. (1987). *Assesing students: How shall we know them?* Kogan Page.

Tekindal, S. (2017). *Okullarda ölçme ve değerlendirme yöntemleri* (6. baskı). Nobel Akademi Yayıncılık.

Turgut, M. F. ve Baykul, Y. (2010). *Eğitimde ölçme ve değerlendirme*. Pegem Akademi.

Tutar, H. (2013). *İşletme & yönetim terimleri ansiklopedik sözlük*. Detay Yayıncılık.

Van De W., John A., Karen S. K., & Jennifer M. Bay-Williams (2021). *İlkokul ve ortaokul matematiği: Gelişimsel yaklaşımsal öğretim [Elementary and middle school mathematics - teaching developmentally]* (S. Durmuş, Çev. 10. baskı). Nobel Akademik (2019).

Efficient abbreviation of lengthy scales using genetic algorithms

Hatice Çiğdem Bulut, Betül Doğan Laçin and Çağla Alpayar

Introduction

To abbreviate scales, there are several options in the literature however each option comes with its limitation. For example, traditional approaches (e.g., selecting the items with higher discrimination or with the highest factor loadings) need so much time to test all meaningful possible item sets and require judgment from researchers in all steps. To overcome these problems, metaheuristic approaches (e.g., genetic algorithms, tabu research, and ant colony optimization) have gained great interest in literature in constructing abbreviated scales. Among these approaches, genetic algorithms show promising solutions in abbreviating lengthy scales (Schroeders et al., 2016).

GA is an automated and heuristic approach that uses a large number of possible abbreviated scales to find an optimal shorter scale that maximizes the variance in the original scale (e.g., Sahdra et al., 2016; Schroeders et al., 2016; Yarkoni, 2010). GA originally mimics the theory of evolution using natural selection and crossover procedures in biology. Many researchers have utilized GA to abbreviate to create psychologically sound shorter scales (e.g., Crone et al., 2020, Eisenbarth et al., 2015; Noetel et al., 2019; Sahdra et al., 2016). In this study, we aimed to abbreviate two scales frequently utilized in psychological research (e.g., Putnam and Rothbart, 2006; Van-Leeuwen and Vermulst, 2004).

Method

The sample consisted of N= 516 parents. Their average age was M = 35.15 years (SD = 18.53), and 74.22% were women. The Temperament Scale in Children (TSC) and the Parenting Behavior Scale (PBS) (Dogan Lacin, 2021) were used for this study. TSC is a 25-item scale and PBS is a 28-item scale. These scales have been shown to have good construct validity and high internal consistency.

Data were analyzed in four steps. We first implemented the genetic algorithm (GA) procedure using the GAabbreviate package (Scrucca and Sahdra, 2015) in R (R Core Team, 2021). The main aim of GAs is to select strong items/variables in a given data set by eliminating relatively weaker items or variables. For the case of scale abbreviation, GAs tries to select a subset of items that maximize possible variance as in the original scale. For this study, we set the item-cost parameter ("itemCost" = .05") to 0.05 and wanted to have up to 6 items per factor ("maxItems = 6") for both scales. Second, we examined

the construct validity of the abbreviated forms of the scales by using confirmatory factor analysis (CFA). In the third step, we examined the criterion-related validity of the abbreviated forms. We examined relationships between all factors in the abbreviated forms of TSC and PBS. In the final step, we calculated alpha coefficients and McDonald's omega for all factors to gather reliability evidence.

Results

We run GAs several times to find the most suitable (i.e, most frequently selected item by GAs) item sets and maintain a similar conceptual structure in the abbreviated forms of the scales. The results revealed 14 items in TSC and ten items in PBS. These items were consistently selected by GAs and showed better factor representability than other items. The abbreviated forms of TSC and PBS performed very similarly in terms of construct validity, criteria-related validity, and reliability coefficients compared to the original scales of TSC and PBS.

References

Crone, D. L., Rhee, J. J., & Laham, S. M. (2020). Developing brief versions of the moral foundations vignettes using a genetic algorithm-based approach. *Behavior Research Methods*, *53*(3), 1179-1187. https://doi.org/10.3758/s13428-020-01489-y

Doğan-Laçin, B. G. (2021). *Okulöncesi dönemde çocuğu olanların ana babalık davranışlarının çocuk ve ana baba özellikleri açısından incelenmesi* (Tez No. 655048). [Doktora tezi, Ankara Üniversitesi] Yükseköğretim Kurulu Tez Merkezi.

Eisenbarth, H., Lilienfeld, S. O., & Yarkoni, T. (2015). Using a genetic algorithm to abbreviate the psychopathic personality ınventory–revised (PPI-R). *Psychological Assessment*, *27*(1), 194. https://doi.org/10.1037/pas0000032

Noetel, M., Ciarrochi, J., Sahdra, B., & Lonsdale, C. (2019). Using genetic algorithms to abbreviate the mindfulness inventory for sport: A substantive-methodological synthesis. *Psychology of Sport and Exercise*, *45*, 101545. doi:10.1016/j.psychsport.2019.101545

Putnam, S. P., & Rothbart, M. K. (2006). Development of short and very short forms of the Children's Behavior Questionnaire. *Journal of Personality Assessment, 87(1)*, 103-113. https://doi.org/10.1207/s15327752jpa8701_09.

R Core Team (2021). *R: A Language and environment for statistical computing* (version 4.0) [Computer software]. https://cran.r-project.org.

Sahdra, B. K., Ciarrochi, J., Parker, P., Basarkod, G., Bradshaw, E., & Baer, R. (2016). Are people mindful in different ways? Disentangling the quantity and quality of mindfulness in latent profiles, and exploring their links to mental health and life effectiveness. *European Journal of Personality, 31*(4), 347–365. https://doi.org/10.1002/per.2108

Schroeders, U., Wilhelm, O., & Olaru, G. (2016). Meta-heuristics in short scale construction: Ant colony optimization and genetic algorithm. *PLoS One*, *11*(11), e0167110. https://doi.org/10.1371/journal. pone.0167110

Scrucca, L., & Sahdra, B. (2015*). GAabbreviate: Abbreviating questionnaires (or other measures) using genetic algorithms* (version 1.0) [Computer software]. https://cran.r-project.org/package=GAabbreviate

Van Leeuwen, K. G., & Vermulst, A. A. (2004). Some psychometric properties of the Ghent parental behavior scale1. *European Journal of Psychological Assessment*, *20*(4), 283-298. https://doi.org/10.1027/1015-5759.20.4.283

Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales with 200 items. *Journal of research in personality*, *44*(2), 180-198. doi:10.1016/j.jrp.2010.01.002

PISA 2018 özyeterlik ölçeği maddelerine verilen tepkiler ile okuma performansı ve başarısızlık kaygısı arasındaki ilişkinin incelenmesi

Ömer Kutlu ve Çağla Alpayar

Anahtar Kelimeler: Özyeterlik, PISA, öğrenci başarısı, okuma becerisi, başarısızlık kaygısı

Giriş

PISA çalışmalarında, katılımcı ülkelerin 15 yaş grubu öğrencilerine okuma, matematik ve fen testleri ile anketler uygulanmaktadır. Anketler ile okul, öğrenci ve öğretmen özelliklerine dayalı belirlemeler yapılmaktadır (OECD, 2019a). PISA anketlerde yer alan bazı maddeler için farklı bir süreç izlemekte, aynı özelliği ölçtüğü kabul edilen az sayıda madde ile tek boyutlu bir yapı oluşturmaktadır (OECD, 2019b). Özyeterlik, bu yaklaşımla oluşturulan yapılardan biridir.

Özyeterlik, bireyin bir görevi gerçekleştirme yeteneğine inanma derecesidir (Bandura, 1977). Jinks ve Morgan (1999), özyeterliği belirli görevlerin başarılmasıyla ilgili bireyin güven algısı olarak değerlendirmiştir. PISA 2018 teknik raporunda özyeterlik ölçeğine yönelik kuramsal bir tanımlama bulunmamakta, ölçekteki beş maddenin genel yapıya bağlı olarak geliştirildiği belirtilmektedir (OECD, 2019c). Önceki uygulamalarda, matematik ve fen dersleri özelinde özyeterlik inançları belirlenirken 2018 uygulamasında genel bir tanımlama yapılmış ve öğrencilere güç durumlarda kaldıklarında sahip oldukları özyeterlik inançları sorulmuştur.

PISA teknik raporunda, ülkelerin özyeterlik düzeyleri hem indeks puan ortalamaları hem de madde düzeylerine dağılımları yönünden raporlanmıştır (OECD, 2019d). Ancak ortalama puanlara dayalı çözümlemelerden elde edilen bulgular tutarsız bilgiler vermekte, madde düzeyindeki çözümlemelerle daha anlamlı bilgiler elde edilebilmektedir (McGrath, 2014). Özellikle puanların normal dağılmadığı durumlarda, veriyi tanımlamak için yanıtların sıklık dağılımının incelenmesi daha bilgi verici olmaktadır (Sullivan ve Artino, 2013). Ayrıca, maddelerin tepki düzeylerine dağılımı farklılık gösteren aynı puan ortalamasına sahip iki madde benzer biçimde yorumlanmaktadır. Ancak düzeylerdeki tepkiler özelinde çözümlendiğinde farklı bulgulara ulaşılabilmektedir (Clason ve Dormody, 1994). Bu nedenle PISA testlerini oluşturan okuma, fen ve matematik başarısı üzerinde etkili olan örneğin özyeterlik gibi değişkenleri ölçen ölçek maddelerinin tepki düzeyleri özelinde incelenmesi, özellikle kültürlerin karşılaştırılmasında önemli bilgiler sunacaktır.

PISA teknik raporunda, özyeterlik ölçeği maddelerinin tepki düzeylerine yönelik yorumlar, öğrenci yüzdelerinin dağılımı ile sınırlıdır (OECD, 2019d). Öğrencilerin maddelerin her bir düzeyini tercih etme durumu ile test puanları ve başka psikolojik özellikler arasındaki ilişkilerin incelenmesi, daha kapsamlı yorumlamalara olanak verecektir. Bu yorumlar, ülkeler özelinde ölçülen özellikle ilgili farklılıklara yönelik ipuçları sunacaktır. Bu bağlamda düzeylere dağılımın ayrıntılı incelenebilmesi için madde ayırt edicilik indeksi hesaplamasına benzer bir yaklaşım izlenebilecek ve bireyin kendine uygun madde düzeyini tercih etme durumunun yorumlanmasında kullanılabilecektir. Böylelikle yalnızca maddenin değil madde düzeylerinin de işleyişi hakkında bilgiler edinilebilecektir.

Klasik Test Kuramı'na (KTK) göre madde ayırt ediciliği, bir maddenin içinde bulunduğu testle arasındaki ilişkinin ölçüsüdür (Crocker ve Algina, 1986). Madde Tepki Kuramı'nda (MTK) ise yanıtlayıcıların düzey seçimi, KTK'dan farklı olarak ilgili model parametreleri kullanılarak kestirilmektedir (Haberman ve diğ., 2015). Muraki (1990), seçenekler arasında hiyerarşinin bulunduğu ölçme araçlarına ait parametre kestirimlerinde; Aşamalı Tepki Modeli, Kısmî Puanlama Modeli veya Sıralı Tepki Model'inin daha uygun olduğunu ifade etmektedir. İki parametreli MTK modellerinin bir karşılığı olan Aşamalı Tepki Modeli, madde parametrelerinin eşik değerlerine bağlı olarak yorumlanmakta (Reise ve Yu, 1990) ve madde parametreleri ile yanıtlayıcının yeteneğine bağlı olarak belli bir düzeyde yanıt verme olasılığı hesaplanmaktadır (Embretson ve Reise, 2000; Symth, (n.d.)). KTK'ya ve MTK'ya dayalı karşılaştırmaların yapıldığı bazı çalışmalarda (Demirtaşlı ve diğ., 2016; Köse, 2015; Uyar ve diğ., 2013), seçenek düzeyine inilmeden yalnızca madde düzeyinde yorumlar yapılmıştır. Ayrıca bu araştırmalara dayalı çözümlemelerde yetenek düzeyi, araçla ölçülen özellikle sınırlı kalmıştır. Bu nedenle düzey tercihini testle ölçülen dışında bir özellikle ilişkilendirebilen KTK çözümlemeleri daha kapsamlı yorumlar verecektir.

PISA gibi uluslararası çalışmalarda ülkeler arası madde düzeyinde yapılacak bir karşılaştırma, aracın psikometrik özelliklerinin yanı sıra hem psikolojik yapı hem de yapının kültürlerdeki işleyiş farklılıkları hakkında bulgular sunacaktır. Maddenin hangi düzeyde daha duyarlı ölçümler sunduğunun ülkeler özelinde yorumlanmasını da sağlayacaktır. Bu araştırmada PISA 2018 öğrenci anketindeki "Özyeterlik Ölçeği"nin her bir maddesine verilen tepkiler ile öğrencilerin özyeterlik, okuma becerisi ve başarısızlık kaygısı düzeyleri arasındaki ilişki incelenerek aşağıdaki sorulara yanıt aranacaktır:

Farklı ülkelerdeki öğrencilerin özyeterlik ölçeğindeki;

1. Maddelere verdikleri tepkilerin düzeylere dağılımları nasıldır?
2. Maddelere verdikleri tepkiler ile ölçek puanları arasında ilişki var mıdır?
3. Maddelere verdikleri tepkiler ile okuma performansları arasında ilişki var mıdır?
4. Maddelere verdikleri tepkiler ile başarısızlık kaygısı arasında ilişki var mıdır?
5. MTK'ya dayalı belirlenen psikometrik özellikler, KTK'ya dayalı olanlardan farklılık göstermekte midir?

Yöntem

Bu araştırma, geçmişte ya da halen var olan bir durumu olduğu şekliyle betimlemeyi hedefleyen tarama modelindedir (Karasar, 2012).

Karşılaştırma için Türkiye'nin yanı sıra Filipinler, Endonezya, Finlandiya ve Hong Kong olmak üzere dört farklı ülke belirlenmiştir. Ülkelerin seçilmesinde, okuma başarısı ve akademik özyeterlik puan ortalamalarının Türkiye'ye göre durumları dikkate alınmıştır.

Türkiye'ye göre;

- Filipinler'in okuma puanı ortalaması düşük, özyeterlik puan ortalaması yüksektir.
- Endonezya'nın okuma puanı ortalaması düşük, özyeterlik puanı ortalaması düşüktür.
- Finlandiya'nın okuma puanı ortalaması yüksek, özyeterlik puanı ortalaması yüksektir.
- Hong Kong'un okuma puanı ortalaması yüksek, özyeterlik puanı ortalaması düşüktür.

PISA 2018 çalışmasında kullanılan ve özyeterlik ölçeğini oluşturan beş madde için, yanıtlayıcılardan bu maddelerin kendilerini ne kadar yansıttığını belirlemeleri istenmiştir (OECD, 2019c). Maddeler, "kesinlikle katılmıyorum", "katılmıyorum, katılıyorum", "kesinlikle katılıyorum" olarak dört dereceli Likert türündedir. Maddeler şunlardır:

1. Genellikle öyle ya da böyle idare ederim.
2. Bir şeyler başardığım için gurur duyuyorum.
3. Aynı anda birçok şeyle başa çıkabileceğimi hissediyorum.
4. Kendime olan inancım beni zor zamanlardan kurtarıyor.
5. İçine girdiğim zor bir durumdan çıkış yolumu genellikle bulabilirim.

Bu maddelerden özyeterlik indeks değeri elde edilmiştir. İndeksteki pozitif değerler, öğrencinin OECD ülkelerindeki ortalama öğrenciden daha yüksek dirençlilik bildirdiği anlamına gelmektedir.

Veriler OECD web sitesinden elde edilecektir. Araştırma sorularının yanıtlanması için betimsel ve korelasyona dayalı çözümlemelerden yararlanılacak, elde edilen bulguların yorumlanması için grafikler kullanılacaktır. Aşamalı Tepki Modeli'ne bağlı olarak hesaplanan madde parametreleri, madde karakteristik eğrisi ve madde bilgi eğrisi üzerinden yorumlanacaktır. Madde grafiklerinin tepe noktaları, maddenin en ayırt edici olduğu düzeyler olarak yorumlanacak (Symth, (n.d.)) ve bulgular diğer çözümleme sonuçlarıyla karşılaştırılacaktır.

Sonuçlar

Maddelerden ve ülkelerden bağımsız olarak öğrencilerin uç düzeyleri tercih etme durumu ile aracın tümünden aldıkları toplam puan arasındaki korelasyonun yüksek olacağı öngörülmektedir. Böylelikle özyeterlik yapısının bir göstergesine yüksek derecede sahip olan öğrencilerin, aracın tümünden

yüksek puan almaları beklenmektedir. Alanyazındaki pek çok çalışma, özyeterlik ve okuma arasında pozitif yönlü ve manidar bir ilişki olduğunu göstermektedir (Gutrie ve diğ., 2009; Niemiec ve Lachowicz-Tabaczek, 2015; Tabrizi ve Jafari, 2015). Bulgulardan hareketle bireylerin maddelerde belirtilen özyeterlik göstergelerine sahip olma olasılığının okuma başarısı arttıkça artması ancak bu değerin ülkeler özelinde farklılaşması beklenmektedir. Buna göre daha yüksek okuma performansına sahip Hong Kong'lu öğrencilerin bir maddeye ilişkin düzey tercihleri ile özyeterlik ölçeğinden aldıkları toplam puan arasındaki korelasyonun diğer ülkelerden daha yüksek olması; Türkiye ve Endonezya'daki öğrencilerin ise seçenek dağılımlarının ve toplam puanla korelasyonunun benzerlik göstermesi beklenmektedir. Alanyazındaki çalışmalara göre kaygı ile özyeterlik arasındaki ilişki negatif ve manidardır (Asayesh ve diğ., 2016; Barrows ve diğ., 2013). Benzer biçimde uç düzeyleri tercih etme durumuna bağlı olarak kaygı düzeyinin de azalacağı tahmin edilmektedir.

Kaynaklar

Asayesh, H., Hosseini, M. A., Sharififard, F., and Kharameh, Z. T. (2016). The relationship between self-efficacy and test anxiety among the paramedical students of Qom university of medical sciences. *Journal of Advances in Medical Education* (JAMED), *1*(3), 14-21.

Barrows, J., Dunn, S., and Lloyd, C. A. (2013). Anxiety, self-efficacy, and college exam grades. *Universal Journal of Educational Research*, *1*(3), 204-208.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191-215. http://doi.org/10.1016/0146-6402(78)90002-4

Clason, D. L., & Dormody, T. J. (1994). Analyzing data measured by individual Likert-type items. *Journal of Agricultural Education*, *35*(4), 4.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Cengage Learning.

Demirtaşlı, R. N., Yalçın, S., & Ayan, C. (2016). The development of IRT based attitude scale towards educational measurement course. *Journal of Measurement and Evaluation in Education and Psychology, 7*(1), 133-144. https://doi.org/10.21031/epod.43804

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Erlbaum.

Guthrie, J. T., Coddington, C. S., & Wigfield, A. (2009). Profiles of reading motivation among African American and Caucasian students. *Journal of Literacy Research*, *41*(3), 317-353. https://doi.org/10.1080/10862960903129196

Haberman, S. J., Liu, Y., & Lee, Y. H. (2019). *Distractor analysis for multiple-choice tests: An empirical study with international language assessment data*. *ETS Research Report Series*, *2019*(1), 1-16. https://doi.org/10.1002/ets2.12275

Jinks, J., & Morgan, V. (1999). Children's perceived academic self-efficacy: An inventory scale. *The Clearing House, 72*(4), 224-230. https://doi.org/10.1080/00098659909599398

Karasar, N. (2012). *Bilimsel araştırma yöntemi*. Nobel Akademik Yayıncılık.

Köse, İ. A. (2015). Aşamalı tepki modeli ve klasik test kuramı altında elde edilen test ve madde parametrelerinin karşılaştırılması. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, *15*(2), 184-197.

McGrath, R. E. (2014). Scale-and item-level factor analyses of the VIA inventory of strengths. *Assessment*, *21*(1), 4-14. https://doi.org/10.1177/1073191112450612

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, *14*(1), 59-71.

Niemiec, T., & Lachowicz-Tabaczek, K. (2015). The moderating role of specific self-efficacy in the impact of positive mood on cognitive performance. *Motivation and Emotion*, *39*(4), 498-505.

OECD (2019a). *PISA 2018 Results: Combined executive summaries volume* I, II & III. https://www.oecd.org/pisa/Combined_Executive_Summaries_PISA_2018.pdf.

OECD (2019b). *PISA 2018: Technical report.* Chapter 9: Scaling PISA Data. https://www.oecd.org/pisa/data/pisa2018technicalreport/Ch.09-Scaling-PISA-Data.pdf.

OECD (2019c). *PISA 2018 results* (Volume III): What school life means for students' lives, PISA. OECD Publishing. https://doi.org/10.1787/acd78851-en.

OECD (2019d). *PISA 2018: Technical report.* Chapter 13: Students' self-efficacy and fear of failure. https://www.oecd-ilibrary.org/sites/2f9d3124-en/index.html?itemId=/content/component/2f9d3124-en.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement*, *27*(2), 133-144.

Sullivan, G. M., & Artino Jr, A. R. (2013). Analyzing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education*, *5*(4), 541-542. https://doi.org/10.4300/JGME-5-4-18

Sytm, R. (n.d.) *Item response theory for polytomous items.* https://www.uwo.ca/fhs/tc/labs/12.PolytomousIRT.pdf

Tabrizi, A. R. N., & Jafari, M. (2015). The relationship among critical thinking, self- efficacy, and Iranian EFL learners' reading comprehension ability with different proficiency levels. *Academic Research International*, *6*(2), 412-427. http://www.savap.org.pk/journals/ARInt./Vol.6(2)/2015(6.2-40).pdf

Uyar, Ş., Öztürk-Gübeş, N. ve Kelecioğlu, H. (2013). PISA 2009 tutum anketi madde puanlarının aşamalı madde tepki modeli ile incelenmesi. *Eğitim ve Öğretim Araştırmaları Dergisi*, *4*(2), 125-134. http://www.jret.org/FileUpload/ks281142/File/15.uyar.pdf

Puanlayıcılar arası uyumun örtük sınıf analizi yöntemi ile incelenmesi

Mediha Korkmaz, Yılmaz Orhun Gürlük, Ömer Emre Can Alagöz ve Gizem Cömert

Giriş

Psikolojik özelliklerin değerlendirilmesinde bireylerin bazı özellikleri puanlayıcılar, yargıcılar tarafından ölçme araçları üzerinden puanlanmaktadır. Dezavantajlı gruplarda, eğitim araştırmalarında ve çocuklar üzerinde yapılan incelemelerde bu tür değerlendirmeler sıklıkla kullanılmaktadır Ebeveyn, akran, uzman, öğretmenler gibi değerlendiricilerin ayrı ayrı değerlendirmeleriyle bu testlerin ölçtükleri özellik üzerinden bireyler hakkında bir karar verilmektedir. Bu nedenle her puanlayıcının uyumlu bir örüntü sağlaması ve tutarlı olması, verilen kararların güvenirliğini arttırmaktadır (Basten, 2016; Raykov ve diğ., 2012).

Değerlendiriciler arasında uyumun (rater agreement) en klasik tekniklerinin başında Pearson korelasyon gelmektedir. Sonraları, Pearson korelasyon temelinde sınıf içi korelasyon katsayısı geliştirilmiştir. Sürekli olmayan veriler için sıklıkla Cohen ve Fleiss'ın kappa katsayılarının yanı sıra Krippendorf alfa istatistiği de kullanılmaktadır (Bıkmaz-Bilgen, 2011; Hayes, 2007). Sıralayıcı ve sınıflayıcı ölçekleme düzeyleri için bilinen bu yöntemler tüm puanlayıcılar, tüm puanlama kategorilerini kullanmadığında tutarlı sonuçlar vermemekte ve anlamsız çıkma eğilimi sergilemektedir. Bazı puanlayıcıların kimi kategori ya da sıraları kullanmaması var olan uyumun gözden kaçmasına sebep olabilmektedir. Böyle durumlarda uyum yüzdesinin kullanılması önerilmektedir (Hayes, 2007; Yarnold, 2016).

Son yıllarda örtük yapı (latent structure) yöntemleriyle de puanlayıcılar arası uyum incelenmektedir. Raykov (2012) tarafından, doğrulayıcı faktör analizi temelli yaklaşım tanımlamıştır. Öte yandan örtük sınıf analizi – ÖSA (latent class analysis) –puanlayıcılar arası uyum için kullanılmaktadır. ÖSA ile puanlayıcılar tarafından sınıflara atanan katılımcıların o sınıflarda bulunma olasılığına odaklanılarak, farklı puanlayıcıların ortak olarak aynı katılımcıyı aynı sınıfa atayıp atamadığı saptanır. Puanlayıcıların, katılımcılara verdikleri puanlar gözlenen değişken olarak ele alınarak puanlayıcıların sınıflanması sağlanır. Puanlayıcıların tek sınıfa atanması uyumun varlığını gösterir. Öte yandan örtük sınıf olasılıklarına bakılarak puanlayıcıların ne kadar aynı kategoriye atama yaptığı yorumlanabilir. Bu yöntemin dezavantajı çok sayıda katılımcı gerektirmesidir (Schuster, 2002). ÖSA,

sıralayıcı/sınıflayıcı değişkenlere dayandığı için sürekli puanlanan veriler için uygun değildir. Bu nedenle sürekli veriler için örtük profil analizi kullanılmaktadır (Major ve diğ., 2018).

ÖSA'nın bir diğer kullanım şekli de farklı puanlayıcıların aynı maddeleri puanladığı durumlar için belirlenen örtük sınıfların profillerinin incelenmesine dayanmaktadır. ÖSA ile bulunan her sınıfta puanlayıcılar, katılımcıları farklı bir örüntüye (patern) göre puanlamaktadır. Bu örüntü her bir madde için uyumlu ya da uyumsuz bir puanlama anlamına gelebilmekte ve buna karar vermek için sınıf profilleri incelenebilmektedir. Buna ek olarak puanlayıcıların, farklı sınıflarda (cluster) yakaladıkları uyumlar da incelenebilmektedir (Basten, 2015; De Los Reyes, 2009; Major, 2018).

Bu çalışmanın amacı örtük değişken modellemesi kapsamında (latent variable modelling) puanlayıcılar arası uyumun (interrater agreement), örtük sınıf analiziyle puanlananların sınıflarının tahminlenmesi, tahminlenen sınıfların profil incelemelerinin yapılması ve farklı sınıflardaki uyum yüzdelerinin de madde bazında hesaplanmasıdır.

Yöntem

Çalışmada Irmak ve arkadaşları (2018) tarafından geliştirilen "Mika ile Kendimi Korumayı Öğreniyorum" cinsel istismarı önleme programı çerçevesindeki bazı veriler, araştırmacıların izni alınarak kullanılmıştır. Adı geçen müdahale programının yan etkisinin olup olmadığını tespit etmek amacıyla araştırmacılar 10 maddelik sıralayıcı ölçek düzeyinde 1-5 (5=en çok, 1=en az) arasında puanlanan bir form oluşturmuşlardır. Bu form 290 çocuk için ebeveyn ve öğretmenleri tarafından değerlendirilmiştir. Böylece iki değerlendirici 10 madde üzerinden 290 katılımcıyı puanlamıştır.

İlk aşamada ebeveynlerin ve öğretmenlerin çocukları ne şekilde puanladıklarını saptamak için ÖSA uygulanmıştır. ÖSA; bireyleri, kategorik ve gözlenen değişkenlere verdikleri yanıtların örüntülerine göre örtük sınıflarda kümelemeyi hedeflemektedir (Lazarsfield ve Henry, 1968). Bu yanıtlama örüntüleri aynı zamanda bulunan örtük sınıfların profillerinin oluşmasına da katkı sağlar. Araştırmacılar, farklı sayıda sınıf içeren bir takım modeli tahminler, elde edilen uyum istatistiklerine göre de en uygun sayıda sınıfa sahip olan modeli seçerler.

Bu çalışmada, ebeveyn ve öğretmenlerin puanları her ön test maddesi için ayrı bir değişken olarak ele alınmış; toplamda 20 madde ÖSA modeline gösterge değişken olarak tanımlanmıştır. Model seçme prosedüründen sonra bireyler "modal" atama kuralına göre sınıflandırılmıştır. Ortaya çıkan her sınıf, ebeveyn ve öğretmenlerin farklı puanlama örüntülerini temsil etmekte; bu örüntü, sınıf profilleri aracılığıyla incelenebilmektedir. Ayrıca, katılımcıların sınıf üyelikleri, onların bu 10 değişkende nasıl bir örüntüyle puanlandıkları hakkında bilgi vermektedir.

Sınıflar içerisinde her bir madde için hesaplanacak bir uyum indeksi ile bu maddenin hangi sınıfta ebeveyn ve öğretmenlerce daha uyumlu puanlandığı görülmektedir. Bu amaçla, ikinci aşamada ÖSA'nın ardından her iki sınıf için madde bazında uyum yüzdesi Cohen kappa sentaksı içerisinde yer alan "puanlayıcıların uyum yüzdesi" değeri kullanılarak incelenmiştir. Temel olarak bir madde için elde edilen

uyum yüzdesi, puanlayıcılar tarafından aynı kategoride puanlanan kişi sayısının toplam kişi sayısına bölünmesi ile elde edilmektedir. Bu istatistik, bir madde için her iki sınıfta da hesaplanarak ÖSA ile elde edilen sınıflardaki uyum, klasik yöntemlerle incelemekte ve sınıflar arasında karşılaştırılabilmektedir.

Sonuçlar

ÖSA ile sınıf sayısı 1'den 5'e artan 5 model analiz edilmiş, Akaike (Nylund, 2007) bilgi kriterlerine göre 2 sınıflı model seçilmiştir. Katılımcıların 182'si birinci sınıfa; kalan 108'i ikinci sınıfa atanmıştır. Profiller incelendiğinde ebeveynlerin ve öğretmenlerin, birinci sınıftaki katılımcıları puanlarken dokuz maddede, ikincidekileri puanlarken dört maddede benzer puanlama örüntüsüne sahip olduğu bulunmuştur. Profil grafiği incelendiğinde, ikinci sınıftaki katılımcıların sistematik olarak iki puanlayıcı tarafından birinci sınıfa göre daha yüksek kategorilerde puanlandığı saptanmıştır. Diğer maddeler, iki sınıfta da hep ebeveynler tarafından daha yüksek puanlanmaktayken "duyguları hakkında konuşma" birinci sınıfta öğretmenler tarafından daha yüksek puanlanmıştır.

Sınıflara atanan katılımcıların madde düzeyinde ne derece uyumlu puanlandığı uyum yüzdesiyle incelenmiştir. "Mızmızlanma" maddesi haricinde birinci sınıftaki katılımcılarda ikincidekilere göre daha yüksek uyum saptanmıştır. Birinci sınıfta "duyguları hakkında konuşma" (%30), "itaatsizlik" (%56) ve "mızmızlanma" (%16) maddeleri haricinde uyum yüzdesi %70-87 arasında bulunmuştur. İkinci sınıfta yalnızca "cinsellik hakkında soru sorma" maddesi kabul edilebilir bir uyum yüzdesine (%71) sahipken diğer maddelerde uyum %17-53 arasında hesaplanmıştır.

Analiz sonucunda, birinci sınıfta yer alan 182 katılımcının yaklaşık 7 maddede yüksek uyumla puanlandığını ve ikinci sınıfta yer alan 108 katılımcının yalnızca 1 maddede yüksek uyumla puanlandığı bulunmuştur. Ayrıca, ÖSA sayesinde ebeveyn ve öğretmen puanlama örüntüleri hakkında daha detaylı bilgiler elde edilebilmiştir.

Kaynaklar

Basten, M.; Tienmeier H.; Althoff, R., van de Schoot, R.; Jaddoe, V. W. V., Hofman, A., Hudziak, J. J.; Verhulst, F. C., and van der Ende, J. (2015). The stability of problem behavior across the preschool years: An emprical approach in general population. *Journal of Abnormal Child Psychology, 44(2),* 393-404. https://doi.org/10.1007/s10802-015-9993-y

Bıkmaz-Bilgen, Ö. (2011). *Üst düzey zihinsel özelliklerin ölçülmesinde puanlayıcılar arası güvenirlik belirleme tekniklerinin karşılaştırılması* (Tez No. 308404). [Yüksek Lisans Tezi, Hacettepe Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

De Los Reyes, A.; Henry, D. B.; Tolan, P. H. T., and Wakschlag, L. S. (2009). Linking informant discrepancies to observed variations in young children's disruptive behavior. *Journal of Abnormal Psychology*, 37(5), 637-652. https://doi.org/10.1007/s10802-009-9307-3

Hayes, A. F., and Krippendorff, K. (2007). Answering the call for a standard reliability measuse for coding. *Communication Methods and Measures*, (1), 77-89. https://doi.org/10.1080/19312450709336664

Lazarsfeld, P. F., and Henry, N. W. (1968) *Latent structure analysis*. Houghton Mifflin.

Major, S., Seabra-Santos, M. J., and Martin, R. P. (2018). Latent profile analysis: Another approach to look at parent-teacher agreement on preschoolers' behavior problems. *European Early Childhood Education Research Journal, 26*(5), 701-717. https://doi.org/10.1080/1350293X.2018.1522743

Nylund, K. L., Asparouhov, T., and Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural equation modeling: A multidisciplinary Journal, 14*(4), 535-569. https://doi.org/10.1080/10705510701575396

Raykov, T., Dimitrov, D. M., von Eye, A., and Marcoulides, G. A. (2012). Interater agreement evaluation: A latent variable modeling approach. *Ducational and Psychological Measurement, 20*(10), 1-20. https://doi.org/10.1177/0013164412449016

Schuster, C., and Smith, D. (2002). Indexing systematic rater agreement with a latent-class model. *Psychological Method, 7*(3), 384-395. https://doi.org/10.1037/1082-989x.7.3.384

Yarnold, P. R. (2016). ODA vs. π and κ: Paradoxes of kappa. *Optimal Data Analysis, 5*, 160-161. https://odajournal.files.wordpress.com/2019/01/v5a38.pdf

Lisansüstü öğrencilerin akademik sahtekârlık eğilim düzeylerinin CHAID analizi ile incelenmesi

Esra Eminoğlu Özmercan, Betül Polat ve Zekeriya Nartgün

Anahtar kelimeler: Akademik sahtekârlık, kopya çekme, CHAID analizi, lisansüstü eğitim

Giriş

Yükseköğretimde kopya çekme, intihal gibi etik dışı akademik sahtekârlık davranışları önemli bir sorun haline gelmiştir. Baran ve Jonason (2020) bu sorunun hem öğrenciler hem de eğitim sistemi için istenmeyen durumlara yol açtığını, Blachnio ve diğerleri de (2021) insan yaşamı, toplumsal değerler ve ekonomi üzerinde ciddi sonuçları olduğunu ifade etmektedir. Akademik sahtekârlık, yükseköğretim öğrencileri arasında yaygın olarak gözlenmekte ve öğrencileri günlük yaşamlarında da sıklıkla dürüst olmayan uygulamalara yöneltmektedir (Baran ve Jonason, 2020). Yükseköğretim üzerinde kötü etkileri olabileceğinden üniversite öğrencileri, öğretmenler ve yöneticiler için bu endişe verici bir durumdur (Whitley ve Keith-Spiegel, 2002).

Akademik sahtekârlık, öğrencilerin başkalarının akademik çalışmalarını kendilerininmiş gibi göstermeleri olarak tanımlanmakta sınav kâğıtlarını değiştirme, sınavlarda kopya çekme, diğer öğrencilerin ödevlerini kopyalama, ödevlerde değişiklik yapma, intihal gibi davranışları içermektedir (Aluede ve diğ., 2006; Jensen ve diğ., 2002). Pavela (1978)'ya göre akademik sahtekârlık dört bileşenden (kopya çekme, uydurma, intihal, akademik sahtekârlığı kolaylaştırma) oluşmaktadır. Bunlardan ilki herhangi bir akademik uygulamada kasıtlı olarak izinsiz bulgu, bilgi veya yardımcı kaynakları kullanmak veya kullanmaya teşebbüs etme olan kopya çekme, ikincisi herhangi bir akademik uygulamada herhangi bir bilgi veya alıntının kasıtlı ve izinsiz tahrif edilmesi veya icadı olan uydurma, üçüncüsü başka bir kişinin fikir, sözcük veya ifadelerini herhangi bir onay olmaksızın, kendisininmiş gibi kasıtlı olarak kullanması veya taklit etmesi olan intihal, dördüncüsü ise bir başkasına bilerek veya isteyerek yardım etmek, yardım etmeye çalışmak, bir tür akademik sahtekârlığa bulaşmak olan akademik sahtekârlığı kolaylaştırmadır. Whitley ve Keith-Spiegel (2002), Pavela (1978)'in bileşenlerine yanlış beyan (bir öğretim görevlisine akademik bir uygulama hakkında yanlış bilgi vermek), ortak bir projeye katkıda bulunmamak ve başkalarının işlerini tamamlamasını engelleyen eylemlerden oluşan sabotajı da eklemiştir.

Yükseköğretim kurumlarının amacı sadece öğrenme çıktılarının gerçekleştirilmesi için eğitim vermek, istendik davranışların gerçekleştirilmesi için çalışmalar yapmak, mesleki anlamda iyi yetişmiş mezunları eğitmek değil aynı zamanda ahlaki ve mesleki anlamda iyi yetişmiş, etik kurallara uyan, adil, sorumluk sahibi, çok yönlü, bireyler yetiştirmektir.

Yükseköğretim kurumlarında öğrencilerin akademik sahtekârlık eğilim düzeylerinin ve bu düzeyleri etkileyen değişkenlerin belirlenebilmesi, elde edilen sonuçlara göre akademik sahtekârlık nedenlerinin öğrenilmesi ve bunlara yönelik önlemlerin alınması, önerilerin sunulmasının eğitim sistemimiz için gerekli olduğu düşünülmektedir. Bu bağlamda bu araştırmanın ilk amacı, üniversite öğrencilerine yönelik Eminoğlu (2008) tarafından geliştirilen Akademik Sahtekârlık Eğilim Ölçeği'nin lisansüstü öğrencilere yönelik uyarlanması, ikinci amacı ise lisansüstü öğrenimine devam eden öğrencilerin akademik sahtekârlık eğilim düzeylerinin belirlenmesi ve etkili olan faktörlerin CHAID Analizi yöntemi ile incelenmesidir.

Yöntem

Araştırmada lisansüstü öğrencilerin akademik sahtekârlık eğilim düzeyleri ile ilişkili olan değişkenlerin belirlenmesi amaçlandığından ilişkisel tarama modelinden yararlanılmıştır. Akademik Sahtekârlık Eğilim Ölçeği'nin lisansüstü öğrencilerine uyarlanabilmesi için 2020-2021 eğitim-öğretim yılında farklı üniversitelerin farklı enstitülerde öğrenim gören 426, CHAID analizi yapılabilmesi için de farklı üniversitelerin Eğitim Bilimleri Enstitülerinde öğrenim gören 336 öğrenci çalışma grubunu oluşturmuştur. Araştırmacılar tarafından hazırlanan kişisel bilgi formu ve Eminoğlu (2008) tarafından lisans öğrencileri için geliştirilen dört alt boyuttan ve 22 maddeden oluşan Akademik Sahtekârlık Eğilimi Ölçeği kullanılmıştır.

Akademik Sahtekârlık Eğilimi Ölçeği'nin Lisansüstü Öğrencilerine Uyarlanması

Lisans öğrencileri için geliştirilen ölçeğin lisansüstü öğrencilerine uygulanabilmesi için geçerlik ve güvenirlik çalışması yapılmıştır. Bu amaçla farklı enstitülerde öğrenim gören 426 lisansüstü öğrencisinden veri toplanmıştır. Ölçeğin yapı geçerliğini sınamak için Doğrulayıcı Faktör Analizi'nden (DFA) yararlanılmıştır. Güvenirliğini sınamak için ise Cronbach alfa iç tutarlılık katsayısı hesaplanmıştır.

Yapı Geçerliği

DFA, ölçeğin dört boyutlu yapısının lisansüstü öğrencilerinden elde edilen 426 veride doğrulanıp doğrulanmadığını belirlemek amacıyla yapılmıştır. DFA sonucunda χ^2/sd oranı 2.66, AGFI .84, GFI .87, RMSEA .072, NFI .94, NNFI, CFI ve IFI ise .96 bulunmuştur. Uyum iyiliği değerleri incelendiğinde, tüm değerlerin model veri uyumu için kabul edilebilir sınırlarda olduğunu ve ölçeğin yapı geçerliğinin sağlandığını ortaya koymaktadır. Cronbach alfa güvenirlik katsayısı alt boyutlar için sırasıyla .728, .734, .673, .720 ve ölçeğin tamamı için ise .876 bulunmuştur.

Öğrencilerin akademik sahtekârlık eğilimlerini açıklayan değişkenler, CHAID (Chi-squared Automatic Interaction Detection) analizi ile belirlenmiştir. CHAID analizi, bir yordanan değişken ile

birden fazla yordayıcı değişken arasındaki ilişkilerin belirlenmesinde kullanılmaktadır (Diepen ve Franses, 2005; Doğan ve Özdamar, 2003). Verilerin analizinde SPSS 21 programı kullanılmış ve anlamlılık düzeyi .05 alınmıştır.

Sonuçlar

Lisansüstü öğrenimine devam eden öğrencilerin Kopya Çekme Eğilimi boyutu üzerinde etkili olan tek önemli değişkenin lisansüstü eğitim düzeyi ($F_{(1,334)}$=11.987, p< .05) olduğu belirlenmiştir. CHAID analizi, tezli ve tezsiz yüksek lisans öğrencilerinin kopya çekme eğilimlerinin doktora öğrencilerinden daha yüksek olduğunu ve her iki grubun da kopya çekme eğilimlerinin çok düşük düzeyde olduğunu göstermektedir.

Ödev-Proje Gibi Çalışmalarda Sahtekârlık Eğilimi boyutu üzerinde etkili olan en önemli değişkenin lisansüstü eğitim düzeyi ($F_{(1,334)}$= 14.246, p< .05), tezli ve tezsiz yüksek lisans öğrencilerin puanlarını açıklayan değişkenin başarı düzeyi olduğu belirlenmiştir ($F_{(1,231)}$=10.100, p< .05). CHAID analizi tezli ve tezsiz yüksek lisans öğrencilerinin ödev-proje gibi çalışmalarda sahtekârlık eğilimlerinin doktora öğrencilerinden daha yüksek olduğunu ve her iki grubun da sahtekârlık eğilim düzeylerinin çok düşük düzeyde olduğunu, orta ve çok düzeyde başarılı olan yüksek lisans öğrencilerinin ödev-proje gibi çalışmalarda sahtekârlık eğilimlerinin başarılı ve başarısız düzeyde olan öğrencilerden daha yüksek olduğunu göstermektedir.

Araştırma Yapma ve Raporlaştırma Sürecinde Sahtekârlık Eğilimi boyutu üzerinde etkili olan tek önemli değişkenin eğitim alma nedeni olduğu belirlenmiştir ($F_{(1,334)}$= 13.623, p< .05). CHAID analizi eğitim alma nedeni alanda yetkin olma ile mesleki alanda ilerleme olan öğrencilerin Araştırma Yapma ve Raporlaştırma sürecinde sahtekârlık eğilimlerinin akademisyen olma olan öğrencilerinden daha yüksek olduğunu ve her iki grubunda çok düşük düzeyde olduğunu göstermektedir.

Atıflara Yönelik Sahtekârlık Eğilimi boyutu üzerinde etkili olan en önemli değişkenin lisansüstü eğitim düzeyi olduğu ($F_{(1,334)}$= 12,234, p< .05), tezli ve tezsiz yüksek lisans öğrencilerin puanlarını açıklayan değişkenin cinsiyet değişkeni olduğu belirlenmiştir ($F_{(1,231)}$=4.509, p< .05). CHAID analizi tezli ve tezsiz yüksek lisans öğrencilerinin atıflara yönelik sahtekârlık eğilimlerinin doktora öğrencilerinden daha yüksek ve çok düşük düzeyde olduğunu, erkek öğrencilerin atıflara yönelik sahtekârlık eğilimlerinin kadın öğrencilerden daha yüksek ve çok düşük düzeyde olduğunu göstermektedir.

Akademik Sahtekârlık Eğilimi puanlarını açıklayan faktörler incelendiğinde, öğrencilerin akademik sahtekârlık eğilimi üzerinde etkili olan en önemli değişkenin lisansüstü eğitim düzeyi ($F_{(1,334)}$=18.243, p< .05), tezli ve tezsiz yüksek lisans öğrencilerin puanlarını açıklayan değişkenin başarı düzeyi olduğu belirlenmiştir ($F_{(1,231)}$=8.555, p< .05). CHAID analizi tezli ve tezsiz yüksek lisans öğrencilerinin akademik sahtekârlık eğilimlerinin doktora öğrencilerinden daha yüksek olduğunu ve her iki grubunda sahtekârlık eğilim düzeylerinin çok düşük düzeyde olduğunu göstermektedir. Orta düzeyde başarılı olan yüksek lisans öğrencilerinin akademik sahtekârlık eğilimlerinin başarısız-başarılı-çok başarılı düzeyde olan öğrencilerden daha yüksek olduğunu ve çok düşük düzeyde olduğunu göstermektedir.

Kaynaklar

Aluede, O., Omoregie, E. O., and Osa-Edoh, G. I. (2006). Academic dishonesty as a contemporary problem in higher education: How academic advisers can help. *Reading Improvement, 43*(2), 97-106. https://link.gale.com/apps/doc/A148856041/AONE?u=anon~ff5e0ce6&sid=googleScholar&xid=9fc31165

Baran L., and Jonason, P. K. (2020). Academic dishonesty among university students: The roles of the psychopathy, motivation, and self-efficacy. *PLoS ONE* 15(8), e0238141. https://doi.org/10.1371/journal.pone.0238141

Błachnio, A., Cudo, A., Kot, P., Torój, M., Asante, K. O., Enea, V., Ben-Ezra, M., Caci, B., Dominguez-Lara, S. A., Kugbey, N., Malik, S., Servidio, R., Tipandjan, A., and Wright, M. F. (2022). Cultural and psychological variables predicting academic dishonesty: A cross-sectional study in nine countries. *Ethics & Behavior, 32*(1), 44-89. https://doi.org/10.1080/10508422.2021.1910826

Diepen, V. M., and Franses, H. F. (2006). Evaluating chi-squared automatic interaction detection. *Information Systems, 31*(8), 814-831. https://doi.org/10.1016/j.is.2005.03.002

Doğan, V. ve Özdamar K. (2003). CHAID analizi ve aile planlaması ile ilgili bir uygulama. *Türkiye Klinikleri Tıp Bilimleri Dergisi, 23*(5), 392-397.

Eminoğlu, E. (2008). *Üniversite öğrencilerinin akademik sahtekarlık eğilimlerinin ölçülmesine yönelik bir ölçek geliştirme çalışması* (Tez No. 221561) [Yüksek lisans tezi, Abant İzzet Baysal Üniversitesi]. Yükseköğretim Kurumu Tez Merkezi.

Jensen, L. A., Arnett, J. J., Feldman, S. S., & Cauffman, E. (2002). It's wrong, but everybody does it: Academic dishonesty among high school and college students. *Contemporary Educational Psychology,* 27(2), 209–228. https://doi.org/10.1006/ceps.2001.1088

Pavela, G. (1978). Judicial review of academic decision-making after Horowitz. *School Law Journal, 55*(8), 55-75.

Whitley, B. E., and Keith-Spiegel, P. (2002*). Academic dishonesty: An educator's guide.* Lawrence Erlbaum Associates, Inc.

Likert tipi ölçeklerde kullanılan farklı tutum ifadelerinin geçerlik ve güvenirlik üzerindeki etkisi

Nuri Doğan, Ceylan Gündeğer ve Meltem Yurtçu

Anahtar kelimeler: Likert tipi ölçek, likert, ölçek, tutum ifadesi, geçerlik

Giriş

Tutum, ilgi gibi psikolojik yapıların ölçülmesinde sıklıkla Likert tipi ölçeklerden yararlanılmaktadır. Likert tipi ölçeklerde, yanıtlayıcı, ölçekteki her maddenin anlamına ilişkin tutumunun derecesini belirtir (Tezbaşaran,1997; Sekreter ve Akyüz, 2003; akt. Başar, 2021). Likert tipi ölçeklerde dereceleme beşli olabildiği gibi, üçlü veya yedili de olabilmekte ve buna bağlı olarak güvenirlik değişebilmektedir. Türkiye'de yayınlanan araştırma yöntemi kitaplarında Likert tipi ölçekler en sık beş seçenekli olarak ele alınmaktadır (İslamoğlu ve Alnıaçık, 2016, Nakip, 2006, Gegez, 2010, Kurtuluş, 2006; akt. Dursun ve Alnıaçık, 2019).

Araştırmalarda, beşli Likert tipi ölçek seçenekleri, "*Kesinlikle katılmıyorum, Katılmıyorum, Kararsızım, Katılıyorum ve Kesinlikle katılıyorum"* şeklinde sıralanır. Çalışmalarda çoğunlukla, ölçek orta noktası olarak, orijinal şekli *"undecided"* olan ifadenin Türkçe çevirisine karşılık gelen *"kararsızım"* ifadesinin kullanıldığı belirtilmektedir (Kağıtçıbaşı, 2010). Bazı pazarlama araştırmalarında orta seçenek "*ne katılıyorum ne katılmıyorum"* olarak (Gegez 2010; Kurtuluş, 2006; Yükselen, 2003), bazı çalışmalarda ise *"fikrim yok"* ifadesi ile etiketlenmiştir (Nakip, 2006). Bazen de ölçeklerde, bu ifadelerle birlikte rakamlar, yalnızca rakamlar ya da ifadeler olmaksızın emojiler de (Kılıç ve diğ., 2021) kullanılabilmektedir.

Farklı tutum ifadelerinin gösterdikleri derecelerin birbirine benzer olup olmadığı; anlambilim (semantik) açısından aralarında farklılık olup olmadığı; bireyler tarafından bu ifadelerin eş-anlamlı olarak anlaşılıp anlaşılmayacağı hakkında bir görüş birliği bulunmamaktadır. Alanyazında bu düzeylerden özellikle orta değerin ne bildirdiğine ilişkin çeşitli tartışmalar gündeme gelmiştir. Başar'a (2021) göre, *"kararsızım, fikrim yok, bir şey söyleyemem"* gibi ifadeler orta derece veya eğilim değil, durum bildirmektedir; dolayısıyla bu tür ifadelerin ölçek uygulamalarında kullanılmaları hatalı olabilmektedir. Başar (2021) bu ifadelerin yerine *"yansızım", "nötrüm" "orta düzeyde katılıyorum"* gibi ifadeye katılma ve katılmama bildirebilecek bir etiket kullanılması gerektiğini belirtmektedir. Bora Semiz ve Altunışık (2016), *"kararsızım", "ne katılıyorum ne katılmıyorum"* etiketlerinin kullanılabileceğini ancak, "fikrim

yok" ifadesinin ölçek orta noktası için tutumda bir yer belirtmediğinden orta nokta için uygun olmayacağını belirtmiştir. Ayrıca başka bir çalışma göstermiştir ki, *"nötr"* anlamında kullanılan orta değer seçeneğine cevaplayıcılar, merkeze yönelme yanlılığı ve sosyal beğenilme arzusu yanlılığı nedeniyle de yönelebilmektedir (Nadler, 2015).

Likert tipi ölçeklerde, yukarıda çeşitlenen tutum ifadelerinden hangisinin kullanılacağına araştırmacılar ölçek geliştirme sürecinde kendileri karar vermektedir. Likert tipi ölçeklerde kullanılacak cevap seçeneğinin sayısı; seçenek etiketlerinin uygulama diline çeviri kolaylığı, seçeneklerin ne ifade ettiğinin net şekilde anlaşılması, seçeneklerin anlam açısından örtüşmemesi, cevaplayıcının vermek istediği cevabın tam karşılığının bulunabilmesi gibi bazı teorik kısıtlar ile geçerlilik ve güvenilirlik gibi psikometrik özelliklerle ilişkilidir (Krosnick ve Presser: 2010; akt. Dursun ve Alnıaçık, 2019).

Bu çalışmada, aynı öğrenci grubundan elde edilen veriyle, aynı maddelerden oluşan fakat farklı tutum ifadelerine sahip ölçek formlarının geçerlik ve güvenirlik kanıtlarını incelemek ve bu kanıtları formlar üzerinden karşılaştırmak amaçlanmıştır. Ölçeklerdeki tutum ifadelerinin farklı kullanımlarına ilişkin ampirik bir kanıt sunması bakımından bu araştırmanın önemli ve gerekli olabileceği düşünülmüştür.

Yöntem

Bu araştırma, aynı maddelerden oluşan ve farklı tutum ifadelerine sahip ölçek formlarının geçerlik ve güvenirliklerinin incelenmesi bakımından betimsel araştırma niteliğindedir. Araştırmanın örneklemini, 2020-2021 eğitim öğretim yılında, Aksaray Üniversitesi, Hacettepe Üniversitesi ve İnönü Üniversitesi'nde öğrenim görmekte olan 377 birey oluşturmaktadır.

Araştırmada veri toplama aracı olarak Matematikle İlgili Düşünceler Ölçeği (MİDÖ) kullanılmıştır. MİDÖ, Baykul (1990) tarafından geliştirilmiş ve 30 maddeden oluşan tek boyutlu bir ölçektir. MİDÖ'nün tek boyutla açıklanan varyans oranı %56 ve güvenirlik katsayısı da 0.96'dır. Ölçekten alınabilecek en düşük puan 30 ve en yüksek puan 150'dir (Nartgün, 2002). Bu araştırma kapsamında, 30 maddelik bu ölçeğin dört ayrı formunun öğrencilere uygulanmasının zaman alacağı düşünülerek ve ölçekteki bazı maddelerin yeterli bulunamaması sebebiyle ölçek kısaltılmıştır. Kısaltılan ölçeğe, Matematiğe yönelik düşünceleri ölçtüğü düşünülen daha ilgi çekici iki madde eklenmiştir. Eklenen maddeler, *"Bir durumu matematiksel olarak ifade etmek beni mutlu eder."* ve *"Matematiksel keşifler beni büyüler."* şeklindedir. Buna göre uygulanan ölçekte yer alan madde sayısı 14'tür. Ölçeğin yalnızca tutum ifadeleri farklılaştırılarak dört ayrı formu oluşturulmuştur. Form-1'de *"Kesinlikle katılmıyorum, Katılmıyorum, Fikrim yok, Katılıyorum ve Kesinlikle katılıyorum";* Form-2'de *"Fikrim yok" yerine "Kararsızım";* Form-3'te *orta derece için "Ne Katılıyorum Ne Katılmıyorum"* ve Form-4'te ise sadece *Kesinlikle katılmıyorum (1) (2) (3) (4) (5) Kesinlikle Katılıyorum* ifadelerine yer verilmiştir.

Veri, çevrimiçi platformda ve bir hafta arayla ölçek uygulamalarını içerecek şekilde toplanmıştır. Veri analizinde, polikorik korelasyon matrisine dayalı olarak, Factor 11.05.01 (Lorenzo-Seva ve Ferrando, 2021) programında Açımlayıcı Faktör Analizi ve R yazılımında (R Core Team, 2013)

Doğrulayıcı Faktör Analizi, varsayımlar test edilerek uygulanmış; analiz sonuçları formlar bazında karşılaştırılmıştır. Formlardan hesaplanan madde faktör yüklerinin formlar arasında farklılık gösterip göstermediği parametrik olmayan yöntemler ile test edilmiştir. Madde faktör yükleri arasındaki ilişki için Spearman Korelasyon katsayısından yararlanılmıştır. Ayrıca güvenirlik için, ölçeğin tek boyutlu bir yapı göstermesi sebebiyle Cronbach Alfa iç tutarlılık güvenirlik katsayısı hesaplanarak formların geçerlik ve güvenirlik kanıtlarının karşılaştırılması sağlanması amaçlanmıştır.

Sonuçlar

Formların dördünün de KMO değerleri 0.90 değerinden yüksektir. Bartlett Küresellik Testi sonuçları ise tüm formlarda 0.01 hata düzeyinde manidar çıkmıştır. Form-1-2-3-4 tek boyutlu bir yapıya işaret etmekte; bu tek boyutla toplam varyansın sırasıyla %76.9; %78.7; %79.7 ve %76.8'ini açıklamaktadır. Formların tümünün iç tutarlılık anlamındaki güvenirliği 0.98 olarak hesaplanmıştır. Madde faktör yükleri Form-1 ve Form-3 arasında *(z* = -2.84; *p* = .005*)* ve Form-3 ve Form-4 arasında (*z* = -2.51; *p* = .012*)* manidar düzeyde farklılaşmaktadır. Cohen'in kriterlerine göre (1988) bu farklılıkların küçük etki büyüklüğü gösterdiği bulunmuştur (*Form-1-3 için r* = 0.10*; Form-3-4 için r* = .09*).* Formlar arası en düşük korelasyonların Form-4 ile diğer formlar arasında olduğu görülmüştür. DFA sonuçlarına göre formların tümünün model veri uyumu iyi düzeydedir. Formlar arası karşılaştırmalar fark (Δ) değerleri ile yürütülmüştür. ΔCFI değerlerinden hiçbiri ±0.01'den büyük değilken; Δ RMSEA değerlerinden yalnızca bir karşılaştırmada bu sınırın aşıldığı dikkat çekmektedir. Orta değerde "*Fikrim yok*" ifadesinin yer aldığı Form 1 ve "*Kararsızım*" ifadesinin yer aldığı Form 2'nin RMSEA değerleri arasındaki fark -0.013 olarak hesaplanmıştır. Buna göre Form 2'nin veriye daha iyi uyum sağladığı görülmüştür. Bu araştırmanın bulguları ışığında araştırmacılara, bu tür araştırmaların tekrarlanması, farklı tutum ifadelerinden oluşan ölçeklerin değişmezlik açısından incelenmesi önerilebilir.

Kaynaklar

Başar. H. (10.7.2021). Araştırmalarda likert yanılgıları. http://docplayer.biz.tr/52855457-adresinden-10.7.2021-tarihinde-erisildi-arastirmalarda-likert-yanilgilari.html

Baykul. Y. (1990). İlkok*ul 5. sınıftan lise ve dengi okulların son sınıflarına kadar matematik ve fen derslerine karşı tutumda görülen değişmeler ve ÖSS sınavındaki başarı ile ilişkili olduğu düşünülen bazı faktörler.* ÖSYM Yayınları.

Bora Semiz. B. ve Altunışık R. (2016). Pazarlama araştırmalarında likert tipi ölçeklerin özelliklerinin cevaplama tarzları üzerindeki etkilerinin incelenmesi. *Bartın Üniversitesi İİBF Dergisi*, *7*(14), 577-598. https://kutuphane.dogus.edu.tr/mvt/pdf.php

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillside, NJ: Lawrence Erlbaum Associates.

Dursun. İ. ve Alnıaçık. Ü. (2019). Likert ölçeklerinde seçenek etiketleme kararları: kullanılan etiketler ölçüm sonuçlarını etkiler mi? *Adıyaman Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 12*(33), 148-196. https://doi.org/10.14520/adyusbd.549447

Gegez. E. (2010). *Pazarlama araştırmaları* (3. baskı). Beta Yayınları.

İslamoğlu. A. H. ve Alnıaçık. Ü. (2016). *Sosyal bilimlerde araştırma yöntemleri* (5. baskı). Beta Yayınları.

Kağıtçıbaşı. Ç. (2010). *Günümüzde insan ve insanlar* (14. baskı). İstanbul: Evrim Yayınevi.

Kılıç. A. F.. Uysal. İ. ve Kalkan. B. (2021). An alternative to likert scale: Emoji. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 182-191. https://doi.org/10.21031/epod.864336

Krosnick. J. A., and Presser. S. (2010). Questionnaire design. In J. D. Wright, and P. V. Marsden (Eds.). *Handbook of Survey Research* (2nd ed., pp. 263-314) Emerald Group Publishing.

Kurtuluş. K. (2006). *Pazarlama araştırmaları* (8. bas). Literatür Yayıncılık.

Lorenzo-Seva, U., and Ferrando, P. J. (2021). *Factor* (version 11.05.01) [Computer software]. Universitat Rovirai Virgili.

Nadler. J. T.. Weston. R., and Voyles. E. C. (2015). Stuck in the middle: The use and interpretation of mid-points in items on questionnaires. *The Journal of General Psychology, 142*(2), 71-89. https://doi.org/10.1080/00221309.2014.994590

Nakip, M. (2006). *Pazarlama araştırmaları: Teknikler ve (spss destekli) uygulamalar.* Seçkin Yayıncılık.

Nartgün, Z. (2002). *Aynı tutumu ölçmeye yönelik likert tipi ölçek ile metrik ölçeğin madde ve ölçek özelliklerinin klasik test kuramı ve örtük özellikler kuramına göre incelenmesi* (Tez No. 113510) [Doktora tezi, Hacettepe Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

R Core Team. (2013). *R: A language and environment for statistical computing* (version 3.0.1) [Computer Software]. http://www.R-project.org/

Sekreter. S. ve Akyüz. G. (2003). Pazarlama araştırmalarında kullanılan ölçeklere ilişkin bir yayın taraması (1995-2002). *Akdeniz Üniversitesi İİBF Dergisi, 6*, 123-150.

Tezbaşaran. A. (1997). *Likert tipi ölçek geliştirme kılavuzu* (2. baskı). Türk Psikologlar Derneği.

Yükselen. C. (2003). *Pazarlama araştırmaları* (2. baskı). Detay Yayıncılık.

İki değişen çeldirici fonksiyonu belirleme yönteminin karşılaştırması

Osman Tat

Giriş

Test geliştirme aşamasında temel amaç, mümkün olan en az madde ile hedeflenen amaç doğrultusunda geçerli ve güvenilir ölçümler yapabilen bir test elde etmektir. Bu durum tipik olarak geniş bir soru havuzundan testin geçerlik ve güvenirliğine en fazla katkı sağlayan maddelerin seçilerek teste dahil edilmesi ile sağlanmaya çalışılır. Testi oluşturan son madde setinin belirlenmesi aşamasında uygulanan istatistiksel yöntemlere madde analizi denir (Crocker ve Algina, 2009). Madde analizinin test geçerliğinin sağlanmasında önemli bir süreç olduğu bilinmektedir. Öncelikle madde analizi ile çalışmanın boyutluluğu hakkında kanıtlar toplanabilir. Ayrıca madde güçlük ve ayırıcılık indekslerinin belirlenmesi yoluyla hangi maddelerin testin nihai formunda yer alacağına karar verilir. Yüksek düzeyde ayırıcılığa sahip maddeler testin güvenirliğini yükseltir. Bu da geçerlik için bir önkoşuldur. Ayrıca madde analizinden elde edilen parametreler, test eşitlemede kullanılabilir ve son olarak çeldirici analizi ile uygun bir biçimde çalışmayan çeldiricilerin düzeltilmesi veya yerlerine daha uygun çeldiricilerin hazırlanması sağlanabilir (Lane ve diğ., 2015). Bu anlamda madde analizinde madde güçlük, madde ayırıcılık ve çeldirici analizlerinin yanında değişen madde fonksiyonu (DMF) ve değişen çeldirici fonksiyonu (DÇF) gibi bu parametrelerin testi alan alt gruplarda farklı davranıp davranmadığına ilişkin analizler de yürütülür.

DMF ve DÇF birbirinin alternatifi analizler olmamakla beraber yakından ilişkilidir ve DMF analizlerinin DÇF ile birlikte yürütülmesinin birçok avantajı vardır. Örneğin, DÇF gösteren maddelerin DMF gösterme olasılığın daha yüksek olması beklenen bir durumdur. Çünkü testi alan bir alt grubun bir çeldirici tarafından daha fazla çekilmesi, doğru cevabın daha az seçilmesi sonucunu doğurabilir. Yani testi alan alt gruplarda aynı yetenek düzeyindeki öğrencilerden biri için bu madde zor iken diğeri için aynı zorlukta olmamasına sebep olabilmektedir. Bu durumda maddelerin DMF gösterme nedenleri incelenirken DÇF'nin göz ardı edilmemesi gerekir (Kato ve diğ., 2009; Middleton ve Laitusis, 2007; Penfield, 2008).

Yöntem

Bu çalışma kapsamında, DMF gösteren madde oranı ve örneklem büyüklüğüne göre iki farlı DÇF analiz yönteminin sonuçları karşılaştırılacaktır. Çalışmanın amacı, Penfield (2008) tarafından geliştirilen

odds oranı yaklaşımı ve Hladká ve Martinková (2020) tarafından geliştirilen multi-nominal log-linear regresyon'a dayalı DÇF analizlerinin DMF'li maddelerdeki performansını gözlemlemektir. Bu amaç doğrultusunda DMF gösteren madde oranının %10, %20 ve örneklem büyüklüğünün 100, 500 ve 1000 olduğu simülasyon koşullarında bu iki yöntemin DÇF gösteren maddeleri ne düzeyde belirleyebildiği ortaya konmaya çalışılacaktır. Çalışmada DMF gösteren maddeler Mantel Haenszel yöntemi kullanılarak belirlenecektir.

Sonuçlar

Çalışmanın sonucunda her iki DÇF analiz yönteminin değişen çeldirici fonksiyonu gösteren maddeleri tespiti noktasında benzer sonuçlar vermesi beklenmektedir.

Kaynaklar

Crocker, L., and Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

Hladká, A., and Martinková, P. (2020). difNLR: Generalized logistic regression models for DIF and DDF detection. *The R Journal*, *12*(1), 300-323. https://doi.org/10.32614/RJ-2020-014

Kato, K., Moen, R. E., and Thurlow, M. L. (2009). Differentials of a state reading assessment: Item functioning, distractor functioning, and omission frequency for disability categories. *Educational Measurement: Issues and Practice*, *28*(2), 28–40. https://doi.org/10.1111/j.1745-3992.2009.00145.x

Lane, S., Raymond, M. R., and Haladyna, T. M. (2015). *Handbook of test development* (2nd ed.). Routledge.

Middleton, K., and Laitusis, C. C. (2007). Examining test items for differential distractor functioning among students with learning disabilities. *ETS Research Report Series*, *2007*(2), i-34. https://doi.org/10.1002/j.2333-8504.2007.tb02085.x

Penfield, R. D. (2008). An odds ratio approach for assessing differential distractor functioning effects under the nominal response model. *Journal of Educational Measurement*, *45*(3), 247–269. https://doi.org/10.1111/j.1745-3984.2008.00063.x

Eğitimde gelişmişlik indeksinin bulanık c-ortalama kümeleme algoritması kullanılarak geliştirilmesi ve PISA 2018 okuma becerileri başarısıyla ilişkilendirilmesi

Özge Altıntaş, Furkan Başer ve Ömer Kutlu

Anahtar Kelimeler: Eğitimde gelişmişlik indeksi, PISA, okuma becerisi, bulanık kümeleme, bulanık c-ortalama kümeleme algoritması

Giriş

Ekonomik açıdan gelişmiş ülkeler, bilimsel bilginin kaynağı olarak eğitimli bireylerin rol oynadığı bilgi ekonomisinin temellerini atmışlardır. Günümüzde bir ülkenin eğitimde gelişmişlik düzeyi, diğer ülkelerle rekabet gücünde belirleyici bir etken haline gelmiş, ülkenin refahı için önemli bir öge olarak kabul edilmiştir. Bu nedenle eğitim, bir ülkenin sosyoekonomik açıdan dünyadaki yerini oluşturmada anahtar bir rol üstlenmiştir. Avrupa Birliği (European Union [EU]), Dünya Bankası (World Bank [WB]), Ekonomik Kalkınma ve İşbirliği Örgütü (Organisation for Economic Co-operation and Development [OECD]), Birleşmiş Milletler Eğitim, Bilim ve Kültür Örgütü (United Nations Educational, Scientific and Cultural Organization [UNESCO]) gibi kuruluşlar, eğitimin dünya ölçeğinde ön plana çıkan önemine bağlı olarak eğitimle ilgili bilgi verici birçok çalışmayı gerçekleştirmiştir. Bu çalışmalardan biri de Birleşmiş Milletler Kalkınma Programı (The United Nations Development Program [UNDP]) tarafından 1990 yılından itibaren yayınlanan İnsani Gelişmişlik Raporu'dur (Human Development Report [HDR]). Bu raporlarda, insanları kalkınma sürecinin merkezine yerleştiren insani gelişme yaklaşımlarının farklı temaları araştırılmıştır.

İnsani gelişme, gayri safi milli hasıladaki büyümeyi gösteren gelir, üretim ve sermaye birikiminden fazlasıdır. Örneğin, uzun ve sağlıklı bir yaşam sürmek, eğitimli olmak, insana yakışır bir yaşam standardı için gerekli kaynaklara erişmek, siyasi özgürlük, insan hakları ve kişisel öz saygı bunlar arasında sayılabilir (UNDP, 1990). Dünya ülkeleri açısından insani gelişmişliğin göstergesi olduğu düşünülen İnsani Gelişmişlik İndeksi (Human Development Index [HDI]) bu değişkenlere bağlı olarak geliştirilmektedir. HDI, *Gayri Safi Yurt İçi Hasıla İndeksi* (Gross Domestic Product Index [GDP]), *Yaşam Beklentisi İndeksi* (Life Expectancy Index [LEI]) ve *Eğitim İndeksi* (Education Index [EI]) olmak üzere üç bileşenden oluşmaktadır. UNDP, eğitim indeksini 2010 yılından itibaren yalnızca "ortalama eğitim süresi" ve "beklenen eğitim süresi" göstergelerine dayalı olarak hesaplamaktadır (UNDP, 2010, 2011). Eğitime

ilişkin gelişmişlik indeksi, ülkelerin birbirlerine göre eğitimdeki gelişmişlik düzeyleri hakkında bilgi vermektedir. Bu indeksin hesaplanmasında kullanılan göstergelerin yanında, farklı göstergelerin de indeks kapsamına eklenmesi daha nesnel, yansız ve etkili seçeneklerin tasarlanması önemlidir (Villar, 2013).

Bu çalışmada yeni bir eğitimde gelişmişlik indeksi hesaplanmış ve iki farklı amaç için kullanılmıştır. Bu amaçlardan ilki, ülkeleri eğitimdeki gelişmişlik düzeylerine göre sıralamaktır. Diğeri ise, PISA 2018 uygulamasına katılmış ülkelerin eğitimde gelişmişlik indeksine göre elde edilen sırası ile bu uygulamanın öncelikli alanını oluşturan okuma becerileri başarı sırasını karşılaştırmaktır. Okuma becerisi, okul öğrenmelerinde merkezi bir rol oynaması, temel bir yaşam becerisi olması, bireyin toplumsallaşmasına katkıda bulunması ve eğitimin tüm alanlarına katılımı kolaylaştırması açısından eğitimde gelişmişliğin önemli bir göstergesidir. Araştırmalar okuma becerisinin, bireyin okulda ya da okuldan sonra eğitimde geçirdiği süreye göre, ekonomik ve toplumsal refahın daha güvenilir bir yordayıcısı olduğunu göstermektedir (OECD, 2010). Bu çalışmada, daha fazla göstergeden yararlanılarak yeni bir eğitimde gelişmişlik indeksi oluşturulmuş ve ülkelerin eğitimdeki başarısı bu indeksle ilişkilendirilerek karşılaştırılmıştır.

Yöntem

Eğitimde gelişmişlik indeksinin hesaplanmasında, UNDP ile Dünya Bankası'nın veri tabanlarında yer alan bazı göstergeler kullanılmıştır (UNDP, 2021; World Bank, 2021). Göstergelerin seçiminde, ülkelerin eğitimde gelişmişliğini gösteren ve son beş yıla ait verisi bulunan ölçütler dikkate alınmıştır. Buna göre, kamu harcamaları içindeki toplam eğitim payı, beklenen eğitim yılı, ortalama eğitim yılı, okulöncesi çağındaki çocukların eğitime katılım yüzdesi ve temel eğitimde öğretmen başına düşen ortalama öğrenci sayısı göstergeleri belirlenmiştir. Bu yolla elde edilen eğitimde gelişmişlik indeksine dayalı olarak ülkelerin eğitim başarıları karşılaştırılmıştır. Daha sonra, bu ülkeler arasından PISA 2018'e katılanlar belirlenmiş ve bu ülkeler okuma becerileri başarısı açısından sıralanmıştır (OECD, 2019; UNDP, 2020). Bu sıralamaya bağlı olarak ülkelerin hem eğitimde gelişmişlik düzeyi açısından durumu hem de okuma başarısı açısından durumu karşılaştırılmıştır.

Bu çalışmada, ülkelerin eğitimde gelişmişlik düzeylerine ilişkin göstergelere göre kümelere ayrılması ve buna bağlı olarak bir eğitimde gelişmişlik indeksi oluşturulması amacıyla bulanık c-ortalama (BCO) kümeleme algoritmasından yararlanılmıştır. Bulanık kümeleme yöntemi, nesnelerin kümelere hangi dereceyle ait olduğunu belirleyen üyelik fonksiyonlarını hesaplamak ve veri seti içerisindeki örtüşen kümeleri saptamak için kullanılmaktadır (De Oliveira ve Pedrycz, 2007). Yaygın olarak kullanılan kümeleme yöntemlerinden biri olan BCO kümeleme algoritması ilk olarak Dunn (1974) tarafından önerilmiş ve Bezdek (1981) tarafından geliştirilmiştir.

Klasik kümeleme algoritmaları, her bir nesnenin bir kümeye kesin sınırlar ile ait olduğu düşüncesine göre oluşturulur. Ancak nesnelerin ait olabileceği sınıfların sınırları her zaman kesin olarak tanımlanamayabilir. Bu gibi durumlarda ve çoklu karmaşık karakteristikler mevcut ise bulanık kümeleme yöntemi, sistemi en iyi temsil edecek bir model oluşturmak üzere etkili bir yol sağlamaktadır. Bu çalışmada, önerilen bulanık tip kümelemeyle örüntülerin üyelikleri hakkında daha fazla bilgi sağlanması amaçlanmaktadır. Bulanık c-ortalama kümeleme yönteminde örüntüler, farklı üyelik dereceleri ile kümelere dahil olabilir (Nayak ve diğ., 2015).

Sonuçlar ve Tartışma

UNDP eğitim indeksini, 2010 yılından itibaren yalnızca "ortalama eğitim süresi" ve "beklenen eğitim süresi" göstergelerine dayalı olarak hesaplamaktadır. Bu çalışmada farklı göstergeler de dikkate alınarak dünya ülkeleri için yeni bir eğitimde gelişmişlik indeksi hesaplanmıştır. Bu çalışmada elde edilen eğitimde gelişmişlik indeksi puanı, UNDP tarafından 2019 yılında hesaplanan eğitim indeksi puanıyla yüksek korelasyon vermiştir. Eğitim kararlarına dikkat çekmesi açısından bu çalışmada elde edilen eğitimde gelişmişlik indeksi bulgularının kullanılması önemlidir.

Türkiye'nin eğitimde gelişmişlik indeksi puanı ortalama değerdedir. Hesaplanan indeks sayesinde Türkiye'nin dünya ülkeleri içindeki yeri daha fazla göstergeye dayalı ortaya koyulmuştur. Türkiye'nin sıralamadaki yerini yukarıya taşıması öncelikle; okulöncesi çağındaki çocukların eğitime katılım yüzdesini arttırmasını, temel eğitimde öğretmen başına düşen ortalama öğrenci sayısını azaltmasını ve kamu harcamaları içindeki toplam eğitim payını arttırmasını gerektirmektedir.

Ayrıca eğitimde gelişmişlik indeksi ile PISA 2018'e katılan ülkelerin okuma becerileri başarısı ilişkilendirilmiştir. Bu çalışma PISA okuma becerisi puanlarındaki artışın aynı zamanda eğitimde gelişmişlik indeksi puanındaki artışa da bağlı olduğunu göstermektedir. Çalışmada kullanılan göstergelerden hareketle, Türkiye'nin eğitimde gelişmişlik düzeyi farklı açılardan tartışılmıştır. Bu sayede uzun yıllardır, PISA gibi uluslararası öğrenci başarısını belirleme çalışmalarının sonuçlarına göre değerlendirilen Türk eğitim sistemine yeni bir bakış açısı sağlayacak bulgular ortaya koyulmaya çalışılmıştır. Bunun yanında, belirlenen göstergeler açısından indeks kapsamı geliştirilerek ülkelerin daha nesnel, yansız ve etkili biçimde karşılaştırılabilirliği arttırılmıştır.

Kaynaklar

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. Plenum Press.

De Oliveira, J. V. & Pedrycz, W. (2007). *Advances in fuzzy clustering and its applications*. Wiley.

Dunn, J. C. (1974). A fuzzy relative ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybern, 3*, 32-57. https://doi.org/10.1080/01969727308546046

Nayak, J., Naik B., & Behera, H. S. (2015). Fuzzy C-Means (FCM) clustering algorithm: A decade review from 2000 to 2014. In *Computational Intelligence in Data Mining (Volume 2)*. L. Jain, H. S. Behera, J. K. Mandal, and D. P. Mohapatra, (Eds.), (pp. 133-149). Springer.

OECD (2010). *PISA 2009 results (Volume I): What students know and can do - Student performance in reading, mathematics, and science.* PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264091450-en

OECD (2019). *PISA 2018 results (Volume I): What students know and can do.* PISA, OECD Publishing. https://doi.org/10.1787/5f07c754-en

UNDP (1990). *Human development report 1990. Concept and measurement of human development.* Oxford University Press. http://www.hdr.undp.org/en/reports/global/hdr1990

UNDP (2010). *Human development report 2010* (20th Anniversary Edition). *The real wealth of nations: Pathways to human development.* Palgrave Macmillan. http://hdr.undp.org/sites/default/files/reports/270/hdr_2010_en_complete_reprint.pdf

UNDP (2011). *Human development report 2011. Sustainability and equity: A better future for all.* Palgrave Macmillan. http://hdr.undp.org/en/content/human-development-report-2011

UNDP (2020). *Human development report 2020. The next frontier - Human development and the Anthropocenel.* AGS, RR Donnelley Company. http://hdr.undp.org/sites/default/files/hdr2020.pdf

UNDP (2021). *Human development data center.* Retrieved July 21, 2021 from http://hdr.undp.org/en/data

Villar, A. (2013). The educational development index: A multidimensional approach to educational achievements through PISA. *Modern Economy, 4,* 403-411. http://dx.doi:10.4236/me.2013.45042

World Bank. (2021). *World Development Indicators database.* http://data.worldbank.org

E-değerlendirme sistemlerinde madde seçim algoritmalarına ilişkin sorunlar

Osman Tat

Giriş

Genel olarak uzaktan eğitim veya uzaktan öğretim olarak adlandırılan paradigma değişimi tüm dünyada olduğu gibi Türkiye'de de beraberinde uzaktan ölçme ve değerlendirme kavramını daha yoğun bir biçimde gündeme getirmiştir. Gelişen internet ve bilişim teknolojisi ile yaygınlaşan uzaktan eğitimin popülerliğinin artmasında, öğrencilere sunduğu işlevsellik ve esnekliğin yanında yüz yüze eğitimde karşılaşılan zaman ve maliyet konularında sunduğu avantajlar da yatmaktadır (Bender, 2003). Uzaktan eğitim kavramı gibi uzaktan-değerlendirme veya çevrimiçi değerlendirme gibi isimler ile de anılan e-değerlendirme kavramı, bilgisayarlar ve günümüzde daha sık kullanılan mobil cihazların değerlendirme yapmak, puanları analiz etmek ve sonuçları raporlamak için kullanılmasını sağlayan teknikler bütünü olarak tanımlanabilir (Graff, 2003). Bennet (1998) bilişim teknolojilerinin ilerlemesi ile birlikte bireylerin kişisel özelliklerine uyarlanmış tekniklerin, simülasyonların, yeni soru biçimlerinin ve sanal gerçeklik uygulamalarının yeni değerlendirme unsuları olarak rol oynayacağını savunmuştur. Bennet'in (1998) düşüncelerini sunduğu tarihin üzerinden geçen yaklaşık yirmi yılda sanal öğrenme ortamları, sanal deneyler, simülasyonlar ve karmaşık modellerin görselleştirilmesini sağlayan yazılımlar artık günümüzün değerlendirme araçları arasına girmiştir. (Fill ve Ottewill, 2006). Uzaktan eğitim ortamlarında değerlendirmede kullanılabilecek ölçme araçlarının dört ana kategoride toplandığı söylenebilir. Bunlar: yazılı ödevler, otomatik puanlanabilen sınavlar, çevrimiçi tartışmalar ve internet yayıncılığıdır. Otomatik puanlanabilen değerlendirme araçları arasında multimedya eşleştirme soruları ve sürükle-bırak soruları içeren sınavlar ve simülasyonlar gibi nispeten yeni yöntemler olduğu gibi çoktan seçmeli, kısa yanıtlı, eşleştirme ve boşluk doldurma benzeri soru türlerinden oluşan sınavlar gibi geleneksel yöntemler hala en yaygın ölçme araçlarıdır (Benson, 2010).

Çoktan seçmeli veya doğru-yanlış soru tipleri, birden fazla durum içerisinden bir veya daha fazla seçeneğin öğrenciler tarafından doğru yanıt olarak seçildiği soru tipleridir. Bu soru tiplerinden oluşan sınavlar aracılığı ile özellikle bilgi ve kavrama gibi alt bilişsel becerilere ilişkin ölçümleri pratik bir şekilde gerçekleştirmek mümkündür (McVey, 2016). Her ne kadar bu sınav türleri uygulama açısından pratik olsa da, hazırlanma sürecinde dikkat edilmesi gereken bir çok yön bulunmaktadır. Diğer tüm ölçme araçlarının geliştirilmesi sürecinde olduğu gibi geleneksel çoktan seçmeli sınavlar ve benzeri ölçe araçları

geliştirme sürecinde de madde ve teste ilişkin analizlerin yürütülmesi ve analiz sonucunda uygun maddelerin teste dahil edilmesi gerekir. Klasik test kuramı veya madde tepki kuramına dayalı olarak ön uygulamadan elde edilen verilerden elde edilen madde güçlük ve ayırıcılık indeksleri, madde çeldirici analizi sonuçları, ortalama test güçlüğü ve ayırıcılığı gibi istatistikler etkili bir ölçme aracı geliştirme aşamasında kullanılan en yaygın ölçütlerdendir. Birçok üniversiteye ve okula uzaktan öğretim ve e-değerlendirme hizmeti sunan servisler madde güçlüklerine dayalı (kolay, orta güçlükte, zor) madde seçim algoritmaları kullanmaktadır. Ancak, çoğu zaman test geliştirme aşamasında pilot uygulama yapılmamakta ve madde güçlük düzeyleri test geliştiricilerin öznel yargılarına dayanabilmektedir. Bu durum e-değerlendirmede uygun ortalama güçlük ve ayırıcılığa sahip olmayan sınavların oluşmasına sebep olabilmektedir. Başka bir ifade ile sadece öğretim elemanlarının maddeye ilişkin öznel yargısı ile oluşturulan testler ya çok zor ya da çok basit olabilmektedir.

Yöntem

Bu çalışmanın amacı birçok üniversiteye uzaktan öğretim ve e-değerlendirme hizmeti sunan sistemlerde test geliştiricilerin öznel yargılarına dayalı olarak belirtilen madde güçlük düzeyleri ile sınav sonucunda ortaya çıkan madde istatistiklerinin ne düzeyde örtüştüğünü belirlemektir. Bu araştırma kapsamında 2020-2021 akademik yılında Van Yüzüncü Yıl Üniversitesi'nde işlenen Eğitimde Ölçme ve Değerlendirme lisans dersinin vize, final ve bütünleme sınavlarından elde edilen madde ve test istatistikleri ile üç ölçme ve değerlendirme uzmanının her bir madde için öngördüğü madde istatistikleri arasındaki uyum betimlenecektir. Bu yönü ile çalışmanın betimsel tarama deseninde (Fraenkel ve diğ., 2012) olduğu söylenebilir.

Araştırmada kullanılacak ölçme aracı Eğitimde Ölçme ve Değerlendirme dersi kapsamında geliştirilen 108 maddelik soru havuzudur. Araştırmanın verileri belirtilen dersin vize, final ve bütünleme sınavına giren, farklı bölümlerden toplam 120 lisans öğrencisinden toplanmıştır. Çalışmada üç ölçme ve değerlendirme uzmanından madde havuzundaki her bir maddenin güçlük düzeyini e-değerlendirme sisteminde tanımlandığı gibi üç kategoriden (basit, orta güçlükte, zor) birini seçecek şekilde belirtesi istenecektir. Ardından bu uzman görüşlerinin hem kendi aralarındaki uyumu hem de sınavlardan elde edilen istatistiklerle uyumu incelenecektir. Uyum analizinde öncelikle uzmanların görüşleri arasındaki uyum incelenecektir. Bu inceleme sonucunda aynı zamanda puanlayıcılar arası güvenirliğin düzeyine ulaşılacaktır. Ardından, sınavlardan elde edilen verilerin analizi ile maddeler, güçlük indekslerine göre 'basit', 'orta güçlükte' ve 'zor' şeklinde kategorize edilecektir. Hem uzman görüşleri arasındaki uyum hem de kategorize edilen gerçek madde güçlükleri ile kategorik olan uzman görüşleri arasındaki uyum Cramer'in V korelasyon katsayısı ile incelenecektir.

Sonuçlar

Araştırma sonucunda öznel yargılara dayalı madde güçlükleri ile sınav sonuçlarına dayalı gerçek madde güçlükleri arasında önemsiz veya küçük düzeyde bir ilişki çıkması beklenmektedir. Buna karşın uzman görüşleri arasında yüksek düzeyde bir ilişkinin çıkması öngörülmektedir. Belirli bir dersi veren

öğretim elemanlarının aşina oldukları derslerin sınav sorularının kolay olduğunu düşünme eğilimde oldukları varsayılmaktadır. Bundan dolayı da oluşturulan sınavların öngörülen ortalama güçlüğünün gerçek güçlük düzeyinden farklı çıkma olasılığının yüksek olduğu düşünülmektedir. E-değerlendirme ortamlarında herhangi bir ön uygulamaya yapmadan sadece öznel düşünceye dayalı olarak madde güçlükleri belirtildiğinde algoritmaların tüm yetenek düzeylerini ölçebilecek sınavlar oluşturma konusunda sorunlar yaşayabileceği düşünülmektedir.

Kaynaklar

Bender, T. (Ed.). (2003). *Discussion-based online teaching to enhance student learning: Theory, practice, and assessment*. Stylus Publishing. https://doi.org/10.1111/j.1467-9647.2006.00283_8.x

Bennett, R. E. (1998). *Reinventing assessment. Speculations on the future of large-scale educational testing. A policy information perspective*. Policy Information Center. Retrieved from https://www.researchgate.net/publication/234731732_Reinventing_Assessment_Speculations_on_the_Future_of_Large-Scale_Educational_Testing_A_Policy_Information_Perspective

Benson, R. (2010). *Online learning and assessment in higher education: A planning guide*. Oxford: Chandos Publishing.

Fill, K., & Ottewill, R. (2006). Sink or swim: taking advantage of developments in video streaming. *Innovations in Education and Teaching International*, *43*(4), 397–408. https://doi.org/10.1080/14703290600974008

Fraenkel, J. R., Wallen, N. E., and Hyun, H. H. (2012). How to design and evaluate research in education (8th ed.). McGraw-Hill Humanities/Social Sciences/Languages.

Graff, M. (2003). Cognitive style and attitudes towards using online learning and assessment methods. *Electronic Journal of E-Learning*, *1*(1), 21-28. https://files.eric.ed.gov/fulltext/EJ1099141.pdf

McVey, M. (2016). Preservice teachers' perception of assessment strategies in online teaching. *Journal of Digital Learning in Teacher Education*, *32*(4), 119–127. https://doi.org/10.1080/21532974.2016.1205460

Yoksunluk içinde başarmanın yordayıcıları: TIMSS 2019 Türkiye örneği

Burcu Parlak vee Ahmet Yıldırım

Anahtar kelimeler: Yoksunluk içinde başaranlar, sosyoekonomik düzey, TIMSS, matematik başarısı, akademik yılmazlık

Giriş

Akademik başarıda sosyoekonomik düzeyden kaynaklanan farklılıklar son yıllarda araştırmacıların ve eğitim alanında karar vericilerin dikkatini çekmektedir. Sosyoekonomik olarak dezavantajlı çocukların bu bakımdan kendilerinden daha avantajlı akranları ile karşılaştırıldığında; okuldan ayrılma, sınıfta kalma, okulda ve uluslararası değerlendirme çalışmalarında (PISA, TIMSS gibi) düşük performans gösterme olasılıklarının daha yüksek olduğu bilinmektedir (Agasisti ve diğ., 2018; Sandoval-Hernandez ve Bialowolski, 2016).

Son yıllarda eğitime erişim anlamında tüm dünyada ciddi ilerlemeler gözlenmektedir. Eğitime erişim oranları arttıkça sosyoekonomik olarak dezavantajlı grupların da eğitime entegrasyonu sağlanmakta ve bu durum, hiç kuşkusuz, eğitime erişimde eşitlik ilkesine hizmet etmektedir. Ancak aynı zamanda bu durum *adalet* ve *eğitim kalitesi* ile ilgili bazı sorunları da beraberinde getirmektedir. Yani düşük sosyoekonomik düzeyden gelen öğrenciler zorlu koşullarda yaşamakta ve akademik çalışmalarını yürütmekte, bu yüzden sosyoekonomik olarak kendilerinden daha avantajlı durumda olan akranlarıyla karşılaştırıldığında daha farklı eğitim gereksinimlerine ihtiyaç duymaktadır (Sandoval-Hernandez ve Cortes, 2012).

Her ne kadar sosyoekonomik bakımdan dezavantajlı olmak, eğitimde başarısızlıkla ilişkilendirilse de sosyoekonomik açıdan dezavantajlı bütün öğrencilerin başarısızlığa mahkûm olduğunu iddia etmek doğru olmaz. Burada *yılmazlık, dayanıklılık (resilience)* kavramı gündeme gelmektedir. Yani sosyoekonomik bakımdan dezavantajlı olmasına rağmen akademik başarı gösteren öğrencileri tanımlamak için *akademik yılmazlık/dayanıklılık* (academic resilience) kavramı kullanılmaktadır (Kalender, 2015; Sandoval-Hernandez ve Cortes, 2012). *Yılmazlık* kavramı; kişisel, mesleki ve akademik hedeflere ulaşmada karşılaşılan zorlukların üstesinden gelme yeteneği olarak tanımlanmakta ve genelde sosyal bilimlerde, özelde de pozitif psikoloji alanında son yıllarda sıklıkla ele alınmaktadır (Coronado-Hijon, 2016). Akademik yılmazlık gösteren öğrenciler yoksunluk içinde başarıyı yakalayan öğrenciler

olarak tanımlanmakta ve bu öğrencilerin okula düzenli katılmaları veya aileleri tarafından sosyal destek görmeleri durumunda başarılarını devam ettirecekleri öne sürülmektedir (Agasisti ve diğ., 2018; Coronado-Hijon, 2016; Mallick ve Kaur, 2016).

Sosyoekonomik bakımdan dezavantajlı olmasına rağmen akademik başarı gösteren ve bu başarıyı sürdürülebilir kılan öğrenciler *akademik olarak yılmaz öğrenciler* biçiminde tanımlanmaktadır. Bu kavram işevuruk olarak farklı şekillerde tanımlanmakla birlikte OECD tarafından *bir öğrencinin sosyoekonomik bakımdan dezavantajlı olmasına rağmen akademik olarak başarı gösterme olasılığı* biçiminde ifade edilmektedir. Burada ifade edilen dezavantaj, uluslararası bir değerlendirme çalışması olan PISA'daki *ekonomik, sosyal ve kültürel statü* indeksi bakımından dezavantajlı olmaya karşılık gelmektedir. OECD yayınları dikkate alındığında, bu indeks bakımından dağılımın birinci çeyreğinin altında yer alan öğrenciler (25. yüzdelik dilim ve altında yer alan öğrenciler) dezavantajlı olarak değerlendirilmektedir (Agasisti ve diğerleri, 2018). Bir öğrenciyi eğitim yaşantıları bakımından *dezavantajlı* yapan koşullar veya değişkenler; *olumsuz aile, okul koşulları/özellikleri ve düşük sosyoekonomik düzey* vb. olarak tanımlanabilir (Aydıner ve Kalender, 2015; Kalender, 2015; Mallick ve Kaur, 2016; Rojas, 2015).

Yapılan çalışmalar incelendiğinde (Aydıner ve Kalender, 2015; Kalender, 2015; Rojas, 2015; Sandoval-Hernandez ve Biaowolski, 2016) akademik yılmazlıkla ilişkili faktörlerin ortaya konulmasının, sosyoekonomik bakımdan dezavantajlı olmasına rağmen yüksek akademik başarı gösteren öğrenci kitlesinin daha iyi anlaşılması ve bu öğrencilerle ilgili çalışmalar yapılabilmesi açısından önemli olduğu düşünülmektedir. Bu nedenle, yakın zamanda gerçekleştirilen ve sonuçları kamuoyuyla paylaşılan bir uluslararası değerlendirme çalışması olan TIMSS 2019 Türkiye verileri kullanılarak yoksunluk içinde başarmayı yani akademik yılmazlığı yordayan değişkenlerin belirlenmesine ihtiyaç duyulmuştur. Bu araştırmanın amacı, yoksunluk içinde başarmayı yani akademik yılmazlığı yordayan değişkenlerin belirlenmesidir.

Yöntem

Bu araştırma, değişkenler arasındaki ilişkilerin incelenmesinin amaçlandığı bir araştırmadır. Bu amaca uygun olacak şekilde bu araştırmada ilişkisel tarama modeli kullanılmıştır. İlişkisel tarama modeli, iki veya daha fazla değişken arasında ilişki olup olmadığını ve/veya ilişkinin derecesini ortaya koymayı amaçlayan bir araştırma modelidir (Karasar, 2008). Araştırma grubunu, TIMSS 2019 uygulamasına Türkiye'den katılan sekizinci sınıf öğrencileri oluşturmaktadır.

Araştırma kapsamında TIMSS 2019 uygulamasına Türkiye'den katılan sekizinci sınıf öğrencilerin matematik testine ait verileri kullanılmıştır. Uygulamaya Türkiye'den sekizinci sınıf düzeyinde 4077 öğrenci katılmıştır (MEB, 2020). Yoksunluk içinde başaran öğrencileri belirlemek amacıyla öncelikle sosyoekonomik düzey indeksi bakımından ilk çeyrekte yer alan öğrencilerin belirlenmesi gerekmektedir (Agasisti ve diğerleri, 2018). TIMSS araştırmasında *evdeki kitap sayısı, evdeki olanaklar ve velilerin eğitim düzeyi* değişkenleri dikkate alınarak her bir öğrenciye ait *sosyoekonomik düzeyi* gösteren 0 ile 10 arasında değişen bir indeks kestirilmektedir (Broer ve diğ., 2019). Bu indeks dikkate alınarak öncelikle dağılımın

ilk çeyreğinde yer alan 1019 öğrenci belirlenmiştir. Daha sonra ise bu öğrenciler arasından TIMSS matematik puanı, TIMSS ölçek orta noktası olarak kullanılan 500 puanın (Mullis, Martin Foy, Kelly ve Fishbein, 2020) üzerinde yer alan 242 öğrenci seçilmiş ve bu öğrenciler *yoksunluk içinde başaran öğrenciler* olarak tanımlanarak analizler bu gruba ait veriler üzerinde yürütülmüştür.

Yoksunluk içinde başaran öğrencilerin başarılarını yordayan değişkenlerin belirlenmesi amacıyla çoklu doğrusal regresyon analizi kullanılmıştır. Çoklu doğrusal regresyon analizi, yordanan değişkenle ilişkili olan iki ya da daha fazla yordayıcı değişkene dayalı olarak yordanan değişkenin tahmin edilmesine yönelik bir analiz türüdür (Büyüköztürk, 2005).

Sonuçlar

TIMSS 2019 uygulama sonuçlarına göre yoksunluk içinde başaran öğrencilerin matematik puanlarını yordayan değişkenleri belirlemek amacıyla gerçekleştirilen çoklu doğrusal regresyon analizi sonuçlarına göre, değişkenler yoksunluk içinde başaran öğrencilerin matematik puanları ile orta düzeyde ve anlamlı bir ilişki vermektedir (R= .457, R^2= .209, p< .05). Yordayan değişkenler, sekizinci sınıf matematik başarısındaki varyansın yaklaşık %21'ni açıklamaktadır. Standardize edilmiş regresyon katsayısına göre yordayıcı değişkenlerin yoksunluk içinde başaran öğrencilerin matematik puanları üzerindeki göreli önem sırası; *matematik dersinde kendine duyulan güven, matematik öğrenmeyi sevme*, *akran zorbalığı*, *okula aidiyet* ve *matematiğe verilen değerdir*. Regresyon katsayılarının anlamlılığına ilişkin t-testi sonuçları incelendiğinde ise sadece *matematik dersinde kendine duyulan güven* ve *akran zorbalığı* değişkenlerinin matematik başarısı üzerinde anlamlı bir yordayıcı olduğu görülmektedir. Diğer değişkenlerin matematik başarısı üzerinde anlamlı bir etkiye sahip olmadıkları belirlenmiştir.

Kaynaklar

Agasisti, T., Avvisati, F., Borgonovi, F. ve Longobardi, S. (2018). *Academic resilience: What schools and countries do to help disadvantaged students succeed in PISA*. OECD Education Working Papers, No. 167.

Aydıner, A., and Kalender, İ. (2015). Student segments based on the factors related to sense of belonging across disadvantaged and resilient groups in PISA 2012. *Procedia - Social and Behavioral Sciences*, *174*, 3299-3305. https://doi.org/10.1016/j.sbspro.2015.01.997

Broer, M., Bai, Y., and Fonseca, F. (2019). *Socioeconomic inequality and educational outcomes. Evidence from twenty years of TIMSS*. Springer.

Büyüköztürk, Ş. (2005). *Sosyal bilimler için veri analizi el kitabı*. Pegem Akademi.

Coronado-Hijon, A. (2016). Academic resilience: A transcultural perspective. *Procedia - Social and Behavioral Sciences*, 237, 594-598. https://doi.org/10.1016/j.sbspro.2017.02.013

Kalender, İ. (2015). An analysis of the resilient students' profile based on PISA 2012. *Journal of Measurement and Evaluation in Education and Psychology*, *6*(1), 158-172. https://doi.org/10.21031/epod.16925

Karasar, N. (2008). *Bilimsel araştırma yöntemi*. Nobel Akademik Yayıncılık.

Mallick, M. K. and Kaur, S. (2016). Academic resilience among senior secondary school students: İnfluence of learning environment. *Rupkatha Journal on Interdisciplinary Studies in Humanities, 8*(2), 20-27. https://doi.org/10.21659/rupkatha.v8n2.03

MEB (2020). *TIMSS 2019 Türkiye ön raporu* (Eğitim Analiz ve Değerlendirme Raporları Serisi No. 15). Milli Eğitim Bakanlığı. http://www.meb.gov.tr/meb_iys_dosyalar/2020_12/08202713_No15_TIMSS_2019_Turkiye_On_Raporu.pdf

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, L. D., and Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science.* Boston College, TIMSS & PIRLS International Study Center.

Rojas, L. F. (2015). Factors affecting academic resilience in middle school students: A case study. *GIST Education and Learning Research Journal*, 11, 63-78. https://files.eric.ed.gov/fulltext/EJ1084404.pdf

Sandoval-Hernandez, A., and Cortes, D. (2012). *Factors and conditions that promote academic resilience: A cross-country perspective.* Paper presented at Comparative and International Education Society, San Juan, Puerto Rico.

Sandoval-Hernández, A., Białowolski, P. Factors and conditions promoting academic resilience: a TIMSS-based analysis of five Asian education systems. Asia Pacific Educ. Rev. 17, 511–520 (2016). https://doi.org/10.1007/s12564-016-9447-4

TIMSS 2019 8. Sınıf matematik maddelerinin Rasch ağacı yöntemi ile değişen madde fonksiyonu açısından incelenmesi

Alperen Yandı ve Hüseyin Yıldız

Giriş

Eğitim öğretim süreçlerinde yapılan ölçme ve değerlendirme uygulamaları, küçük ölçekli-sınıf içi ve geniş ölçekli olarak sınıflandırılabilir. Eğitim politikalarının belirlenmesi bağlamında düşünüldüğünde geniş ölçekli test uygulamalarından elde edilen sonuçlar, ülkeler son derece önemlidir. Geniş ölçekli test uygulamaları, bilişsel alan alt testleri ve eğitim öğretim süreci paydaşlarına yönelik sosyal, demografik ve psikolojik özellikleri ölçmeye yönelik araçların birlikte uygulanması ile gerçekleştirilir. Bu uygulamaların, eğitim kalitesinin izlenmesi, eğitim öğretim programlarının değerlendirilmesi, öğrencilerin çeşitli alanlara yönelik okuryazarlık düzeylerinin belirlenmesi gibi amaçları mevcuttur.

ABİDE, PISA ve PIRLS gibi farklı uygulamaları mevcut olan bu test uygulamalarının bir örneği de TIMSS uygulamalarıdır. Uluslararası Matematik ve Fen Eğilimleri Araştırması her dört yılda bir farklı ülkelerde öğrenim görmekte olan dördüncü ve sekizinci sınıf öğrencileri katılmaktadır. 2019 yılında yapılan son uygulamaya dördüncü sınıf düzeyinde 58 ülke, sekizinci sınıf düzeyinde ise 39 ülke katılmıştır.

Eğitim öğretim sürecine ilişkin önemli kararlar alınmasında rol oynayan tüm sınavların geçerlik ve güvenirlik özelliklerinin tartışılmayacak şekilde kanıtlanmış olması gerekmektedir. Bu tür sınavlardan elde edilen sonuçlar doğrultusunda verilen kararların isabetli olması, sınavların psikometrik özellikleri ile yakından ilişkilidir (Yalçın, 2018). Geçerliğin bir test uygulamasını anlamlı kılan en önemli özellik (Zumbo, 1999) olduğu göz önüne alındığında, bir testin ölçmek istediği özellikleri, farklı değişkenlerle karıştırmadan ölçebilmesi (Thorndike ve Hagen, 1961) olarak tanımlanan bu özelliğe ilişkin kanıtların elde edinilmesi zorunluluk haline gelmektedir. Ölçme uygulamalarda, amaç dışı bir değişkenin ölçme sürecine dahil olma durumu, ölçme aracının yapı geçerliğini olumsuz yönde etkilemektedir (Doğan & Öğretmen, 2005). Bireylere ait demografik değişkenlerin (cinsiyet, kültür vb.) aynı yetenek düzeyindeki bireylerin maddeleri doğru yanıtlama olasılığı üzerinde farklılaşma yaratması, amaç dışı değişkenlerin ölçme sürecine dahil olmasının en önemli örneklerindedir. Sistematik hata sınıfında değerlendirilen bu tür örnekler, katılımcı grubundaki belirli bir alt küme için yanlılık oluşturmaktadır (Camilli 2006). Bu nedenle yapılan ölçme uygulamalarında, ölçme araçlarının geçerliğine ilişkin soru işaretlerinin ortadan

kaldırılabilmesi için yanlılık olup olmadığı ve yanlılık tespit edilmesi durumunda olası kaynaklar tespit edilmeye çalışmaktadır.

Yanlılık belirleme sürecinde gerçekleştirilen istatistiki incelemeler değişen madde fonksiyonu (DMF) analizlerini kapsamaktadır. DMF analizleri, bir ölçme uygulamasına katılan ve belirli bir demografik değişkene (cinsiyet, kültür vb.) göre farklı alt grupta yer alan, aynı yetenek düzeyindeki bireylerin bir maddeyi doğru yanıtlama olasılığında bir farklılık olup olmadığını tanımlamak için yapılmaktadır (Hambleton ve Rogers, 1989). DMF analizlerinde farklı kuramları temel alan birçok yöntem mevcuttur. Yaygın yöntemlerin analiz süreçleri incelendiğinde belirli sınırlılıkları olduğu gözlenebilir. Bu yöntemlerin kullanılabilmesi için odak ve referans grupları oluşturulması bir gerekliliktir. Bir başka ifadeyle DMF'ye yol açabilecek olası alt katılımcı grupları tanımlanmasına ihtiyaç duyulmaktadır. DMF yöntemleri için bu özellik bakımından yapılacak sınıflamada ön tanımlı olarak nitelendirilen bu yöntemlerle elde edilen sonuçların analizleri nispeten kolay olsa da bu durum bir sınırlılık olarak ele alınmaktadır. Yaygın olarak kullanılan bu yöntemlerin ön tanımlı olma durumlarına ilişkin sınırlılık için ön tanımlı olmayan örtük sınıf modelleri kullanılabilir. Ancak bu yöntemlerinde yorumlanma süreci araştırmacılar için zor olabilmektedir. Bu yöntemler için bir diğer sınırlılık, sürekli yapılı bir değişkenin analiz sürecine doğrudan dahil edilmesi mümkün olmamasıdır. Bir sürekli değişken ancak kategorik hale getirilerek analize dahil edilebilmektedir ki bu durum bilgi kaybına yol açmaktadır. Ayrıca bu yöntemlerle birden fazla değişkenin birlikte analize dahil edilmesi söz konusu olmamaktadır. Strobl ve diğ. (2015) tarafından bu sınırlılıklardan en az etkilenerek DMF analizi yapılmasına imkan tanıyan bir yöntem olan Rasch ağaçları yöntemi ortaya atılmıştır. Rasch ağacı yöntemi, araştırmacılara örtük sınıf modelleri gibi ön tanımlı olmayan bir değişken için analiz yapabilme ve ön tanımlı yöntemler gibi kolay yorumlama olanağı sağlamaktadır.

Yapılan bu çalışmada 2019 TIMSS uygulamasında uygulanan Kitapçık-1'de yer alan maddelerin Rasch Ağacı yöntemi ile birden fazla kategorik ve sürekli değişken için DMF analizlerinin yapılması amaçlanmaktadır. Ayrıca görece yeni bir yöntem olan Rasch ağacı yöntemi ile ilgili katılımcıların bilgilendirilmesi hedeflenmektedir.

Yöntem

Var olan bir durumu betimleme amacıyla yapılan bu çalışma tarama modeline göre planlanmıştır. Çalışma grubu için, 8. sınıf düzeyinde gerçekleştirilen, TIMMS 2019 elektronik uygulamalarına katılan ülkelerden beş tanesi seçilmiştir. Yapılan ülke seçiminde Türkiye (TUR), Fransa (FRA), Singapur (SGP), Amerika Birleşik Devletleri (USA), Ontario/Kanada (ONT) ve Şili (CHL) ülkelerinden katılan öğrenciler çalışma grubuna dahil edilmiştir. Ülkelerin matematik öğrenme alanları ortalama ölçek puanları göz önüne alınarak yapılan seçimde, farklı düzeyleri temsil edebilecek nitelikte ülkeler belirlenmiştir. Ontario/Kanada ise karşılaştırma birimi olarak belirlenmiş olan ülkeleri temsil etmesi amacıyla alınmıştır. Bu altı ülkeden 2020 öğrenciye ait Kitapçık-1 yanıtları veri setini oluşturmaktadır. Toplamda 48 maddeye ait yanıtlar analizlerde kullanılmıştır. Buna göre beş değişken, bireylere ait belirli

bilgileri (ülke, cinsiyet, bilgisayar/tablet sahipliği, matematiğe karşı tutum ve ortalama cevaplama süresi) içermekte iken; 43 madde ise matematik öğrenme alanları ile ilgilidir. Matematik öğrenme alanında bulunan tüm maddeler içinde 1-0 şeklinde puanlanan maddeler seçilmiştir. Tüm veriler kontrol edilerek, ilk beş sorudan herhangi birinde kayıp değer içeren ve matematik öğrenme alanlarında %10'dan fazla yanıtsız bırakılmış madde olan katılımcılar veri setinden çıkarılmıştır.

Yapılan DMF analizlerinde, ülke (COUNTRY), cinsiyet (GENDER), bilgisayar/tablet sahipliği (COMPUTER_TABLET), matematiğe karşı tutum (ATTITUDE) ve ortalama cevaplama süresi (Z_TIME) değişkenleri bir arada kullanılmıştır. Ortalama cevaplama süresi değişkeni, sonuçların doğru şekilde yorumlanabilmesi için standart Z puanlarına dönüştürülerek analizlere dahil edilmiştir. Rasch ağacının en önemli avantajı olarak, birbirinden farklı sürekli ve süreksiz yapıdaki değişkenler birlikte analize dahil edilmiştir. Rasch ağaçları, referans ve odak grup kavramlarından bağımsız şekilde analizlerin gerçekleştirilmesine olanak sunmaktadır. Analize dahil edilen değişkenleri önem sırasına koyarak, alt kümelere bölmektedir. Buna ek olarak sürekli değişkenler için kırılma noktaları tespiti sağlamaktadır. Rasch ağacı analizleri için R açık kaynak kodlu platformda bulunan *"psychotree"* paketi (Strobl, Wickelmaier & Zeileis, 2015) kullanılmıştır. Paketin içerisinde bulunan *"raschtree"* fonksiyonu işe koşularak yapılan analizler sonrasında Mantel Haenszel yöntemiyle, madde düzeyinde DMF varlığına ve düzeyine ilişkin incelemeler yapılmıştır.

Sonuçlar

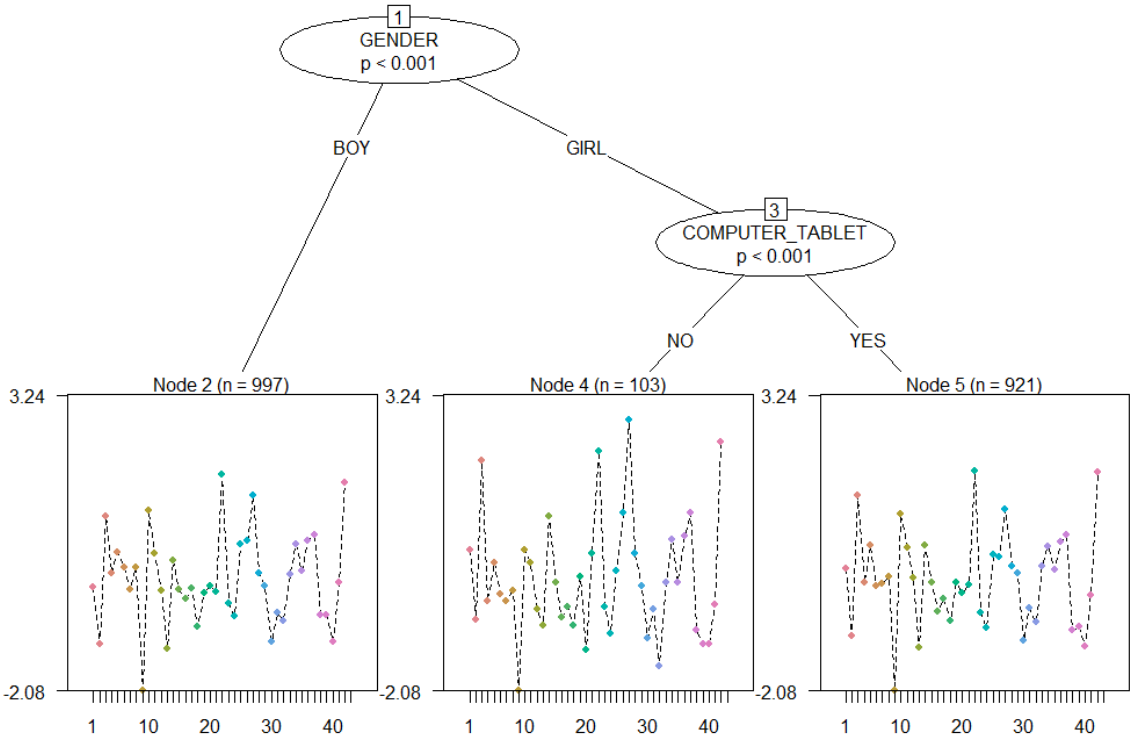
Rasch ağacı yöntemi ile yapılan analiz sonucunda her bir değişken için elde edilen DMF olduğu tespit edilen alt gruplara ve DMF büyüklüklerine ilişkin bulgular bu bölümde paylaşılmıştır. Her bir değişken için elde edilen sonuçlar ayrı tablolar ve rasch ağaçları şeklinde sunulmuştur. Cinsiyet ve bilgisayar tablet sahipliği değişkenleri bir arada analize dahil edilmiş ve sonuçlar bu doğrultuda rapor edilmiştir. İlk olarak bu iki kategorik değişkene ait sonuçlar Tablo 1'de verilmiştir.

Tablo 1

Cinsiyet ve Bilgisayar / Tablet Sahipliği Değişkenine İlişkin Bulgular

	Cinsiyet		Bilgisyara Sahip Olma	
	İstatistik	p	İstatistik	p
Düğüm 1 - Erkek-Kadın	110.074	.000	93.459	.000
Düğüm 2 - Erkek (Yes-No)	-	-	53.449	.092
Düğüm 3 - Kadın (Yes-No)	-	-	75.446	.001

Tablo 1'e göre kız ve erkek grupları, bilgisayar sahibi olan ve olmayan kız katılımcılar arasında DMF tespit edilmiştir. Erkek katılımcılar için bilgisayar sahibi olma durumunun, DMF'ye yol açmadığı belirlenmiştir. Bu bulgulara ait sonuçlar grafik 1'de sunulmuştur.



Grafik 1. Cinsiyet ve bilgisayar/tablet sahipliği değişkenine ilişkin rasch ağacı

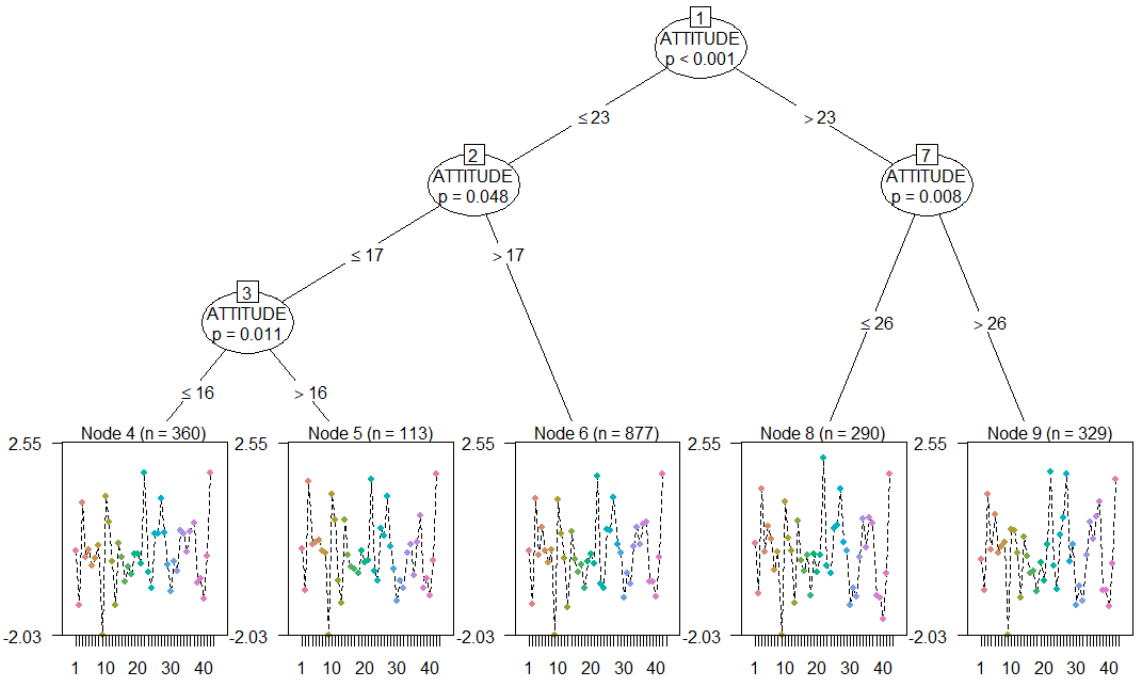
Matematiğe karşı tutum değişkeni için elde edilen bulgular ve rasch ağacı sırasıyla Tablo 2 ve Grafik 2'de paylaşılmıştır.

Tablo 2

Matematiğe Karşı Tutum Değişkenine İlişkin Bulgular

	İstatistik	p
Düğüm 1. (< 23) - (> 23)	118.323	.000
Düğüm 2. (<17) - (17 , 23)	73.621	.048
Düğüm 3. (< 16) - (16 , 17)	80.205	.011
Düğüm 7. (23 ,26) - (> 26)	81.548	.008

Tablo 2'ye göre sırasıyla, 16, 17, 23 ve 26 tutum puanlarının alt ve üstünde kalan gruplar arasında DMF olduğu tespit edilmiştir.



Grafik 2. Matematiğe karşı tutum değişkeni rasch ağacı

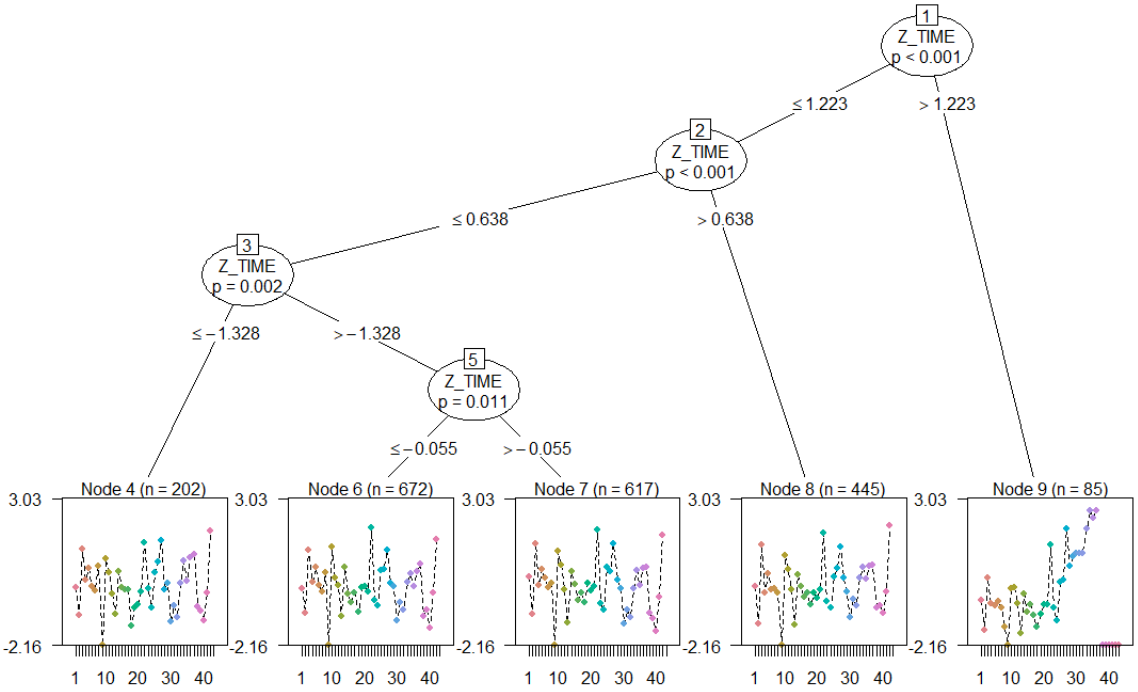
Ortalama cevaplama süresi standart değerleri değişkeni için elde edilen bulgular ve rasch ağacı sırasıyla Tablo 3 ve Grafik 3'de sunulmuştur.

Tablo 3

Ortalama Cevap Süresi Standart Değerleri Değişkenine İlişkin Bulgular

	İstatistik	p
Düğüm 1 - (> 1.223)- (< 1.223)	332.342	.000
Düğüm 2 - (< 0.638) - (0.638 , 1.223)	113.317	.000
Düğüm 3 - (< -1.328) - (-1.328 , 0.638)	87.675	.002
Düğüm 5 - (-1.328 ,-0.055) - (-0.055 , 0.638)	80.461	.000

Tablo 3'e göre sırasıyla, -1.328, -0.055, 0.638, 1.223 standart değerleri alt ve üstünde kalan gruplar arasında DMF olduğu sonucuna ulaşılmıştır.



Grafik 3. Ortalama cevaplama süresi standart değerleri değişkeni rasch ağacı

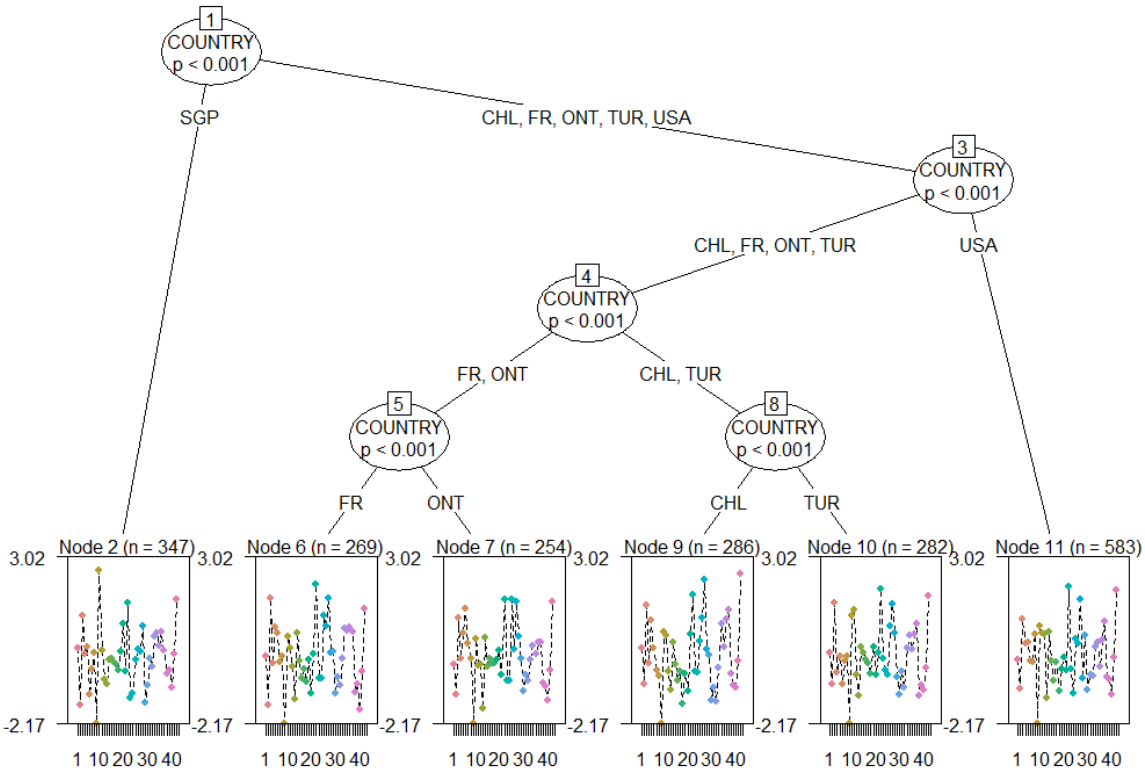
Farklı ülke gruplarına ilişkin DMF tespitine ait istatistiksel veriler Tablo 4'de verilmiştir.

Tablo 4

Ülke Değişkenine İlişkin Bulgular

	İstatistik	p
Düğüm 1. SGP - Others	1,724.674	.000
Düğüm 3. USA - (FRA, ONT, TUR, CHL)	1,106.239	.000
Düğüm 4. (FR, ONT) - (CHL - TUR)	650.785	.000
Düğüm 5. FR - ONT	133.191	.000
Düğüm 8. CHL - TUR	166.852	.000

Tablo 4'deki sonuçlara göre hangi ülke grupları arasında DMF varlığı olduğunun görsel olarak sunumu Grafik 4'de paylaşılmıştır.



Grafik 4. Ülke değişkeni için Rasch ağacı

Kaynaklar

Camilli, G. (2006). Test Fairness. In R.L. Brennan (Ed.), *Educational measurement* (4th edi., pp. 221-256).Westport, CT: Praeger.

Doğan, N. ve Öğretmen, T. (2005). Test ve madde yanlılığı. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 5*(1), 89-103.

Hambleton, R. K., and Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel Haenszel methods. *Applied Measurement in Education, 2*(4), 313-334.

Strobl C., Kopf J., and Zeileis A. (2015). Rasch trees: A new method for detecting differential item functioning in the rasch model. *Psychometrika, 80*(2), 289-316. https://doi.org/10.1007/s11336-013-9388-3).

Thorndike, R. L., and Hagen, E. (1961). *Measurement and evaluation in psychology and education* (2nd ed.). Wiley.

Yalçın, S. (2018) 21. yüzyıl becerileri ve bu becerilerin ölçülmesinde kullanılan araçlar ve yaklaşımlar. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 51*(1), 183-201. https://doi.org/10.30964/auebfd.405860

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential ıtem functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) ıtem scores.* Directorate of Human Resources Research and Evaluation, Department of National Defense.

Küçük örneklem büyüklüklerinde açımlayıcı faktör analizi üzerine bir tanıtım ve tartışma

Alperen Yandı

Giriş

Ölçme araçları, insanların katılımcı olduğu, farklı alanlarda gerçekleştirilen araştırmaların birçoğundan örtük özelliklerin keşfedilmesi için kullanılmaktadır. Ölçme araçları, örtük özelliklere ilişkin göstergeler olan, gözlenen değişkenler olarak da adlandırılan maddeler bütününden oluşmaktadır. Ölçme aracını oluşturan maddeler istatistiksel teknikler ve teorik incelemeler sonrasında psikometrik açıdan incelenerek araştırmacıların kullanımına sunulmaktadır.

Ölçme araçlarının geliştirime sürecinde uygulanan istatistiki incelemelerden birisi faktör analizidir. Faktör analizi, açımlayıcı ve doğrulayıcı faktör analizi olarak iki ana başlık altında incelenebilir. Bu başlıklardan açımlayıcı faktör analizi, bireylerden toplanan veriler üzerinde gerçekleştirilen analizlerle, maddelerin oluşturduğu yapının keşfedilmesi için işe koşulmaktadır. Doğrulayıcı faktör analizi ise daha önceden keşfi yapılmış olan bir yapının model-veri uyumu bağlamında incelenmesine olanak tanımaktadır. Ölçme aracı türlerinden ölçeklerin geliştirme sürecinde açımlayıcı faktör analizi geçerlik kanıtı elde edilmesi için kullanılan önemli isatistiksel teknikler arasındadır. Açımlayıcı faktör analizi, teorik olarak bir temele dayandırılarak, bireylerden alınan yazılar, uzman görüşleri ve alanyazındaki araştırmalar ışığında yazımı gerçekleştirilen maddelerin, kendi aralarındaki kovaryans ve korelasyon matrislerine dayalı olarak faktörler oluşturulması işlemi olarak tanımlanabilir. Açımlayıcı faktör analizi sonucunda elde edilen istatistiksel sonuçlar, araştırmacıların faktörlere ilişkin fikir sahibi olmalarında, maddelerin örtük özellikle olan ilişkilerine yönelik istatistiki kanıtlar elde etmelerine imkân tanımaktadır. Açımlayıcı faktör analizi ile gözlenen değişkenler ile örtük değişkenlerin karşılıklı olarak birbirileri üzerinde açıklayabildikleri varyanslara dair bilgi elde edilebilmektedir.

Açımlayıcı faktör analizi sürecinde kestirilen parametrelere ilişkin güvenirliği etkileyen belirli varsayımlar vardır. Normallik, doğrusallık, çoklu bağlantı varsayımı bunlara örnek olarak verilebilir. Analiz sürecindeki bir diğer önemli hususta örneklem büyüklüğüdür. Alanyazında yer alan kaynaklarda AFA sürecinde örneklem büyüklüğüne ilişkin farklı görüşler mevcuttur. Bu görüşlerde genel olarak madde sayısı ile örneklem büyüklüğü arasında ilişki kurulmaktadır. Bu görüşler arasında madde sayısının en az beş katı (Child, 2006; Doğan ve Başokçu, 2010), madde sayısının en az on katı (Kline, 2005)

şeklinde olması gerektiğini belirten görüşler mevcuttur. Buna ek olarak örneklem büyüklüğünü madde sayısından bağımsız şekilde ele alarak, en az 150-200 (Guilford, 1954; Tabachnick ve Fidell, 2006; Kline, 2005) en az 300 katılımcı (Comrey ve Lee, 1992; akt. Tabachnick ve Fidell, 2006) olması gerektiğini belirten kaynaklarda mevcuttur. Örneklem büyüklüğüne ilişkin farklı görüşler değerlendirildiğinde araştırmacıların madde sayısında bağımsız olarak belirli bir katılımcı sayısına ulaşması gerektiği önemli bir gereklilik olarak dikkati çekmektedir. Açımlayıcı faktör analizi sürecinde parametre kestirimi ve faktör geri kazanımı (recovery) süreçlerindeki örneklem büyüklüğünden etkilenme durumları bu gerekliliği daha da önemli hale getirmektedir (MacCallum ve diğ., 1999; Preacher ve MacCallum, 2002).

AFA sürecinde kestirim yöntemi olarak sıklıkla maksimum olabilirlik (ML) yöntemleri tercih edilmektedir (Fabrigar ve diğ., 1999). Bununla birlikte, bu kestirim yönteminin güvenilir ve kararlı çözümler elde etmek için büyük örneklemlere ihtiyaç duyulmaktadır (MacCallum ve diğ., 1999). Veri toplama süreçlerinde yukarıda belirtilen örneklem büyüklüklerine ulaşmanın zorlukları düşünüldüğünde, küçük örneklemlerde AFA sürecinin gerçekleştirilebilmesi için alternatif kestirim yöntemleri üzerinde araştırmalar gerçekleştirilmektedir (Bentler ve de Leeuw, 2011). Ağırlıklandırılmamış en küçük kareler ortalaması (ULS), ML için küçük örneklemlerde iyi bir alternatif olarak önerilmektedir (Budaev, 2010). Bu iki kestirim yöntemine ek olarak alanyazında yer alan araştırmalar incelendiğinde küçük örneklem büyüklüğü olan durumlarda AFA gerçekleştirilebilmesi için iki farklı alternatif yaklaşım ortaya konduğu belirlenmiştir. Bu yaklaşımlar (1) Düzenlenmiş AFA (Regularized Exploratory Factor Analysis-REFA) (Jung ve Lee, 2011), (2) Genelleştirilmiş AFA (Genaralized Exploratory Factor Analysis GEFA) (Trendafilov ve Unkel, 2011) şeklindedir. Yapılan araştırmalar Düzenlenmiş AFA analizi ile küçük örneklemlerde daha iyi sonuçlar elde edildiğini göstermektedir (Jung ve Lee, 2011; Trendafilov ve Unkel, 2011; Jung, 2013). Küçük örneklem büyüklüklerinde ağırlıklandırılmamış en küçük kareler yöntemine kıyasla düzenlenmiş ve genelleştirilmiş AFA süreçlerinin parametre kestirimi açısından belirli farklılıkları vardır. Jung (2013) tarafından gerçekleştirilen araştırmada ULS, REFA ve GEFA yaklaşımları karşılaştırılmış ve küçük örneklemlerde ULS ve REFA yaklaşımlarının faktör geri kazanımı bakımından anlamlı şekilde iyi sonuçlar verdiği rapor edilmiştir.

Yapılan bu araştırmada küçük örneklemlerde AFA sürecinde araştırmacılara REFA ve ULS yaklaşımları ile gerçekleştirilen analiz sonuçlarına ilişkin bilgi sunulması ve bu iki yaklaşımının tanıtılması amaçlanmıştır. Buna ek olarak bu analiz yaklaşımlarının özellikle belirlenen örneklemin, evreni temsil edebilme bağlamında görüş alınması planlanmaktadır. İstatistiki açıdan küçük örneklemlerde AFA sürecinin gerçekleştirilme durumunu sağlanabilir olsa da bu durumun getirebileceği sınırlılıkların tartışılmasının sağlanması önem arz etmektedir. Bu araştırmanın amaçlarından birisi de farklı görüşlerin alınarak bu durumun örneklemin temsil gücü açısından değerlendirilmesidir.

Yöntem

Araştırma tarama araştırması olarak planlanmıştır. Varolan bir durumu herhangi bir değişkene araştırmacı müdahalesi olmadan olduğu haliyle ortaya koyma amacı olması (Karasar, 2018) araştırma modelinin tarama olarak nitelendirilmesinin en önemli nedenidir. Araştırma kapsamında MPlus programı 8.3 sürümü yardımıyla üretilmiş veriler kullanılmıştır. Yapılan bu araştırma başlangıç niteliğinde olduğunda ve örneklem büyüklüğü koşulu özelinde yöntemlerin karşılaştırılması ve tartışılmasını amaçladığından sadece örneklem büyüklüğü değişkeni farklılaştırılarak veri setleri üretilmiştir. Veri setlerine ait koşullara ilişkin bilgiler Tablo 1'de sunulmuştur.

Tablo 1

Üretilen Verilere İlişkin Koşullara Ait Bilgiler

Koşul adı	Özellik	Replikasyon ve koşul sayısı
Puanlama	5'li likert	Toplam 3 koşul altında
Boyut sayısı	2	
Madde sayısı	3x2 = 6	20 replikasyon
Faktör yükleri	0.80	
Örneklem büyüklüğü (n)	10 – 50 – 200	
Faktörler arası korelasyon	0.30	

Veri setleri üzerinde ağırlıklandırılmamış en küçük kareler (ULS) ve düzenlenmiş açımlayıcı faktör analizi (REFA) yaklaşımları ile analizler gerçekleştirilmiştir. REFA analiz sürecindeki parametre tahminlerini ULS yaklaşımına kıyasla daha basitleştirmektedir. REFA, gerçek benzersiz varyansların, geçici tahminleriyle orantılı olduğunu varsayar. Yani REFA, "orantı"yı temsil eden tek bir parametrenin tahmin edilmesini içerir. Benzersiz varyanslar, geçici miktarlarının bir oran tahmini ile çarpılmasıyla tahmin edilir. Benzersiz varyans tahminleri elde edildikten sonra faktör yükleri kapalı bir formülle hesaplanır (Jöreskog, 1977). Parametre kestirimini basitleştirmesine ek olarak REFA parametre tahminin parça parça yaklaşımını esas almaktadır. Örneklem büyüklüğünün küçük olduğu durumlarda yinelemeli sayısal yöntemlerin uygulanabilir olsa da hesaplama süreci zor olabilmektedir. Bu nedenle bu tür durumlarda parametre tahminlerinin birkaç alt kümeye bölünerek yapılması önerilmektedir (Bollen, 1996). REFA parametre tahmin sürecini bu doğrultuda gerçekleştirmektedir. REFA'nın sahip olduğu bu avantajlar karşısında, belirli sınırlılıkları da mevcuttur. Bunlardan ilki REFA sonucunda model veri uyumuna ilişkin herhangi bir uyum iyiliği testi yapılamamasıdır. ULS, model uyum iyiliğini değerlendirmek ve ayrıca istatistiksel olarak modelin anlamlılığını değerlendirmek için imkân tanımaktadır (Browne, 1984; Akt: Jung, 2013). Bir diğer önemli sınırlılık ise REFA uygulamaları için yazılım programlarının olmamasıdır. ULS yaklaşımı ile SPSS, Mplus gibi kapalı kaynak kodlu programlarla gerçekleştirilmesi mümkündür. Buna ek olarak R açık kaynak kodlu platform aracılığıyla da bu yaklaşımla analizler yapılabilmektedir. Ancak REFA için program çeşitliliği bu düzeyde değildir. Araştırma kapsamında yapılan analizler R platformu üzerinde yayınlanmış fungible paketindeki fareg ve fals fonksiyoları kullanılarak gerçekleştirilmiştir.

Sonuçlar

Araştırma kapsamında yapılan analizler sonrasında karesel ortalama hata (mean square error - MSE) değerleri ve göreceli yanlılık değerlerinin (relative bias-RB) incelenmiştir. Karesel ortalama hata bir parametre ile tahmini arasındaki ortalama kare farkı olarak tanımlanabilir. Karesel ortalama parametre değeri tahmin edilen ve gerçek parametre arasındaki ortalama kare farkıdır. Bu değerin küçük olması tahmin edilen parametrenin, gerçek parametreye yakın olduğunu gösterir. MSE hesaplama süreci aşağıdaki denklemde verilmiştir:

$$MSE(\hat{\theta}j) = E[(\hat{\theta}j - \theta j)^2] = E[(\hat{\theta}j - E(\hat{\theta}j))^2] + (E(\hat{\theta}j) - \theta j)^2$$

Denklem incelendiğinde bir tahminin karesel ortalama hatası, varyansının ve kare yanlılığının toplamıdır. Bu değer tahminin hem yanlılığını hem de değişkenliğini hesaba katmaktadır (Mood, Graybill & Boes, 1974). Diğer bir kriter olan göreceli yanlılık, mutlak değerde 10'dan büyük olduğu durumlar kabul edilemez düzey olarak yorumlanabilir (Lei, 2009). Bu iki inceleme kriterine göre elde edilen sonuçlar Tablo 2'de verilmiştir.

Tablo 2

Gerçekleştirilen Faktör Analizine İlişkin Kriter Değerlere Ait Sonuçlar

n	REFA (Düzenlenmiş AFA)		ULS (Ağırlıklandırılmamış En küçük Kareler Yöntemi)	
	RB (Görecelik Yanlılık)	MSE (Karesel Ortalama Hata)	RB (Görecelik Yanlılık)	MSE (Karesel Ortalama Hata)
10	7.900	0.021	7.502	0.031
50	4.705	0.005	4.440	0.009
200	5.282	0.003	5.222	0.004

Tablo 2.'de sunulan sonuçlar incelendiğinde karesel ortalama değerlerinin her iki kestirim yöntemi içinde yakın olduğu görülmüştür. Buna ek olarak en düşük karesel ortalama hata değerleri 200 örneklem büyüklüğündeki veri setleri; en yüksek hata değerleri ise 10 örneklem büyüklüğüne sahip veri setleri için için elde edilmiştir. Göreceli yanlılık değerleri incelendiğinde ise tüm koşullar altında elde edilen değerlerin kabul edilebilir (<10) aralıkta olduğu sonucuna ulaşılmıştır. Buna ek olarak 50 örneklem büyüklüğüne sahip veri setlerinde yanlılık değerlerinin en düşük olması dikkat çekmektedir. Elde edilen bu sonuçlar ve alınan görüşler doğrultusunda ağırlıklandırılmamış en küçük kareler ortalaması yönteminin küçük örneklem büyüklüklerinde güçlü bir yöntem olduğu ancak düzenlenmiş AFA kestirimlerinin bu yönteme iyi bir alternatif olduğu ileri sürülebilir. Ortaya konan bu sonuçlar alanyazında bulunan ve REFA üzerinde yapılmış araştırmalarla benzerlik göstermektedir (Jung, 2013; Jung & Lee, 2011). Bu bağlamda yapılacak olan ve koşullara farklı değişkenlerin eklenerek yapılacağı devam araştırmalarında REFA özelliklerinin araştırmacılara daha ayrıntılı şekilde sunulması ve sosyal bilimler alanına kazandırılması planlanmaktadır. Bu sayede özellikle ölçek geliştirme çalışmaları sürecinde gerçekleştirilmesi önerilen ve iş yükü, zaman kaygısı, eleman eksikliği gibi nedenlerle uygun

örnekleme yönteminin yanlış yorumlanarak kullanılması sonrasında belirlenen gelişigüzel örneklemler üzerinde yapılan deneme uygulamalarının daha doğru bir biçimde gerçekleştirilebilmesine katkı sunulacağı düşünülmektedir. Bu katkı, araştırmacıların doğru örnekleme yöntemi seçerek, zaman ve iş yükü kaygılarından kurtulması aracılığıyla sağlanabilir. Küçük örneklemler yapılacak analizlerde yöntemlerin kullanışlılığı ve özelliklerinin bilinirliğinin arttırılması bu nedenle önemli görülmektedir. Buradan hareketle devam araştırmalarında REFA ve MLS'nin karşılaştırılması sürecinde faktör yük dağılımı, faktör madde sayısı dağılımı, faktörler arası korelasyon, boyut sayısı, puanlama aralığı değişkenlerinin farklılaştırılarak simülatif veriler üzerinde çalışılması önerilebilir. Buna ek olarak bu araştırmada kullanılan ve diğer kestirim yöntemlerinin R platformu aracılığıyla kullanışlı bir arayüze sahip şekilde araştırmacılara sunulması sağlanabilir.

Kaynaklar

Bentler, P. M., & de Leeuw, J. (2011). Factor analysis via components analysis. *Psychometrika*, *76*(3), 461-470. https://doi.org/10.1007/s11336-011-9217-5

Bollen, K. A. (1996). An alternative two stage least squares (2SLS) estimator for latent variable equations. *Psychometrika*, *61*(1), 109-121. https://doi.org/10.1007/BF02296961

Budaev, S. V. (2010). Using principal components and factor analysis in animal behaviour research: caveats and guidelines. *Ethology*, *116*(5), 472-480. https://doi.org/10.1111/j.1439-0310.2010.01758.x

Child, D. (2006). *The essentials of factor analysis*. A&C Black.

Doğan, N. ve Başokçu, T. O. (2010). İstatistik tutum ölçeği için uygulanan faktör analizi ve aşamalı kümeleme analizi sonuçlarının karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *1*(2), 65-71. https://dergipark.org.tr/tr/download/article-file/65985

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., and Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, *4*(3), 272-299. https://doi.org/10.1037/1082-989X.4.3.272

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). McGraw-Hill.

Jöreskog, K. G. (1977). Factor analysis by least squares and maximum likelihood methods. In Statistical Methods for Digital Computers. http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-22118

Jung, S. (2013). Exploratory factor analysis with small sample sizes: A comparison of three approaches. *Behavioural processes*, *97*, 90-95. https://doi.org/10.1016/j.beproc.2012.11.016

Jung, S., & Lee, S. (2011). Exploratory factor analysis for small samples. *Behavior research methods*, *43*(3), 701-709. https://doi.org/10.3758/s13428-011-0077-9

Karasar, N. (2018). *Bilimsel araştırma yöntemi* (33. baskı). Nobel Akademik Yayıncılık.

Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. Sage.

Lei, P. W. (2009). Evaluating estimation methods for ordinal data in structural equation modeling. *Quality and Quantity*, *43*(3), 495-507. https://doi.org/10.1007/s11135-007-9133-z

MacCallum, R. C., Widaman, K. F., Zhang, S., and Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84-99. https://doi.org/10.1037/1082-989X.4.1.84

Mood, A.M., Graybill, F.A., and Boes, D.C. (1974). *Introduction to the Theory of Statistics*. McGraw-Hill.

Preacher, K. J., and MacCallum, R. C. (2002). Exploratory factor analysis in behavior genetics research: Factor recovery with small sample sizes. *Behavior genetics*, *32*(2), 153-161. https://doi.org/10.1023/A:1015210025234

Tabachnick, B. G., and Fidel, L. S. (2006). *Using Multivariate Statistics* (5th ed.). MA, Allyn, Bacon, Inc.

Trendafilov, N. T., and Unkel, S. (2011). Exploratory factor analysis of data matrices with more variables than observations. *Journal of Computational and Graphical Statistics*, *20*(4), 874-891. https://doi.org/10.1198/jcgs.2011.09211

Veri madenciliği teknikleri kullanılarak 8. sınıf öğrencilerinin matematik başarılarını sınıflandırma analizi çalışması: 2019 TIMMS Türkiye örneği

Simge Ceylan ve Tuncay Öğretmen

Anahtar kelimeler: Eğitimde veri madenciliği, WEKA, TIMMS, sınıflandırma analizi, makine öğrenme algoritmaları

Giriş

Her geçen gün yapılan çalışmalarla birlikte birçok veri toplanmaktadır. Bu verilerin çokluğu veri madenciliği ve makine öğrenmesi gibi çalışmalara yol açmıştır. Makine öğrenmesi eğitim, ekonomi, tıp gibi farklı alanlarda sıklıkla kullanılmaktadır. Eğitim alanında makine öğrenmesi ile veri madenciliği uygulaması Eğitim Veri Madenciliği (EVM) olarak adlandırılır. EVM eğitim alanında uzmanlara karmaşık veri kümelerini anlamlandırma imkânı sunmaktadır. Bu bağlamda, Uluslararası Eğitim Başarılarını Değerlendirme Birliği (International Association for the Evaluation of Educational Achievement [IEA]), katılımcı ülkelerin eğitim sistemlerinin politikalarının ve uygulamalarının etkilerini izlemek amacıyla güvenilir veri setlerini elde etmek ve raporlamak için birçok karşılaştırmalı çalışma yürütmektedir. Yürütülen çalışmalardan biri olan Uluslararası Matematik ve Fen Eğilimleri Araştırması (Trends in International Mathematics and Science Study [TIMSS]), öğrenci başarısını uluslararası düzeyde değerlendiren en büyük IEA projesidir ve 60'tan fazla ülke bu projeye katılmaktadır. Proje her dört yılda bir dördüncü ve sekizinci sınıf matematik ve fen öğrencileriyle gerçekleştirilmektedir. TIMSS, yalnızca eğitim politikalarının etkileri ve uygulamaları ile ilgili güvenilir bilgi sağlamakla kalmaz, aynı zamanda katılımcı ülkelerin öğrenci başarısı açısından, katılımcı ülkelerin sonuçları arasında karşılaştırma yapmalarını sağlar(Mullis et al., n.d.). Çalışmada kullanılan TIMMS 2019 Türkiye 8.sınıf verilerinin analizinde EVM'nin tercih edilme sebebi normallik, değişkenler arası doğrusallık ve değişkenlerin homojenliği gibi varsayımları sağlama zorunluluğu olmamasıdır (Han ve diğ, 2012).

EVM kapsamında sınıflandırma, önemli veri sınıflarını tanımlayan modelleri çıkaran bir veri analizi biçimidir. Sınıflandırıcılar olarak adlandırılan bu tür modeller, kategorik (ayrık, sırasız) sınıf etiketlerini tahmin eder. Bu tür analizler, genel olarak verileri daha iyi anlamamıza yardımcı olabilir. Araştırmacılar tarafından makine öğrenmelerinde, örüntü tanıma ve istatistik alanlarında birçok sınıflandırma yöntemi önerilmiştir. Algoritmaların çoğu, genellikle küçük bir veri boyutu varsayılarak

bellekte yerleşiktir. Son zamanlardaki veri madenciliği araştırmaları, büyük miktardaki yerleşik veriyi işleyebilen ölçeklenebilir sınıflandırma ve tahmin teknikleri geliştirerek bu tür çalışmalar üzerine inşa edilmiştir (Han Jiawei et al., n.d.) TIMMS gibi büyük örneklem içeren verilerin analizinde uygun EVM yöntemini tespit etmek araştırma açısından önemlidir. Bu bağlamda araştırmanın amacı TIMMS Türkiye 8.sınıf öğrencilerinin matematik başarılarının sınıflandırmasında hangi EVM yönteminin daha uygun olduğunu belirlemektir. Bunun için literatürde en çok tercih edilen K-En Yakın Komşu Algoritması, Naive Bayes Algoritması, Yapay Sinir Ağları, Karar Ağacı Algoritmaları ve Lojistik Regresyon Analizi tercih edilmiştir.

Makine Öğrenme Algoritmaları

Veri madenciliği yöntemleri temel olarak tahmin, sınıflandırma, ku¨meleme ve birliktelik kuralları şeklinde 4 bölümden oluşur. Tahmin yöntemleri olarak Regresyon analizi, Bayes ağları, LR, KA ve YSA; sınıflandırma için k-NN, NB, YSA, KA, DVM ve LR; ku¨meleme için k-means, modele dayalı ku¨meleme, tam bağlantı ku¨meleme; birliktelik kuralları için Apriori, Carma, Sequence, GRI, Eclat, FP-Growth gibi yöntemler kullanılır.(Filiz ve Öz, 2019) Çalışmada tahminleme yöntemine dayalı olan algoritmalardan Naive Bayes ve Lojistik regresyon, Sınıflandırma yöntemine dayalı algoritmalardan ise En Yakın Komşu (IBk), Karar Ağacı (j48), Yapay Sinir Ağları yaklaşımlarından da Multilayer Perceptron kullanılmıştır.

Tahminleme Yöntemine Dayalı Algoritmalar

Makine öğrenmelerinde farklı tahmin yöntemleri kullanılır bunlardan en çok bilinenlerine örnek olarak regresyon analizi ve Bayes ağları ve lojistik regresyon, karar ağacı verilebilir.

Lineer Regresyon; sonuç veya sınıf, sayısal (numeric) olduğunda ve tüm nitelikler sayısal olduğunda, doğrusal regresyon dikkate alınması gereken doğal bir tekniktir. Bu istatistikte temel bir yöntemdir. Buradaki temel fikir; sınıfları, önceden belirlenmiş ağırlıklarla niteliklerin doğrusal bir kombinasyonu olarak ifade etmektir. Lojistik regresyon ise 0 ve 1 gibi dönüştürülmüş bir hedef değişkene dayalı doğrusal bir model oluşturan regresyon modeli türüdür (Witten ve diğ., 2017).

Naive Bayes; en çok tercih edilen Bayes ağlarındandır. Temeli matematikteki olasılık teorisine dayanır. Naive Bayes algoritması, olası sınıflandırmaların her birinin olasılığını sırayla hesaplamak için kullanabilecek tek bir formülde, önceki olasılık (prior) ve koşullu olasılıkları (conditional) birleştirmenin bir yolunu sunar ve sonra en büyük değere sahip sınıflandırmayı seçer (Max Bramer, 2007).

Sınıflandırma Yöntemine Dayalı Algoritmalar

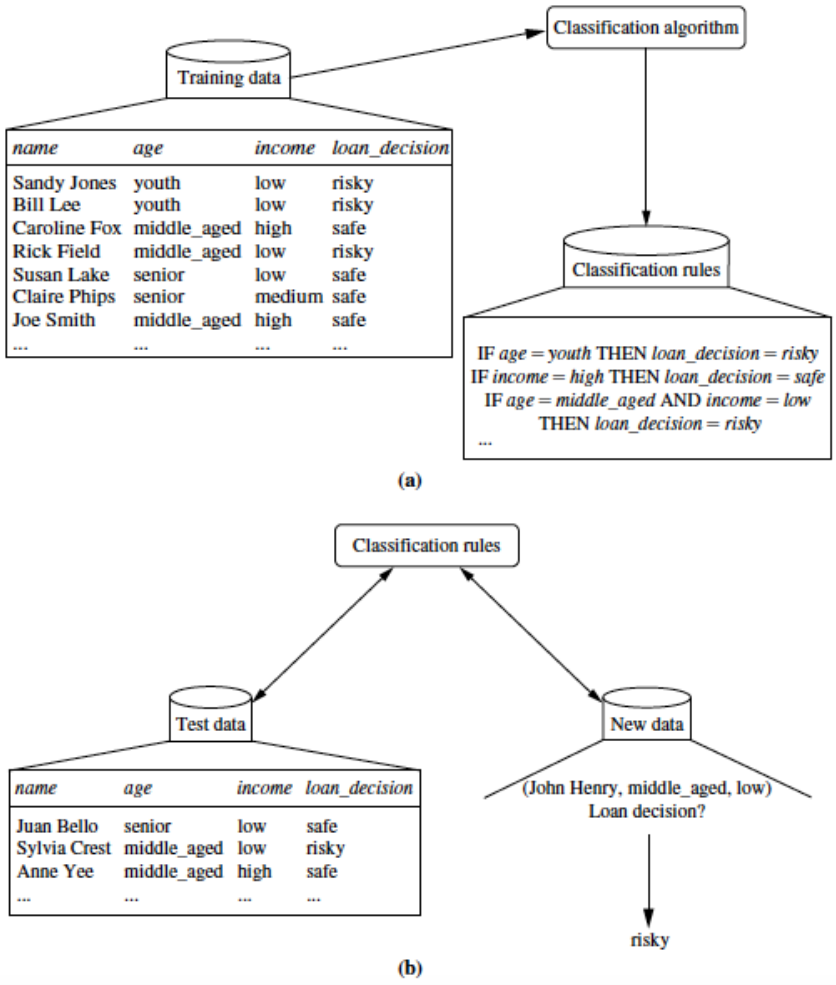
Veri sınıflandırma, bir öğrenme adımından (bir sınıflandırma modelinin oluşturulduğu) ve bir sınıflandırma adımından (modelin, verilen veriler için sınıf etiketlerini tahmin etmek için kullanıldığı) oluşan iki aşamalı bir süreçtir. İlk adımda, önceden belirlenmiş veri sınıfları veya kavramları kümesini tanımlayan bir sınıflandırıcı oluşturulur. Bu, bir sınıflandırma algoritmasının, veri tabanı demetlerinden ve bunlarla ilişkili sınıf etiketlerinden oluşan bir eğitim setini analiz ederek veya "öğrenerek"

sınıflandırıcıyı oluşturduğu öğrenme adımıdır (veya eğitim (training) aşamasıdır). Sınıf etiketi özelliği ayrık değerlidir ve sırasızdır. Her değerin bir kategori veya sınıf olarak hizmet etmesi bakımından kategoriktir (veya nominal). Başka bir deyişle sınıflandırma sürecinin bu ilk adımı, belirli bir veri kümesinin ilişkili sınıf etiketini (y) tahmin edebilen bir eşleme ya da fonksiyonu öğrenmesi olarak adlandırılabilir. (Y= f(x)). Tipik olarak bu eşleme ya da sınıflandırma kuralları, karar ağaçları veya matematiksel formüller şeklinde temsil edilir. Şekil1 de sınıflandırma aşamaları verilmiştir (Han ve diğ, 2012).

En Yakın Komşu sınıflandırması; kategorik özniteliklerle başa çıkmak için değiştirilebilmesine rağmen, esas olarak tüm öznitelik değerleri sürekli olduğunda kullanılır. Temel anlamda bu algoritmayla sınıfını bilmediğimiz örneğe en yakın sınıflandırmayı kullanarak, sınıflandırmasını tahmin etmeye çalışmaktır (Max Bramer, 2007).

Karar ağaçları; veri madenciliğinde kullanılan sınıflandırma tekniklerinden bir başkasıdır. Karar ağaçlarında önce bir karar ağacı oluşturulur ardından karar ağacından u¨retilen kurallar ile veri tabanında bulunan değişkenler sınıflandırılmaktadır (Silahtaroğlu, 2008). Karar ağaçlarında ID3, C4.5, Sliq, Sprint, CART, REP Tree, Random Forest, Logistic Model Tree gibi birçok algoritma kullanılabilmektedir. (Thornton ve diğ., 2013)Veriler önce bu karar ağacı algoritmalarından birine gönderilir. Algoritma bu verileri işler ve bir karar ağacı oluşturur. Oluşturulan bu karar ağacı sınıfı bilinmeyen veriler u¨zerine uygulanarak bu verilerin sınıflarının tespit edilmesi sağlanır (Aksu ve Karaman, 2017).

Yapay sinir ağları yaklaşımı; Genellikle "sinir ağı" (NN) (Lateef ve Adenubi, 2013) olarak adlandırılan bir yapay sinir ağı (YSA), biyolojik sinir ağlarına dayanan bir hesaplama modelidir, başka bir deyişle, insan sinir sisteminin bir temsilidir(Refaat, 2007). Birbirine bağlı bir yapay nöron grubundan oluşur ve bilgiyi hesaplamaya bağlantıcı bir yaklaşım kullanarak işler. Pratik anlamda sinir ağları, doğrusal olmayan istatistiksel veri modelleme araçlarıdır (Kamruzzaman ve Sarkar, 2011). Yapay sinir ağları yaklaşımlarının en popüler biçimi çok katmanlı algılayıcıdır (Multi Layer Perceptron MLP). MLP, özellikle vektör öznitelik değerleri tarafından belirlenen örneği bir veya daha fazla sınıfa yerleştiren bir sınıflandırma fonksiyonuna ihtiyaç duyulduğunda uygundur (Sebastian ve Puthiyidam, 2015).



Şekil 1: Veri sınıflandırma süreci: (a) Öğrenme: Eğitim verileri bir sınıflandırma algoritması ile analiz edilir. Burada, sınıf etiketi özniteliği kredi kararıdır ve öğrenilen model veya sınıflandırıcı, sınıflandırma kuralları biçiminde temsil edilir. (b) Sınıflandırma: Test verileri, sınıflandırma kurallarının doğruluğunu tahmin etmek için kullanılır. Doğruluğun kabul edilebilir olduğu düşünülürse, kurallar yeni veri gruplarının sınıflandırılmasına uygulanabilir(Han Jiawei et al., n.d.).

Yöntem

TIMSS, matematik ve fen başarılarını değerlendirmek için 4 yılda bir yürütülmektedir ve hedef kitlesi 4. ve 8. sınıf öğrencileridir. TIMSS 2019'ta 39 katılımcı ülke 8. sınıf değerlendirmesine katıldı. TIMSS örnekleme prosedürü, iki aşamalı bir tabakalı küme örnekleme tasarımıdır. İlk aşama, okul örneklemini orantılı olarak seçmek, ikincisi ise örneklenen okullar arasından rastgele sınıfları seçmektir(Laroche et al., n.d.). Bu araştırma, kişilerin bildirdiği anketlere dayalı bir çalışmadır. Anketler, öğrencilerin kendileri hakkındaki görüşlere, fen ve matematik derslerine karşı tutumlarına, ev ve okul yaşantılarına, evde eğitim ve öğretim ile ilgili araç ve eğitimsel kaynakların durumuna yönelik maddelerden oluşmaktadır.

Bu çalışmada, öğrenci performanslarına ilişkin veriler, TIMSS'ın resmi sitesindeki (https://timssandpirls.bc.edu/) veri dosyalarından internet aracılığı ile elde edilmiştir. Araştırmada kullanılan TIMMS 2019 veri setinde 9 adet ölçek bulunmaktadır. Bu ölçekler; ev eğitim kaynakları (BSBGHER), öğrencilerin okula ait olma hissi (BSBGSSB), öğrenci zorbalığı (BSBGSB), öğrencinin matematik öğrenmeyi sevmesi (BSBGSLM), tahmin için matematik başarısı çok düşük (BSDMLOWP) ve matematik ödevinde haftalık harcanan zaman (BSDMWKHW)'dir. Çalışma için alanda yapılan çalışmalar göz önünde bulundurularak ev eğitim kaynakları ve öğrencinin matematikte özgüveni seçilmiştir (Bulut ve diğ., 2017; Filiz ve Öz, 2019; Polat, 2019). Matematikte ev eğitim kaynakları ölçeğini oluşturan sorular ve cevap alternatifleri Tablo1'de verilmiştir.

Tablo 1

Matematikte Ev Eğitim Kaynakları Ölçeği

Soru kodu	Soru	Cevap
BSBG04	Evdeki kitap sayısı	0-10 11-25 26-100 101-200 200'den fazla
BSDG06S	Ev çalışmalarındaki desteklerin sayısı	Yok İnternet bağlantısı ya da kendi odası olma ikiside
BSDGEDUP	Ebeveynlerin eğitim düzeyi	İlkokul, tamamlanmamış ortaokul ya da hiç okula gitmemiş Ortaokul Lise Lise sonrası (yükseköğrenim olmayan) Üniversite ya da üzeri

Tablo 2

Matematikte Öğrencinin Özgüveni Ölçeği

Soru kodu	Soru	Cevap
BSBM19A	Matematikte genellikle başarılıyımdır	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM19B	Birçok sınıf arkadaşıma göre matematikte daha çok zorlanırım	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM19C	Matematik güçlü yanlarımdan biri değildir	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum

(devam ediyor)

Tablo 2 *(devam)*

Soru kodu	Soru	Cevap
BSBM19D	Matematikteki şeyleri kolaylıkla öğrenirim	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM19E	Matematik beni tedirgin eder	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM19F	Matematikte zor problemleri çözmekte başarılıyım	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM19G	Öğretmenim matematikte iyi olduğumu söyler	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM19H	Matematik benim için diğer konulardan daha zordur	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM19I	Matematik beni şaşırtır	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum

Araştırmanın ilk aşamasında TIMMS 2019 Türkiye veri setinde seçilen bağımsız değişkenler olan Ev eğitim kaynakları (BSBGHER) ve matematikte öğrencinin özgüveni (BSBGSCM) ile bağımlı değişken olarak matematik birinci değerlendirme sonucu (BSMMAT01) seçilmiştir. Bunların dışında kalan tüm veriler silinmiştir. Bağımlı değişken z puanlarına dönüştürüldükten sonra +1 standart sapma üstü başarılı, -1 standart sapma altı başarısız olarak etiketlenmiştir.

Araştırmanın başında veri düzenleme başlığı altında yapılan bir diğer çalışma ise kayıp verilerdir. Kayıp veriler incelendiğinde anlamlılık düzeyi .698 çıkmıştır. Kayıp verilere ortalama değer atanmaya karar verilmiştir.

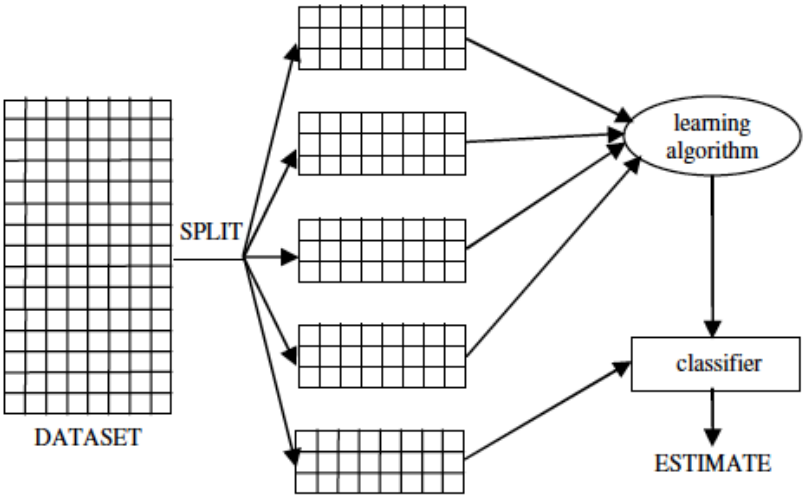
Çalışmada verilerin analizi için WEKA isimli veri madenciliği programı kullanılmıştır. Program Waikato u¨niversitesi tarafından geliştirilmiş ve ismi Waikato Environment for Knowledge Analysis kelimelerinin baş harflerinden oluşturulmuştur. Java programa dili tabanlı bir yazılımdır.(Frank et al., 2016).

Veriler düzenlendikten sonra R programı ile Sav dosyasında kayıtlı olan veriler, WEKA programı için gerekli olan arff dosyasına dönüştürülmüştür. Ardından WEKA programı pencerelerinden biri olan KnowledgeFlow ile Sınıflandırma analizleri türlerinden literatürde en çok tercih edilen K-En Yakın

Komşu Algoritması, Naive Bayes Algoritması, Yapay Sinir Ağları, Karar Ağacı Algoritmaları ve Lojistik Regresyon Analizi tercih edilmiştir. Bu çalışmada, k değeri için sıklıkla kullanılan 10 değeri alınmıştır yani 1327 öğrencinin verileri ile 10-katlı çapraz doğrulama yapılmıştır. 10 parçanın 9 tanesi eğitim seti, 1 tanesi test grubu olarak seçilmiştir. 1194 öğrencinin verisi eğitim setinde, 133 öğrencinin verisi de test grubunda yer almaktadır. Bu işlem her defasında test grubunu değiştirmek suretiyle 10 kez tekrar edilir. Çapraz doğrulama yöntemi olarak sıklıkla k-katlı ve n-katlı kullanılır. k-katlı çapraz doğrulama örneklem sayısı az olduğunda (ve birçoğunun boyutu ne olursa olsun kullanmayı tercih ettiği) sıklıkla benimsenen alternatif bir "eğit ve test et" yaklaşımıdır. Veri kümesi N örneklem içeriyorsa, bunlar k eşit parçaya bölünür, k tipik olarak 5 veya 10 gibi küçük bir sayıdır. K'ya ayrılan parçanın her biri sırayla test seti olarak, diğer k -1 parça ise eğitim seti olarak kullanılır (Max Bramer, 2007).

Şekil 2.1

K Katlı Çapraz Doğrulama (Max Bramer, 2007)



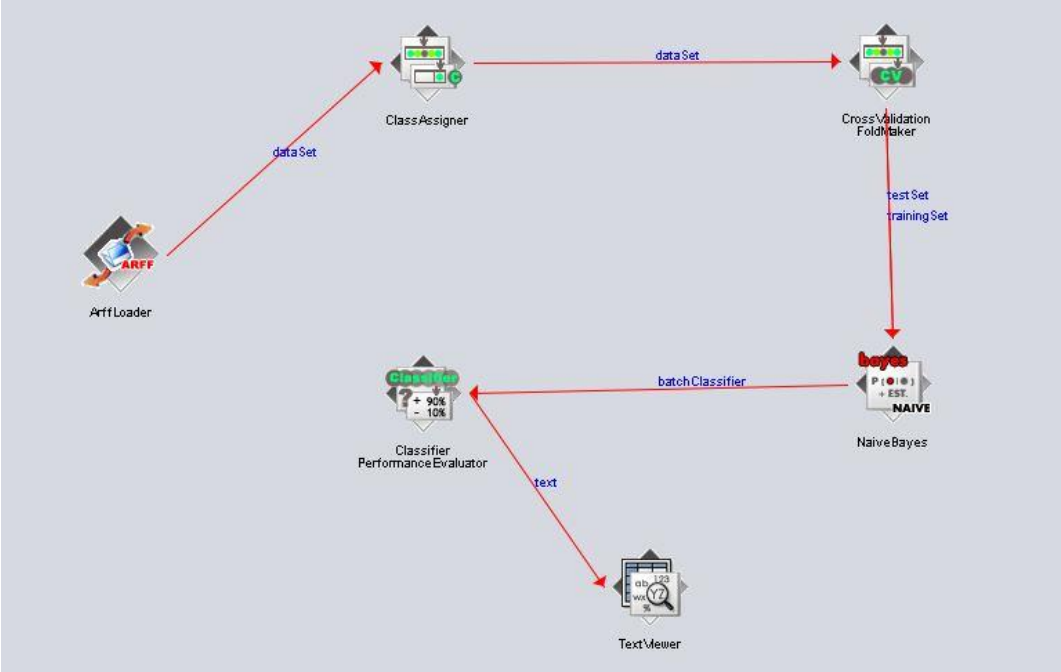
N-katlı çapraz doğrulama, genellikle *birini dışarıda bırak* çapraz doğrulama veya jack-knifing olarak bilinen, veri kümesinin örneklem kadar çok parçaya bölündüğü, k-katlı çapraz doğrulamanın aşırı bir durumudur ve her örneklem etkin bir şekilde bir test seti oluşturur. Her biri N – 1 olan örneklemden N sınıflandırıcı oluşturulur ve her biri tek bir test örneklemini sınıflandırmak için kullanılır. Dahil edilen büyük miktarda hesaplama, N katlı çapraz doğrulamayı büyük veri kümeleriyle kullanım için uygun hale getirir. Pratikte, sınıflandırıcıyı eğitmek için mümkün olduğu kadar çok verinin kullanılması gereken çok küçük veri kümeleri ile yöntemin faydalı olması da muhtemeldir (Max Bramer, 2007).

Verilerin Analizi için Weka Programında farklı pencereler bulunmaktadır. Explorer penceresinde verilere uygun olan analizler kendiliğinden belirmekteyken, eğer veri uygun değilse uygulanamayacak analizler kendiliğinden aktifliği kapanmaktadır. Bir de programda Knowledge Flow bölümü ile analizler yapılabilmektedir. KnowledgeFlow, WEKA'ya veri akışından ilham alan bir ara yüzdür. Kullanıcıya, bir

araç çubuğundan WEKA bileşenlerini seçme, bunları bir yerleşim tuvaline yerleştirebilme ve verileri işlemek ve analiz etmek için bir bilgi akışı oluşturma amacıyla bunları birbirine bağlama imkanı sunar. WEKA'nın Explorer penceresinde bulunan tüm sınıflandırıcıları bazı ekstra araçlarla birlikte KnowledgeFlow'da da mevcuttur (Hall ve Reutemann, 2008). Şekil 2.2'de bu pencerede yapılan Naive Bayes analizinin şeması örnek olarak verilmiştir.

Şekil 2.2

Knowledge Flow Bölümünde Naive Bayes Algoritması Kullanılarak Yapılan Analizin Şeması



Sonuçlar

Bu bölümde Weka programıyla yapılan analiz sonuçları sınıflandırma kriterlerine göre incelenmiştir.

Sınıflandırma Algoritmaları için Kriterler

Algoritmaların etkiliğini belirlemek adına birçok kriter kullanılmaktadır. Bu kriterlere örnek olarak doğru pozitif (DP) oranı (duyarlılık) , yanlış pozitif (YP) oranı, hassasiyet (precision)(doğru sınıflandırılmış pozitif örneklerin toplam pozitif örneklerin sayısına oranıdır), f- ölçütü (f-measure)(hassasiyet ve DP oranının harmonik ortalamasıdır) , alıcı işlem karakteristiği eğrisi (ROC)(bu kriter görsel olarak da yorumlama imkanı sunar), Kappa istatistiği (k)(ki-kare tablosuna dayalı bir değerdir), Ortalama mutlak hata(MAE)(tahmin edilen ve gözlemlenen değerler arasındaki mutlak farkların ortalamasıdır), kök ortalama kare hata (RMSE)(MAE'deki kare farklarının ortalama karekköküne eşittir), Matthews korelasyon katsayısı (MCC)(-1 ile +1 arasında değer alır, +1 mükemmel

tahmin anlamına gelir) verilebilir. Ayrıca karşılaştırma tablosu Gerçek değer ve tahminlenen değer arasında doğru pozitif, yanlış pozitif, yanlış negatif ve doğru negatif değerleri vermek için kullanılabilir (Çığşar ve Ünal, 2019; Nasa ve Suman, 2012).

Ev eğitim kaynakları (BSBGHER) ve matematikte öğrencinin özgüveni (BSBGSCM) değişkenleri kullanılarak öğrencilerin başarılı ve başarısız sınıflarına dahil olma durumlarını gösteren karşılaştırma matrisleri Tablo 3.1'de verilmiştir. Bu matrise göre IBk 1327 öğrenci verisinin 1155 tanesini (592+563) doğru sınıflandırmıştır. Ayrıca J48 1169(589+580), NB 1174 (567+607), MLP 1167 (574+593), LR 1176 (580+ 596) ve K star 1178 (575+ 603) tane doğru sınıflandırma yapmıştır.

Tablo 3.1

Karşılaştırma Tablosu

	ıBk		J 48		NB	
	a	b	a	b	a	b
a=başarılı	592	73	589	76	567	98
b=başarısız	99	563	82	580	55	607
	MLP		LR		K star	
	A	b	a	b	a	b
a=başarılı	574	91	580	85	575	90
b=başarısız	69	593	66	596	59	603

Tablo 3.2'de sınıflandırma sonuçları verilmiştir. Sınıflandırma sonuçları incelendiğinde, DP oranına göre LR (0,886) ve K star (0.888) en iyi algoritma belirlenmiştir. K star için Precision (.889) ve MCC (.776) değerleri; LR için MAE (.162) ve RMSE (.285) ve ROC alanı (.957) bu sonucu desteklemektedir. Ayrıca diğer algoritmalar incelendiğinde benzer performansların olduğu görülebilir.

Tablo 3.2

Sınıflandırma Sonuçları

	Sınıflandırma Algoritmaları					
Kriterler	IBk	J48	NB	MLP	LR	K star
DP oranı	0.870	0.881	0.885	0.879	0.886	0.888
YP oranı	0.130	0.119	0.115	0.120	0.114	0.112
Precision	0.871	0.881	0.886	0.880	0.887	0.889
F ölçütü	0.870	0.881	0.885	0.879	0.886	0.888
K istatistik	0.740	0.762	0.769	0.759	0.772	0.776
MAE	0.157	0.165	0.178	0.163	0.162	0.194
RMSE	0.337	0.302	0.288	0.288	0.285	0.291
ROC Area	0.921	0.933	0.956	0.955	0.957	0.955
MCC	0.741	0.762	0.771	0.759	0.773	0.776

Ayrıca J48 karar ağacına yönelik bulgular aşağıda verilmiştir. Karar ağacının boyutu 25, yaprakların sayısı 13 çıkmıştır.

Şekil 3.1

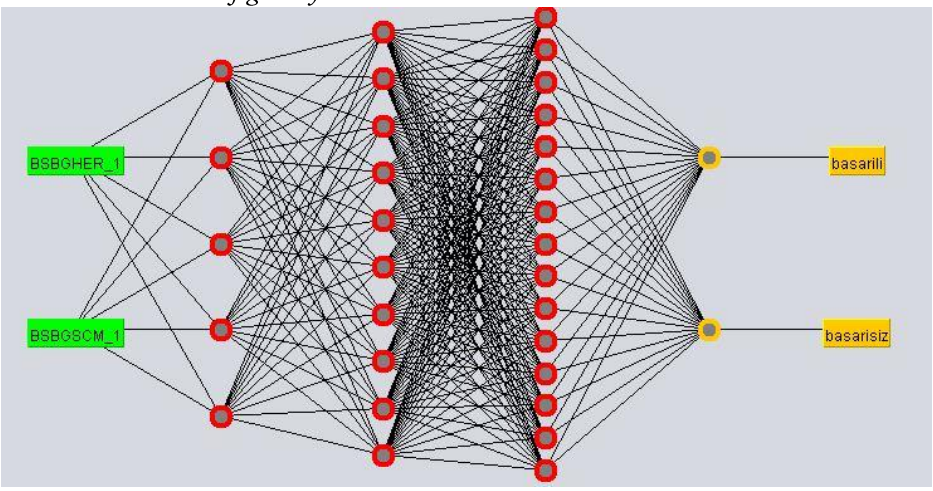
J48 Karar Ağacı Modeli

```
J48 pruned tree
------------------

BSBGSCM_1 <= 10.5867
|   BSBGHER_1 <= 10.05509
|   |   BSBGHER_1 <= 8.99877: basarisiz (411.0/14.0)
|   |   BSBGHER_1 > 8.99877
|   |   |   BSBGSCM_1 <= 9.89515: basarisiz (162.0/22.0)
|   |   |   BSBGSCM_1 > 9.89515
|   |   |   |   BSBGSCM_1 <= 10.35323: basarili (24.0/8.0)
|   |   |   |   BSBGSCM_1 > 10.35323: basarisiz (2.0)
|   BSBGHER_1 > 10.05509
|   |   BSBGHER_1 <= 11.19079
|   |   |   BSBGSCM_1 <= 9.834: basarisiz (70.0/25.0)
|   |   |   BSBGSCM_1 > 9.834: basarili (24.0/7.0)
|   |   BSBGHER_1 > 11.19079
|   |   |   BSBGHER_1 <= 12.1587
|   |   |   |   BSBGHER_1 <= 11.48712: basarili (37.0/7.0)
|   |   |   |   BSBGHER_1 > 11.48712: basarisiz (5.0/1.0)
|   |   |   BSBGHER_1 > 12.1587: basarili (55.0/4.0)
BSBGSCM_1 > 10.5867
|   BSBGHER_1 <= 9.64761
|   |   BSBGHER_1 <= 7.34568
|   |   |   BSBGSCM_1 <= 13.19086: basarisiz (18.0/2.0)
|   |   |   BSBGSCM_1 > 13.19086: basarili (5.0/1.0)
|   |   BSBGHER_1 > 7.34568: basarili (184.0/27.0)
|   BSBGHER_1 > 9.64761: basarili (330.0/4.0)

Number of Leaves  :     13

Size of the tree :      25
```

Yapay sinir ağları yaklaşımından MLP algoritması yapılırken 10 kat çapraz doğrulama yerine percentage split %66 seçilerek oluşturulan Neural Network Grafiğinin ara yüzü aşağıda verilmiştir. Şekilde yeşil renkle gösterilen kısım giriş katmanını, kırmızı kısım gizli katmanı, Sarı renkli kısım ise çıkış katmanı temsil etmektedir. Gizli katmandaki kırmızı layerslar hidden layers seçeneğinde a olarak gelmektedir. Çalışmada bu bölüm 5-10-15 olarak değiştirilmiştir.

Şekil 3.3

Neural Network Grafiği arayüzü



Sonuçlar

Araştırmada TIMMS 2019 verilerinde yer alan Matematik birinci değerlendirme sonucu üzerinde etkisi olduğu düşünülen bağımsız değişkenlere bağlı olarak öğrencilerin hangi sınıfta yer alacağı K-En Yakın Komşu Algoritması, Naive Bayes Algoritması, Yapay Sinir Ağları, Karar Ağacı Algoritmaları ve Lojistik Regresyon gibi farklı istatistiksel yöntemlerle tahminlenmiştir. En iyi tahminleyicilerin K star ve lojistik regresyon olduğu tespit edilmiştir. Alan incelemesi yapıldığında da LR, öğrencilerin akademik performansını belirlemek için yapılan çeşitli veri madenciliği yöntemleri arasında en iyi sınıflandırma algoritması olduğu görülmüştür. Çalışmada k-star'ın DP oranı .888 olduğu görülmektedir. Bunun anlamı; bağımsız değişkenlerin değerlerinin bilinmesi durumunda herhangi bir öğrencinin başarılı olup olamayacağının belirlenmesi %88.8 oranında doğru bir şekilde yapılabilecektir. Bu çalışmada literatürde yer alan alışagelmiş bakış açılarından ziyade farklı bir istatistiksel yöntem olan veri madenciliği kullanılmıştır. Bunun sebebi standart istatistiksel yöntemlerin varsayımlarının çokluğu ve karşılanmasının zorluğudur.

Kaynaklar

Aksu, M. Ç. ve Karaman, E. (2017). Karar ağacları ile bir web sitesinde link analizi ve tespiti. *ACTA INFOLOGICA*, *1*(2), 84-91. https://dergipark.org.tr/tr/download/article-file/398965

Bulut, H. C., Demirtaşlı, N., Yalçın, S. ve İlgün-Dibek, M. (2017). Türk öğrencilerin TIMSS 2007 ve 2011 matematik başarısında öğrenci ve öğretmen özelliklerinin etkileri. *Eğitim ve Bilim*, *42*, 27–47. http://dx.doi.org/10.15390/EB.2017.6885

Çığşar, B., and Ünal, D. (2019). Comparison of data mining classification algorithms determining the default risk. *Scientific Programming*, *2019*, 8706505. https://doi.org/10.1155/2019/8706505

Filiz, E. ve Öz, E. (2019). *Makine öğrenmesi yöntemleri ve eğitim verisi üzerine bir uygulama: uluslararası matematik ve fen eğilimleri araştırması 2015 Türkiye örneği* (Tez No. 598315) [Doktora Tezi, Yıldız Teknik Üniversitesi] Yükseköğretim kurumu Tez Merkezi.

Hall, M., & Reutemann, P. (2008). *WEKA knowledgeflow tutorial for version 3-5-8*. http://software.ucv.ro/~eganea/AIR/KnowledgeFlowTutorial-3-5-8.pdf

Han J., Micheline, K., and Pei J. (2012). *Data Mining* (3th ed.). Elsevier Inc. http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

Kamruzzaman, S., and Sarkar, A. (2011). A new data mining scheme using artificial neural networks. *Sensors (Basel, Switzerland)*, *11*, 4622–4647. https://doi.org/10.3390/s110504622

Laroche, S., Joncas, M., and Foy, P. (2019). Sample design in TIMSS 2019. In M. O. Martin, M. von Davier, and I. V. S. Mullis (Eds), *Methods and procedures: TIMSS 2019 technical report* (pp. 3.1-3.33). TIMSS & PIRLS International Study Center.

Lateef, U., and Adenubi, A. (2013). Artificial neural network (ANN) model for predicting students' academic performance. *Journal of Science and Information Technology (JOSIT-TASUED)*, *13*(2), 61–71.

https://www.researchgate.net/publication/275644377_Artificial_Neural_Network_ANN_Model_For_Predicting_Students'_Academic_Performance

Max Bramer. (2007). *Principles of data mining*. Springer.

Mullis, I. V. S., Martin, M. O., Foy, P., and Arora, A. (2012). *TIMSS 2011 International Results in Mathematics*. TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2011/downloads/T11_IR_Mathematics_FullBook.pdf

Nasa, C., & Suman, S. (2012). Evaluation of different classification techniques for WEB data. *International Journal of Computer Applications*, *52*, 34–40. https://doi.org/10.5120/8233-1389

Polat, M. (2019). *TIMSS-2015 matematik ve fen duyuşsal özellik modellerinin kültürlere, cinsiyete ve bölgelere göre ölçme değişmezliğinin incelenmesi* (Tez No. 564356) [Yüksek lisans tezi, Hacettepe Üniversitesi]. Yükseköğretim Kurumu Tez Merkezi.

Refaat, M. (2007). Review of data mining modeling techniques. In M. Refaat (Ed.), *Data preparation for data mining using SAS* (pp. 15–27). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-012373577-5/50005-3

Sebastian, S., and Puthiyidam, J. (2015). Evaluating Students Performance by Artificial Neural Network using WEKA. *International Journal of Computer Applications*, *119*, 36–39. https://doi.org/10.5120/21380-4370

Silahtaroğlu, G. (2008). *Kavram ve algoritmalarıyla temel veri madenciliği*. Papatya.

Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013). *Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms*. https://arxiv.org/pdf/1208.3719.pdf

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *The WEKA workbench: data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann. https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf

Meta analiz çalışmalarında heterojenlik testlerinden tau kare kestirim yöntemlerinin karşılaştırılması

Görkem Ceyhan ve Ceren Mutluer

Anahtar kelimeler: Meta analiz, heterojenlik testleri, tau kare, moment kestirim yöntemleri, en küçük kareler yöntemleri, bayes metotlar, maksimum olabilirlik yöntemleri

Giriş

Meta-analiz, bağımsız çalışmalardan elde edilen bilgilerin bütünleştirildiği istatistiksel bir tekniktir. Meta analiz çalışmalarında istatistiksel süreç etki büyüklüklerinin hesaplanması ile başlamaktadır (Field, 2001). Araştırmaya dahil edilen her bir etki büyüklüğü ile bütüncül sonuca erişebilmek için gerçek etki büyüklüğünü tanımlamak gerekmektedir. Araştırmada gerçek etkinin evrende tek olup olmadığına karar verebilmek içim model seçimini doğru yapabilmek gerekir. Meta analiz çalışmalarında iki model bulunmaktadır. Bunlar sabit etkiler modeli ve rastgele etkiler modelidir (Borenstein ve diğ., 2010). Sabit Etki (FE) Modelindeki temel varsayım, meta analizine alınacak olan her bir çalışmanın tamamen ortak bir etki büyüklüğüne sahip olmasıdır. Yani, etki büyüklüğünü etkileyen bütün faktörler meta analizine dahil edilen tüm çalışmalarda aynı olduğu için gerçek etki büyüklüğü bütün çalışmalarda sabittir (Borenstein ve diğ., 2009). Rastgele Etkiler (RE) modelinde ise bazen evrende farklı alt gruplar olabilir. Yani gerçek etki sabit bir etki olmayabilir. Farklı alt gruplarda farklı farklı etki büyüklükleri vardır. Bir meta analiz çalışması için sabit etkiler modeli mi? Rastgele etkiler modeli mi kullanılması gerektiğine karar vermemiz gerekiyor. Bunun için öncelikle teorik olarak desteklemek gerekmektedir. Teorik temeli daha önceden yapılmamış çalışmalar için heterojenlik testleri bu iki modelin hangisinin uygun olduğunu göstermede kullanılmaktadır. Heterojenlik için güven aralıkları, p, Cohran'ın Q istatistiği, Tau kare ve I^2 istatistikleri kullanılmaktadır. Bu çalışmada sadece Tau kare istatistiği üzerinde durulacaktır.

τ^2 gerçek etki büyüklüğünün varyansı olarak tanımlanmaktadır. Diğer bir deyişle, eğer sonsuz sayıda geniş örnekleme sahip çalışmalarla analiz yapılıyorsa her bir çalışmanın etki tahmini gerçek etkiye eşit olacak ve bu etkilerin varyansı gerçek varyans ile gösterilecektir (Borenstein ve diğ, 2009). Bu varyans $\tau^{2\prime}$ dir. τ^2çalışmalar arası varyansdır ve meta analizindeki rastgele etki modelindeki gibi tahmin edilir.

τ^2 çok sayıda yöntem kullanılarak tahmin edilir. Bu yöntemler moment kestirim (DerSimonian ve Laird, Two-step DerSimonian ve Laird, Hartung-Makambi, Hunter-Schmidt, Cochran's ANOVA, Two-step Cochran's ANOVA), en küçük kareler (Sidik- Jonkman, Alternative Sidik-Jonkman), Bayes(Rukhin Bayes, Positive Rukhin Bayes,Rukin's Bayesian estimator with zero prior), non parametrik (Non-Parametric bootstrap DerSimonian ve Laird, Malzahn, B**o**¨hning and Holling) ve maksimum likelihood yöntemler (Maximum likelihood, Restricted maximum likelihood) olarak beş kategoriye ayrılmaktadır (Thorlund ve diğ, 2011; IntHout ve diğ., 2014; Langan ve, diğ., 2019). Belirtilen τ^2 kestirim yöntemlerinin karşılaştırılması bu araştırmanın genel amacını oluşturmaktadır. Bu genel amaç doğrultusunda aşağıdaki sorulara yanıt aranacaktır:

1-İkili veriler ve sürekli veriler için farklı kestirim yöntemlerinden elde edilen τ^2 değerleri farklılık göstermekte midir?

2-Çalışma sayısındaki değişime göre farklı kestirim yöntemlerinden elde edilen τ^2 değerleri farklılık göstermekte midir?

3-Çalışmalarda yer alan örneklem sayısına göre farklı kestirim yöntemlerinden elde edilen τ^2 değerleri farklılık göstermekte midir?

Yöntem

Bu araştırmada nicel araştırma desenlerinden meta analiz deseninde kullanılan bir heterojenlik testi olan tau kare değerinin çeşitli kestirim yöntemlerine göre değişimini ele almaktadır. Araştırma τ^2 kestirim yöntemlerinin karşılaştırılmasını amaçlaması yönüyle betimsel bir çalışmadır. Betimsel çalışmalarda ele alınan durum olabildiğince tam ve dikkatli bir şekilde tanımlanır (Büyüköztürk ve diğ., 2012).

Araştırma sürecinde YÖK Ulusal Tez Merkezi, Google Akademik, Web of Science veri tabanları kullanılmıştır. 'Tamamlayıcı' kelime yerine 'alternatif' kelimesi kullanımı sıklıkla kullanılan bir kavram yanılgısı olarak veri tabanlarında tez, makale ve bildiri olarak ele alınmaktadır. Bu durumu da dikkate alarak belirtilen veri tabanlarında 2000-2021 yılları arasında yapılmış çalışmalar arasında 'tamamlayıcı ölçme ve değerlendirme', 'alternatif ölçme ve değerlendirme', bu ölçme ve değerlendirme tekniklerinin temel alınıp 'akademik başarı' kavramının incelendiği çalışmalar taranmıştır. Meta-analiz çalışmasına dahil edilen araştırmaların seçilmesinde aşağıdaki ölçütler kullanılmıştır:

i. TÖD/AÖD tekniklerinin öğrencilerin akademik başarıları üzerindeki etkisinin incelendiği bir araştırma olması,
ii. Türkiye'de yayımlanan yüksek lisans, doktora tezi, makale ya da bildiri olması,
iii. Araştırmada etki kavramını incelemek için kontrol ve deney gruplarının yer alması,
iv. Etki büyüklüğünün hesaplanabilmesi için araştırmalarda yeterli bilginin (örneklem büyüklüğü, ortalama, standart sapma) yer alması

Belirtilen ölçütler bağlamında toplamda çalışmaların etki büyüklükleri kullanılarak tau kare için moment kestirim, en küçük kareler, Bayes, non parametrik ve maksimum likelihood yöntemleri uygulanacaktır.

Heterojenlik varyans kestirim yöntemlerinin performansı beş ölçüm kullanılarak değerlendirilmiştir. Bu değerlendirme kritlerleri şunlardır; etki büyüklüğü, Mean Absulate Estimation Errror (MAEE), θ için Covarega Probability, θ için Güven Aralığı ve güçtür. Kestirim yöntemleri bu beş açıdan k, θ ve $\tau 2$ 'nin bütün kombinasyonlarında, İkili veriler için OR, sürekli veriler için MD ve SMD hesaplanarak değerlendirme yapılacaktır.

Sonuçlar

Araştırma sürecine öncelikle meta analiz çalışması için belirtilen ölçütleri sağlayan çalışmalar seçilecektir. Seçilen araştırmalara ilişkin bir çizelge oluşturulacaktır. Ölçütlere uyan çalışmalar için ayrı ayrı etki büyüklükleri hesaplanacaktır. Her bir etki büyüklüğü için ağırlıklandırma yapılacaktır. Heterojenlik testlerinden tau kare kestirim yöntemi için moment kestirim DerSimonian ve Laird, Two-step DerSimonian ve Laird, Hartung-Makambi, Hunter-Schmidt, Cochran's ANOVA, Two-step Cochran's ANOVA), en küçük kareler (Sidik- Jonkman, Alternative Sidik-Jonkman), Bayes(Rukhin Bayes, Positive Rukhin Bayes, Rukin's Bayesian estimator with zero prior), non parametrik (Non-Parametric bootstrap DerSimonian ve Laird, Malzahn, B**o**¨hning and Holling) ve maksimum likelihood yöntemleri (Maximum likelihood, Restricted maximum likelihood) uygulanacaktır.

Belirtilen araştırmanın amaç cümlesi doğrultusunda tau kare için hesaplanan değerlerdeki farklılıklar incelenecek, çalışma sayısı ve örneklem büyüklüğü değişkenlerine göre tau kare değerlerindeki farklılık etki büyüklüğü, Mean Absulate Estimation Errror (MAEE), θ için Covarega Probability, θ için Güven Aralığı ve güç açısından incelenecektir.

Kaynaklar

Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research synthesis methods, 1*(2), 97-111. https://doi.org/10.1002/jrsm.12

Büyüköztürk, Ş., Çakmak, E. K., Akgün, Ö. E., Karadeniz, Ş. ve Demirel, F. (2012). *Bilimsel araştırma yöntemleri*. Pegem Akademi.

IntHout, J., Ioannidis, J. P., & Borm, G. F. (2014). The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC medical research methodology, 14*(1), 1-12. https://doi.org/10.1186/1471-2288-14-25

Langan, D., Higgins, J. P., Jackson, D., Bowden, J., Veroniki, A. A., Kontopantelis, E., Viechtbauer, W., and Simmonds, M. (2019). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research synthesis methods*, *10*(1), 83-98. https://doi.org/10.1002/jrsm.1316

Thorlund, K., Wetterslev, J., Awad, T., Thabane, L., and Gluud, C. (2011). Comparison of statistical inferences from the DerSimonian–Laird and alternative random-effects model meta-analyses–an empirical assessment of 920 Cochrane primary outcome meta-analyses. *Research synthesis methods*, *2*(4), 238-253. https://doi.org/10.1002/jrsm.53

A software comparison on item parameter recovery in multistage adaptive testing

Rabia Karatoprak Erşen

Giriş

There are commercial and non-commercial softwares to conduct psychometric analysis using the Item Response Theory (IRT) such as Cai (2017), Zimowski et al. (2003), Partchev and Maris (2017) or Chalmers (2012). flexMIRT (Cai, 2017) is one of the widely used IRT computer programs for operational or research purposes Lee & Lee (2018). It is a user-friendly software with capability to run large datasets in a short time. Also, it can accommodate many IRT models such as explanatory, mutidimensional or multilevel IRT models as well as unidimensional ones. A minor point about flexMIRT is that it is a commercial software. It limits the accessibility and common usage.

Irtoys (Partchev and Maris, 2017) and mirt (Chalmers, 2012) are free software used on R platform. R (R Core Team, 2018) is a free program with many psychometric packages maintained by individuals and comes with no warranty (see Schumacker (2019) for the list of these packages). Analysis of irtoys uses only unidimensional whereas mirt uses unidimensional and multidimensional IRT models. Analysis capacity of mirt has been expanded since its first version. Current version can support explanatory IRT or cognitive diagnostic models (Chalmers, 2020).

There are studies compared software output in terms of parameter recovery using multidimensional or unidimensional IRT models Wang (2018). Common finding was that different software produce different item parameter estimates. However, all these studies used paper-pencil tests and it is unknown if their results are applicable for computerized multistage adaptive testing (MST) exams. Since item parameter estimates are used for scoring, equating or differential item functioning, it is important to eliminate confounding effect of software from estimates. So, the purpose of this study is to compare item parameter estimates obtained from FlexMIRT (Cai, 2017), irtoys (Partchev & Maris, 2017) and mirt (Chalmers, 2012) in terms of item parameter recovery in MST. Item parameter recovery, in this study, refer to evaluation of estimated parameters of new items when an MST item pool is aimed to be expanded. Therefore, items should be regarded as pretest items.

Method

Several studies have examined item calibration and linking in multistage adaptive testing (MST) designs (Chuah et al., 2006; Eggen and Verhelst, 2011; Glas, 1988; Zheng, 2012). Related to calibration and linking of new items, Karatoprak Ersen (2019) investigated various ways in a simulated 1-3 MST design. Two pretesting models mimicking the operational practices were employed: all pretest items administered within a module referred to as the embedded-section (ES) model or distributed to all modules referred to as the embedded-items (EI) model. As an IRT software, flexMIRT (Cai, 2017) was utilized. For comparability purposes, these pretesting models, MST design and calibration methods are kept the same in this study.

MST design is 1-3. Routing points are determined through approximate maximum information (Luecht et al., 2006). The examinee responses are scored using expected a posteriori estimation. The total number of items is 40 as 20 items in each stage. Unidimensional 3PL IRT model is used.

Calibration methods are fixed and separate calibration with the Stocking-Lord (Stocking & Lord, 1983) linking method. Since purpose of this study, compare and understand results produced from available IRT software, irtoys (Partchev and Maris, 2017) and mirt (Chalmers, 2012) are utilized additional to FlexMIRT (Cai, 2017).

Suggestions of existing literature (e.g., Ban et al., 2001; Lee and Ban, 2009) are also followed to determine the study factors. Therefore, ability distributions and sample size are included. Conditions of ability distributions are N(0, 1), N(.5, 1), and N(.5, .64). Conditions of sample size of the examinees assigned to the routing module is 800, 1500 and 3000.

Findings

Accurate estimation of item parameters is crucial for score validity. This study aims to contribute by comparing software results in terms of item parameter recovery. Findings are expected to guide practitioners or researchers deciding which software to use for item calibration. If comparable item parameter estimates are obtained from the irtoys (Partchev & Maris, 2017), mirt (Chalmers, 2012) or FlexMIRT (Cai, 2017), free software might standout compared to the commercial one especially when there is not a funding for the analysis. Also, interpretation of findings produced by these software can be done as parallel to each other. If each software shows some strengths or weaknesses such as estimation time or cost, then identifying them for specific conditioins will be valuable to choose one of them for analysis.

References

Ban, J.-C., Hanson, B. A., Wang, T., Yi, Q., and Harris, D. J. (2001). A comparative study of on-line pretest item: Calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement*, *38*(3), 191–212. http://www.jstor.org/stable/1435120

Cai, L. (2017). *flexMIRT: Flexible multilevel multidimensional item analysis and test scoring* (version 3.51) [Computer Software]. Vector Psychometric Group.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P. (2020). *Mirt: Multidimensional item response theory* (version 1.33.2) [Computer Software]. https://cran.r-project.org/web/packages/mirt/mirt.pdf

Chuah, S. C., Drasgow, F., and Luecht, R. M. (2006). How big is big enough? Sample size requirements for CAST item parameter estimation. *Applied Measurement in Education*, *19*(3), 241–255. https://doi.org/10.1207/s15324818ame1903_5

Eggen, T. J., and Verhelst, N. D. (2011). Item calibration in incomplete testing designs. *Psicologica: International Journal of Methodology and Experimental Psychology*, *32*(1), 107–132.

Glas, C. A. W. (1988). The Rasch model and multistage testing. *Journal of Educational Statistics*, *13*(1), 45. https://doi.org/10.2307/1164950

Karatoprak-Ersen, R. (2019). *Pretest item calibration in multistage adaptive testing* (Publication No. 9983779597902771) [Doctorate dissertation, University of Iowa]. https://doi.org/10.17077/etd.005206

Kim, K. Y., and Lee, W. (2017). The impact of three factors on the recovery of item parameters for the three-parameter logistic model. *Applied Measurement in Education*, *30*(3), 228–242. https://doi.org/10.1080/08957347.2017.1316274

Lee, W. C., and Ban, J. C. (2009). A comparison of IRT linking procedures. *Applied Measurement in Education*, *23*(1), 23–48. https://doi.org/10.1080/08957340903423537

Lee, W. C., and Lee, G. (2018). IRT linking and equating. In P. Irwing, T. Booth, D. J. Hughes (Eds.), *The wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 639–673). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118489772.ch21

Li, T. (2019). A comparison between BMIRT and IRTPRO: A simulation study of a multidimensional item response model. *American Journal of Educational Research*, *7*(11), 865–871. https://doi.org/10.12691/education-7-11-17

Partchev, I., and Maris, G. (2017). *Irtoys: A collection of functions related to item response theory (IRT).* [Computer Software]. https://cran.r-project.org/package=irtoys

R Core Team. (2018). *R: A language and environment for statistical computing* [Computer Software]. https://www.R-project.org/

Schumacker, R. (2019). Psychometric packages in r. *Measurement: Interdisciplinary Research and Perspectives*, *17*(2), 106–112. https://doi.org/10.1080/15366367.2018.1544434

Stocking, M. L., and Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210. https://doi.org/10.1177/014662168300700208

Wang, J. (2018). *Does it matter which IRT software you use? Yes.* [Technical Report]. Assessment Systems. https://assess.com/docs/Which_IRT_Software.pdf

Yavuz, G., and Hambleton, R. K. (2017). Comparative analyses of MIRT models and software (BMIRT and flexMIRT). *Educational and Psychological Measurement*, *77*(2), 263–274. https://doi.org/10.1177/0013164416661220

Zheng, R. C. (2012). Item calibration/recalibration in computerized multistage testing (MST)–evaluation of some practical solutions. *AICPA Summer Internship Project Report*.

Zimowski, M., Muraki, E., Mislevy, R., and Bock, R. (2003). BILOG-MG 3: Item analysis and test scoring with binary logistic models. https://www.semanticscholar.org/paper/BILOG-3-%3A-item-analysis-and-test-scoring-with-Mislevy-Bock/f6f301dacf8c37b0cfbfbf2fb9c8abb819ec4099#citing-papers

Puanlayıcılar arası uyumun farklı ölçekleme düzeyleri, farklı puanlayıcı sayısı ve farklı puanlanan sayısı açısından incelenmesi

Yılmaz Orhun Gürlük, Gizem Cömert, Mediha Korkmaz ve Ömer Emre Can Alagöz

Anahtar kelimeler: Puanlayıcılar arası uyum, sınıf-içi Korelasyon, Cohen kappa, Fleiss kappa, Krippendorff alfa

Giriş

Psikolojik özellikler değerlendirilirken çeşitli testler birçok araştırmacı ve alan çalışanı tarafından kullanılmaktadır. Bazı testlerin bazı maddeleri veya tamamı dereceli puanlama anahtarına uymamaktadır. Söz gelimi; alınan ölçümlerde açık uçlu maddeler kullanılmakta ve bu puanlar çeşitli kategorilere atanmakta veya maddeler doğrudan testi uygulayanlar tarafından belirli bir puan ranjı çerçevesinde puanlanmaktadır. Alan çalışanlarının puanlamalarına veya diğer bir ifade ile kararlarına göre bireyler tanı almakta, sınıflara yerleştirilmekte veya bir tema etrafında gruplandırılmaktadır. Bu durum araştırma ve uygulamaları önemli oranda etkilemektedir (Bıkmaz, 2011; Raykov, 2012).

Cohen(1960), bu amaçla nominal ölçekleme için kappa katsayısını geliştirmiştir. Cohen katsayısı çapraz tablolara dayanmakta ve puanlayıcıların aynı kategoride puanlama olasılığını göstermektedir. Elde edilen katsayı -1 ile +1 arasında değer almakta ve yüksek değerleri yüksek uyum anlamına gelmektedir. Orijinal formül 2 puanlayıcı için geliştirilmiş olsa da daha sonra çoklu puanlayıcıya uyarlanmıştır. Ancak buradan elde edilen değerler prevalans problemine sahiptir. Prevelans problemi, puanlayıcıların belli örüntüde puanlamasının; örneğin, her puanlayıcının aynı denekler olmasa dahi deneklerin %50'sine aynı puanı vermesinin, yüksek kappa değerlerine yol açması olarak tanımlanmaktadır. Ayrıca marjinal puanlayanlar değerin düşük tahminlenmesine yol açmaktadır (Hallgren, 2012). Landis ve Koch'a (1977) göre 0'dan küçük değerler uyumsuzluğu, 0.01-0.20 zayıf uyumu, 0.21- 0.40 orta düzeyde uyumu, 0.41-0.60 kabul edilebilir uyumu, 0.61-0.80 iyi uyumu ve 0.81-1.00 ise mükemmel uyumu göstermektedir.

Daha sonra Fleiss (1971) tarafından bir diğer kappa katsayısı geliştirilmiştir. Fleiss'ın katsayısı doğrudan çoklu puanlayıcı için tasarlanmış olması nedeniyle Cohen'in katsayısına göre daha tutarlı sonuçlar vermektedir. Fleiss kappa, puanlayıcıların, puanlayıcı evreninden seçkisiz olarak seçildiğini varsayar. Landis ve Koch'un (1977), değer sınıflandırması Fleiss kappa için de kullanılmaktadır. Çoklu puanlayıcı için Cohen kappa değeri anlamsız çıkma eğilimindeyken Fleiss kappa değeri 2 ve daha fazla puanlayıcı için daha tutarlıdır (Hallgren, 2012).

Ordinal ölçekleme düzeyine gelindiğinde ise Krippendorff alfa katsayısının sıklıkla kullanıldığı görülmektedir (Bıkmaz, 2017). Bu katsayı ağırlıklandırma prosedürü sayesinde ordinal, nominal ve sürekli değişken düzeylerine uyarlanabilmektedir. Örneklem büyüklüğü ve puanlayıcı sayısı ile ilgili düzeltmeler yapmaktadır. Gözlenen uyumsuzluk/beklenen uyumsuzluk oranına dayanır (Krippendorff, 2004). Krippendorff alfa'nın.67'den küçük değerleri zayıf, .67-.80 arası orta ve.80'den büyük değerleri ise yüksek uyumu göstermektedir (Hayes, 2007; Bıkmaz-Bilgen, 2017).

Uyum incelemelerinde en yaygın kullanılan yöntemlerden biri de sınıf içi korelasyon katsayısıdır (SKK). Fisher tarafından 1954 yılında Pearson korelasyon katsayısı temel alınarak geliştirilmiş olmasına rağmen günümüzdeki katsayı ortalama kareler aracılığıyla hesaplanmaktadır (Ateş, 2009). SKK en basit anlamıyla puanlayıcı varyanslarının birbirilerine olan oranıdır. Sınıf içi korelasyon katsayısı da tıpkı diğer benzerleri gibi 0 ile 1 arasında standart bir değer alır. SKK'nın 1'e yakın değerler alması puanların birbiri ile yüksek derecede uyumlu olduğunu gösterir. Sıfıra yakın bir sınıf içi korelasyon puanlayıcıların verdiği değerlerinin benzer olmadığına işaret eder, diğer bir ifadeyle puanlayıcılar arası uyum yoktur (Koo ve Li, 2010; Ateş, 2009).

Bu araştırmada ölçek maddelerinin sınıflayıcı, sıralayıcı ve sürekli ölçekleme düzeylerine göre yöntemler arasındaki değerlendiricler arası tutarlık inceleniştir. Aynı zamanda farklı sayıda puanlayıcının, farklı sayıda katılımcıyı puanladığı koşullarda farklı uyum yöntemlerinin farklılaşıp farklılaşmayacağı araştırılmıştır.

Yöntem

Araştırmanın verilerini Bender Gestalt II testinin Kopyalama puanları oluşturmaktadır. Bender Gestalt II Kopyalama maddeleri görsel motor bütünleştirme yeteneğini ölçmek üzere toplamda 16 farklı figürden oluşmaktadır, orijinal şeklin benzerliği temelinde likert formatında 0-4 (0-benzerlik yok ve 4 mükemmel çizim kalitesi) aralığında puanlanmaktadır (Korkmaz ve diğ., 2019). Bu araştırmada 35 katılımcının birbirine kör 8 puanlayıcı (6 kadın, 2 erkek) tarafından puanlandığı Kopyama maddelerinin 9 kartı veri olarak kullanılmıştır.

Araştırma kapsamında incelenen temel soru belirlenen ölçekleme düzeyine göre seçilen uyum değerlendirmelerinin anlamlı derecede farklılaşıp farklılaşmadığıdır. Bu amaç doğrultusunda kart puanlamalarının sürekli kabul edildiği durum için sınıf içi korelasyon katsayısı kullanılmıştır. Sınıflayıcı ölçekleme düzeyi için hem Cohen'in hem de Fleiss'ın kappa değerleri, son olarak sıralayıcı ölçekleme düzeyi için de Krippendorff alfa değeri tespit edilmiştir. Cohen Kappa ve Fleiss kappa tekniklerinin aynı anda kullanılma amacı; Cohen'in ilk istatistik olmasının yanı sıra Fleiss'ın marjinal değer ve çoklu puanlayıcı durumları için daha tutarlı sonuçlar verdiği yönündeki literatürün (Hallgren, 2012) desteklenip desteklenmediğini görmektir. Sınıf içi korelasyon katsayısı ve Fleiss kappa değerleri doğrudan SPSS 26.0 paket programı kullanılarak hesaplanmıştır. Cohen kappa değeri için David Nichols (1997) tarafından yazılan SPSS sentaksı kullanılmıştır. Krippendorff'un alfa değerleri için ise Hayes (2018) tarafından geliştirilen Kalpha 4.0 makrosundan yararlanılmıştır.

Araştırmada, 3 farklı puanlayıcı düzeyi (4 puanlayıcı, 6 puanlayıcı ve 8 puanlayıcı) ve 3 farklı katılımcı sayısı düzeyi (10, 20 ve 35 katılımcı) üzerinden değerlendiriciler arası uyum değerleri hesaplanmıştır. Tüm durumlar için sınıf içi korelasyon katsayısı, Cohen'in kappa istatistiği, bu değerin bir formu olan ancak marjinal puanlamalardan daha az etkilenen Fleiss'ın kappa değeri ve son olarak Krippendorff'un alfa güvenirlik katsayısı kullanılmıştır. Yöntemler arasındaki farklılaşmaların istatistiksel anlamlılığını incelemek amacıyla genel doğrusal modeller kullanılmıştır.

Sonuçlar

Ölçekleme düzeylerine göre yöntemler tablolar üzerinden incelendiğinde sürekli durum için kullanılan SKK; 10-20-35 puanlanan ve 4-6-8 puanlayıcının olduğu 9 farklı durum için her zaman en yüksek uyumu yakalamıştır. Tüm durumlar incelendiğinde ordinal ölçekleme düzeyini hesaplayan Krippendorff alfa; SKK'yi takip etmiştir. En düşük değerler Cohen kappada hesaplanmıştır. Tüm tablolalar incelendiğinde Cohen kappa'nın çoklu puanlayıcı için anlamsız çıkma eğiliminde olduğu görülmüştür. Çoklu puanlayıcı için Fleiss'ın kappası daha tutarlıdır. Ortalama düzeyinde SKK, Cohen ile Fleiss'ın kappaları ve Krippendorff alfanın farklılaşma durumunu görmek amacıyla tekrarlı ölçümler varyans analizi yapılmış ancak küresellik varsayımı karşılanmadığı (p< .001) için Huynh-Fedlt yöntemi kullanılmıştır. Farklı yöntemlerden elde edilen ortalama uyum düzeylerinin istatistiksel olarak anlamlı farklılaşma gösterdiği tespit edilmiştir, F(1.38, 109.23)= 323.90, η^2=0.80, *güç*=1.00, p< .001. Ortalama uyum değerleri düşükten yükseğe; Cohen κ ($\bar{X}$= 0.13, σ= 0.13), Fleiss κ ($\bar{X}$=0.22, σ= 0.11), Krippendorf α ($\bar{X}$=0.45, σ=0.19) ve son olarak SKK ($\bar{X}$= 0.59, σ= 0.18) şeklinde sıralanmaktadır.

Ardından, puanlanan sayısı ve puanlayıcı sayısına göre kullanılan uyum yöntemlerinin değişimi çok değişkenli varyans analiziyle incelenmiştir. Box M testine göre varyans homojenliğinin karşılanmıştır, F(80, 6050.19)= 1.06, M= 108.25, p= .33. Pillai'nin izi kriterine göre genel model anlamlı bulunmuştur, F(4, 68)= 318.51, η^2= .95, *güç*= 1.00, p< .001. Düzeltilmiş model tablosu incelendiğinde farklılaşmanın yalnızca Fleiss κ değerinde anlamlı olduğu görülmüştür, F(8, 71)= 2.61, η^2= .23, *güç*= .90, p= .01.

Kaynaklar

Ateş, C., Öztuna, D. ve Gen. Y.(2009). Sağlık araştırmalarında sınıf içi korelasyon katsayısının kullanımı. *Türkiye Klinikleri 1*(2), 59-64. https://www.proquest.com/openview/257d12f5de4c7cb0d1d09b5e7a4b85d0/1.pdf?pq-origsite=gscholar&cbl=236263&casa_token=AKutcwsQutoAAAAA:SkRWk-mmCSU3H1AaX_vm7nKGcWPSS7fKAbQeU0vghGfMH_CU--Dp0XoAmgFYwWi8aXUCMSUB

Bıkmaz-Bilgen, Ö. (2011). Üst düzey zihinsel özelliklerin ölçülmesinde puanalyıcılar arası güvenirlik belirleme tekniklerinin karşılaştırılması. *Yüksek Lisans Tezi. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Eğitim Bilimleri Anabilim Dalı Eğitimde Ölçme ve Değerlendirme Bilim Dalı*

Bıkmaz-Bilgen, Ö. ve Doğan, N. (2017). Puanlayıcılar arası güvenirlik belirleme tekniklerinin karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 8(1),* 63-78. https://doi.org/10.21031/epod.294847

Cohen (1960). A coefficient of rater agreement for nominal scales. *Educational and Psychological Measurement, 20(1),* 37-46.

Fleiss, J. L. (1971). Measuring agreement for multinomial data. *Psychological Bulletin, 76(5),* 378-382.

Hallgren, K. (2012). Computimg inter-rater reliability for observational data: An overview and tutorial. Torials in Quantitative Methods for Psychology, 8(1), *23-34.* https://doi.org/https://doi.org/10.20982/tqmp.08.1.p023

Hayes, A. F., and Krippendorff, K. (2007). Answering the call for a standard reliability measuse for coding. *Communication Methods and Measures, 1*(1), 77-89. https://doi.org/10.1080/19312450709336664

Korkmaz, M., Demiral, N., Sapmaz-Yurtsever,S., Kaçar-Başaran,S. ve Çabuk,T. (2019). *Bender Gestalt-II testinin Koppitz II ve Bender Gestalt II puanlama sistemlerine göre (4-7 ve 8-18 Yaş) ön norm çalışması.* Ege Üniversitesi Bilimsel Araştırma Projesi. https://avesis.ege.edu.tr/proje/206f3f76-b64f-4460-8030-e51bdcdda20c/bender-gestalt-ii-testinin-koppitz-ii-ve-bender-gestalt-ii-puanlama-sitemlerine-gore-4-7-ve-8-18-yas-on-norm-calismasi

Koo, T. K., and Li, M. Y. (2010). A guideline of selecting and reporting intraclass correlation coefficients for reliability. *Research. Journal of Chiropractic Medicine, 15*(2), 155-163. https:doi.org/10.1016/j.jcm.2016.02.012

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Sage Publication

Landis, J. R., and Koch, G. G. (1977) An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33(2)*, 363-374.

Raykov, T., Dimitrov, D. M., von Eye, A., and Marcoulides, G. A. (2012). Interater Agreement Evaluation: A latent variable modeling approach. *Ducational and Psychological Measurement, 20(10).* 1-20. https://doi.org/10.1177/0013164412449016

Von Eye, A. Ve Mun, E. Y. (2005). *Analyzing rater agreement manifest variable methods* (1st ed.). Lawrence Erlbaum Associates.

Determining differential item functioning using explanatory item response models and various methods

Serap Büyükkıdık ve Elif Özlem Ardıç

Keywords: Explanatory item response models (EIRM), transformed item difficulties (TID), Breslow day, Mantel Haenszel, logistic regression, standardization

Introduction

Determining the differential item functioning (DIF) is important as it affects the validity of the test (French & Finch, 2015). DIF can be defined as the change in the probability of individuals who differ in terms of characteristics like gender, ethnicity, socioeconomic level, creed, etc. to respond correctly to test items the same ability level (Shepard ve diğ., 1985). There are basically two reasons for the emergence of DIF. These reasons are item effect which show the real difference between subgroups and item bias. (Camilli and Shepard, 1994). There are many methods for determining DIF, such as mantel-haenszel, logistic regression, standardization, Transformed item difficulties (TID), and Breslow-day. In addition, DIF can be determined with explanatory item response models. A completely similar study has not been found in the literature regarding the comparison of all the mentioned DIF determination methods and the findings obtained from the explanatory item response models. For this purpose, differential item functioning analyzes of reading fluency items in the first booklet in the Programme for International Student Assessment (PISA) 2018 according to gender with mantel-haenszel, logistic regression, standardization, transformed item difficulties (TID), breslow-day methods done and explanatory item response model (response ~ -1 + item * gender) and the results were compared.

Method

Within the scope of the research, the data of 241 students, 117 male and 124 female, who responded to booklet 1 in PISA 2018, were used. Since it measures the same structure, the 22 reading fluency items in this booklet is included in the analysis.

In the analysis of the data, in the first stage, missing and invalid data were cleaned from the data set. The difR and lme4 (Bates et al., 2015) packages were used in the R Studio software. Mantel-Haenszel is a chi-square-based statistic. For this, chi-square values and significant values were calculated. The

transformed item difficulties method, called Angoff's Delta method, was calculated separately for the delta values and probabilities of the focus and reference groups. Statistics and significance levels were obtained for logistic regression, breslow-day, and standardization methods. "response ~ -1 + item * gender + (1 | id)" for DIF analysis within the scope of EIRM, "response ~ -1 + item + gender + (1 | id)" model was used to examine whether gender had a significant effect on students' performance. Model data fit of all these models was examined with Akaike's information criterion (AIC), Bayesian information criterion (BIC), log likelihood and deviance values.

Results

In the "response ~ -1 + item + gender + (1 | id)" model in EIRM, gender did not have a significant effect on students' performance. In the model in which the gender x item interaction is handled, the interactions for all items of "response ~ -1 + item * gender + (1 | id)" were not statistically significant. As a result of the research, 22 items do not contain DIF as a result of mantel haenszel, logistic regression, transformed item difficulties and DIF analyzes with EIRM, while one item contains DIF in the Breslow day method. In the standardization method, four items contain DIF at B level according to the Dorans, Schmitt & Bleistein (DSB) criterion. As a result of these results, researchers are recommended to detect DIF with at least two methods.

References

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Camilli, G., and Shepard, L. A. (1994). *Methods for identifying biased test items.* Sage Publications.

French, B. F., & Finch, W. H. (2015). Transforming SIBTEST to account for multilevel data structures. *Journal of Educational Measurement, 52*(2), 159-180. https://doi.org/10.1111/jedm.12071

Shepard, L. A., Camilli, G., and Williams, D. M. (1984). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22*, 77–105. https://doi.org/10.1111/j.1745-3984.1985.tb01050.x

TIMSS 2019 Fen ve Matematik başarısına etki eden faktörlerin açıklayıcı madde tepki modeli ve hiyerarşik genelleştirilmiş doğrusal modeller ile incelenmesi

Elif Özlem Ardıç ve Serap Büyükkıdık

Anahtar kelimeler: Hiyerarşik doğrusal modelleme (HLM), açıklayıcı madde tepki modeli, cinsiyet, madde türü, bilişsel alan

Giriş

Özellikle sosyal bilimlerde hiyerarşik verilerle sıklıkla karşılaşılmaktadır. Eğitim bilimlerinde öğrenciler; sınıf, okul, şehir, bölge, ülke gibi iç içe geçmiş bir yapı içerisinde bulunmaktadır. Bu hiyerarşik yapıdan ötürü gözlemlerin birbirinden tamamen bağımsız olması beklenemez. Bu tarzda yapılarda doğrusal regresyon modelleri yerine hiyerarşik doğrusal modeller tercih edilmelidir (Atar, 2010). Hiyerarşik doğrusal modeller, çeşitli tahminler, nedensel yorumlar ve veri indirgeme gibi çeşitli amaçlar için kullanılan regresyon yöntemlerinin hiyerarşik yapıda verilerde kullanımına izin veren genelleştirilmiş halidir (Raudenbush ve Bryk, 2002).

Birey ve madde özellikleri öğrencilerin performansını etkileyebilmektedir. "The Programme for International Student Assessment (PISA)" and "The Trends in International Mathematics and Science Study (TIMSS)" gibi uluslararası uygulamalarda madde türü, bilişsel alan gibi madde özellikleri ve cinsiyet gibi birey özellikleri bireylerin performanslarını etkileyebilen değişkenler olarak karşımıza çıkmaktadır. Geleneksel madde tepki kuramı yöntemleriyle tüm bu özellikler gözetilerek madde analizi yapılamaz. Açıklayıcı madde tepki modelleri geleneksel madde tepki kuramının bu sınırlılığını ortadan kaldırabilmektedir. Açıklayıcı madde tepki modelleri genelleştirilmiş doğrusal ve doğrusal olmayan modeller çerçevesinde, tanımlayıcı madde tepki modelleri olarak ele alınan geleneksel madde tepki kuramı modellerine, maddesel farklılıkları açıklamak üzere madde özelliklerinin ve/veya bireysel farklılıkları açıklamak üzere birey özelliklerinin eklenmesiyle elde edilmektedir (Atar, 2011). Açıklayıcı madde tepki kuramı modelleri ve hiyerarşik doğrusal modeller ile birey ve madde özellikleri gözetilerek çeşitli modeller kurulabilmektedir. Bu çerçevede TIMSS 2019 fen bilimleri ve matematik maddelerine verilen yanıtlar kullanılarak madde ve birey özelliklerinin öğrenci performansına etkisi hiyerarşik genelleştirilmiş doğrusal modeller ve açıklayıcı tepki modelleri ile incelenmiş ve elde edilen bulgular karşılaştırılmıştır. Hiyerarşik genelleştirilmiş doğrusal modeller kullanılarak gerçekleştirilen araştırmalar

incelendiğinde birey ve madde özelliklerinin birlikte ele alındığı modelleri kullanan tamamen benzer bir araştırmaya rastlanılmamıştır. Hiyerarşik genelleştirilmiş doğrusal modeller ve açıklayıcı tepki modellerini aynı veri üzerinde uygulayarak karşılaştırmalar yapan bir araştırma ile de karşılaşılmamıştır. Araştırmanın bu yönüyle ilgili alanyazına katkı sağlayacağının düşünülmesi araştırmayı yapmak için motivasyon kaynağı olmuştur.

Yöntem

Araştırmada, Uluslararası Matematik ve Fen Eğilimleri Araştırması (TIMSS) 2019 uygulaması kapsamında 8. sınıf düzeyindeki 26 ülke öğrencisinin kitapçık 1'e verdiği cevaplar kullanılmıştır. Kitapçık 1'de yer alan matematik ve fen bilimleri sorularını öğrenci cevaplayan 1951'i kız ve 1790'ı erkek olmak üzere 3741 öğrencinin cevapları kullanılmıştır. Kitapçık 1'de yer alan 44 matematik sorusunun 26'sı çoktan seçmeli ve 18'i yapılandırılmış madde türündedir. Soruların bilişsel düzeylerine göre dağılımı ise 13 soru bilme, 18 soru uygulama ve 13 soru akıl yürütme şeklindedir. Fen bilimleri kitapçığındaki soruların ise 28'i çoktan seçmeli ve 15'i yapılandırılmıştır. Kitapçıktaki soruların %34.9'u (15 soru) bilme, 32.6'sı (14 soru) uygulama ve %32.6'sı (14 soru) akıl yürütme bilişsel düzeyindedir. Çoklu puanlanan maddelerde doğru cevaplar 1, kısmi doğru ve yanlış cevaplar ise 0 şeklinde kodlanarak, iki kategorili hale getirilmiştir.

Bağımlı değişken iki kategorili olduğu için verilerin analizinde hiyerarşik genelleştirilmiş doğrusal modeller (HGLM) ve açıklayıcı madde tepki modelleri kullanılmıştır. Bu araştırmada HGLM analizi yapmak için HLM8 yazılımı kullanılmıştır. Açıklayıcı madde tepki modeli analizinde ise R Studio yazılımında lme4 (Bates ve diğ., 2015) paketinden yararlanılmıştır.

Hiyerarşik genelleştirilmiş doğrusal modellemede fen ve matematik verisi için tamamen koşulsuz model, koşulsuz düzey-2 modeli (madde türü ve bilişsel alan etkisi), koşulsuz düzey-1 modeli (cinsiyet etkisi) ve koşullu model (düzey-1 için madde türü, bilişsel alan ve düzey-2 için cinsiyet etkisi) analiz edilmiştir.

Açıklayıcı madde tepki modelleri kapsamında hem fen bilimleri hem de matematik maddeleri için madde özelliklerinden madde türü, bilişsel alan, madde türü x bilişsel alan etkileşimi ve birey özelliklerinden cinsiyet ve cinsiyet x madde türü x bilişsel alan etkileşimleri birlikte incelenmiştir.

Sonuçlar

Tamamen koşulsuz modele ait sabit etkilerin tahminine dayanarak elde edilen değerler, bireyin matematik ve fen bilimleri maddelerini doğru cevaplama olasılığının yanlış cevaplama olasılığından daha fazla olduğunu göstermektedir. Kestirilen ortalama bilişsel alan etkisi sıfırdan anlamlı olarak farklıdır ($p<$ .001). Bilişsel alan etkisine göre bireyler arasında gerçek farklılıklar bulunmaktadır. Matematik okuryazarlığı için çoktan seçmeli ve yapılandırılmış maddelerin cevaplama olasılıklarının eşit olduğu, fen bilimleri için hesaplanan odds değerinin 0.42 olması ise çoktan seçmeli maddelerin doğru cevaplanma olasılığının yapılandırılmış maddelerin 0.42 katı olduğu bulunmuştur. Madde türünün matematik ve fen

bilimleri cevapları üzerinde istatistiksel olarak anlamlı bir etkisi olmadığı bulunmuştur (p> .05). Matematik ve fen bilimleri için bilme bilişsel düzeyindeki çoktan seçmeli maddelerin doğru cevaplama olasılıklarının yanlış cevaplama olasılıklarından daha yüksek olduğu bulunmuştur. Erkeklerin maddeyi doğru cevaplama olasılığının her iki ders için kızlara göre daha yüksek olduğu bulunmuştur. Cinsiyetin bilişsel alan-cevap ve madde türü-cevap eğimleri üzerindeki etkisi ise her iki ders için de anlamlı bulunmamıştır. Cinsiyet kontrol altına alındıktan sonra bilme bilişsel düzeyindeki maddeler için madde türü etkisi ise istatiksel olarak anlamlı değildir (p> .05).

Açıklayıcı tepki modeli 1'de matematik ve fen bilimleri için çoktan seçmeli madde türü yapılandırılmış maddelere göre daha kolaydır. Her iki dersi için çoktan seçmeli ve yapılandırılmış maddeler arasındaki bu fark istatistiksel olarak anlamlıdır (p< .001). Model 2'de matematik dersi için bilme bilişsel alanındaki maddeler en kolay iken akıl yürütme alanındaki maddeler en zor maddelerdir. Fen bilimleri dersi için ise akıl yürütme bilişsel alanındaki maddeler en kolay, bilme düzeyindeki maddeler ise en zordur. Bilişsel alana göre katılımcıların cevaplarının istatistiksel olarak farklılaştığı bulunmuştur (p< .05). Model 3'te her iki ders için bilişsel alan ve madde türü etkileşimleri istatistiksel olarak anlamlıdır (p<.001). Model 4'te birey özelliklerinin etkisi incelendiğinde, matematik ve fen bilimleri maddelerinin kız öğrencilere göre erkek öğrenciler için daha kolay olduğu görülmektedir (p< .001). Model 5'te "*Cinsiyet x Madde Türü x Bilişsel Alan"* etkileşimleri incelendiğinde yalnızca "*Erkek x Yapılandırılmış x Akıl yürütme"* etkileşiminin istatistiksel olarak anlamlı olduğu görülmektedir (p< .001).

Kaynaklar

Atar, B. (2010). Basit doğrusal regresyon analizi ile hiyerarşik doğrusal modeller analizinin karşılaştırılması. *Journal of Measurement and Evaluation in Education and Psychology*, *1*(2), 78-84. https://dergipark.org.tr/tr/download/article-file/65987

Atar, B. (2011). Tanımlayıcı ve açıklayıcı madde tepki modellerinin TIMSS 2007 Türkiye matematik verisine uyarlanması. *Eğitim ve Bilim*, *36*(159), 255-269. http://egitimvebilim.ted.org.tr/index.php/EB/article/view/811/252

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. https:doi.org/10.18637/jss.v067.i01

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Sage.

Öğretmenler sınıf içinde uyguladıkları testleri oluştururken ve değerlendirirken dikkat ettikleri özelliklerin belirlenmesi

Nilgün Mısır ve Nuri Doğan

Anahtar kelimeler: Ölçme, değerlendirme, başarı testi, sınıf içi değerlendirme, madde yazarlığı, bilişsel seviye.

Giriş

Değerlendirme, eğitim sürecinin ayrılmaz bir parçası olması yanında, eğitim-öğretim sürecinin ne derece işlediğinin belirlenmesi yönünden de oldukça önemlidir. Ölçme ve değerlendirmenin asıl amacı, eğitim kalitesinin ölçülmesi ve iyileştirilmesidir. Bu doğrultuda yinelenebilir ve güvenilir değerlendirme yöntemlerinin kullanılması gerekmektedir. Okullardaki ölçme değerlendirme uygulamalarında öğretmenler tarafından hazırlanan sınavlar her ne kadar işlenen konuların anlaşılmasının tespitinde en uygun yol olarak görülse de, bu tarz sınavlarda "ölçme ve değerlendirme" kurallarının tam olarak işlediği konusunda endişeler vardır. Baştürk (2014) başarı testlerinin yanlış kullanılmasından dolayı ölçme hataları, değerlendirme problemleri ve puanlama hatalarının yaşandığını belirtmiştir.

Öğretmenler sınıf içinde uyguladıkları testleri oluştururken test maddelerinin psikometrik özellikleri, istatistiksel işlemler, test puanlarının geçerlik ve güvenirlik kanıtlarının iyi ele alınması testin amacına uygunluğu açısından çok önemlidir. Aynı zamanda öğrencilerin bazı merkezi sınavlara hazırlanırken karşısına çıkan çok sayıda soru içerikli kaynakların doğru şekilde ölçüp ölçmediği noktasında ders öğretmeninin rehber rolünde olması da faydalıdır. Bu tür durumlarda doğru madde ve doğru test olması adına öğretmenlerin alan uzmanlığının yanında bazı ölçme değerlendirme özelliklerinde de yeterli olması gerekir. Ancak internetten hazır kaynakların kullanımı, piyasada hızlı hazırlanan binlerce sorunun varlığı, soru yazma pratiklerinin ve teorik araştırmaların yeterince yapılmaması gibi durumlardan ötürü öğrencinin karşısına gelen ölçeklerin ölçme özelliğinin yeterli olmadığı durumlar söz konusu olabilmektedir. Bu araştırmada öğretmenlerin sınıf içinde uyguladıkları testleri oluştururken hangi madde özelliklerine dikkat ettikleri, hangi noktalarda zorlandıkları ve kendilerini nasıl geliştirmeyi düşündüklerini belirlemek amaçlanmıştır. Öğrencilerin öğrenme durumlarının doğru şekilde belirlenmesinde temel etkenlerden biri olan ölçeklerin doğru şekilde hazırlanması ve kullanılması, yine öğrencilerin karşılaştıkları birçok soru kaynağının amacına uygunluğunun ele alınması açısından öğretmenin ölçme değerlendirme açısından belirli yeterlikte olması

gerekmektedir. Bu bağlamda araştırmadan elde edilecek sonuçlar mevcut sıkıntıların belirlenmesi ve öğretmenlerin seçimlerinin değerlendirilmesi açısından önem arz etmektedir.

Araştırmanın genelleme evreni Türkiye genelindeki ilk, orta ve lise öğretmenleridir. Pandemi sürecinde yapılan bu araştırmada uygulama ve kişilere erişim güçlüğünden dolayı yakın çevre ve tanıdıklara ulaşılabilmiştir. Bu nedenle kolay/elverişli örnekleme yöntemi kullanılmıştır. Bu yöntemle çalışmanın örneklemi Trabzon, Rize ve Gümüşhane illerinde MEB'e bağlı okullarda görev yapan ve seçkisiz olarak belirlenen toplam 263 ilk, orta ve lise öğretmeni örneklem olarak belirlenmiştir.

Araştırmada veri toplama aracı olarak öğretmenlerin sınıf içinde uyguladıkları testleri oluştururken ve değerlendirirken maddelerin hangi özelliklerine dikkat ettiğini belirlemeye yönelik 22 maddelik anket hazırlanmıştır. Anketin hazırlık aşamasında öncelikle devlet okullarında her düzeyde görev yapmakta olan öğretmenlerin ölçme değerlendirme testlerini geliştirme ve değerlendirme sürecinde yaşadıkları durumlar hakkında görüşleri alınmıştır. Bunun yanı sıra Balcı ve Tekkaya (2000)'nın "Çoktan seçmeli madde tipleri ve fen eğitiminde kullanılan örnekleri" başlıklı araştırması, Kutlu ve diğ. (2010)'nın "İlköğretim programında yer alan kazanımlara dayalı soru yazma ve puanlama çalışması", Anıl ve Acar (2008)'ın "Sınıf öğretmenlerinin ölçme değerlendirme sürecinde karşılaştıkları sorunlara ilişkin görüşleri" gibi araştırma ile çalışmalar incelenmiştir. Oluşturulan bu maddeler arasından, ölçme ve değerlendirme ve program geliştirme bilim dalından iki akademisyen görüşü alınarak, ankette yer alması uygun maddeler seçilmiştir. Böylelikle demografik bölümle birlikte test hazırlarken dikkat edilen hususlara ilişkin toplam 22 maddelik ankete son şekli verilmiştir.

Ankette iki bölüm yer almaktadır; birinci bölümde demografik bilgiler, ikinci bölümde soru hazırlarken dikkat edilen hususlara ilişkin sorular bulunmaktadır. Öğretmenlerin sınıf içinde uyguladıkları testleri oluştururken ve değerlendirirken maddelerin hangi özelliklerine dikkat ettiğini belirlemeye yönelik uygulanan anketten elde edilen veriler SPSS 25 Programında betimsel analiz yapılarak, hesaplanan frekans ve yüzdeler tablo ve grafikler yardımıyla sunulmuştur.

Sonuçlar

Ankete katılan öğretmenlerin büyük çoğunluğunun (%95.4) sınavlarında orta güçlükte sorular kullandığını, çok zor ve çok basit soru kullanan öğretmen oranının manalı olmayacak kadar az olduğu tespit edilmiştir. Bir sorunun zor olmasının nedenine ilişkin bulgulardan öğretmenlerin çoğunun sorunun birden fazla kazanım içermesi, bununla birlikte sorunun da bilgi yığını dolayısıyla soru uzunluğu olduğunu düşündüğü sonucuna ulaşılmıştır. Katılımcı öğretmenlerin yarısından fazlası sınav içeriklerinin basit bilişsel seviyelerde olması, hazırlanma ve uygulanmanın kullanışlılığı açısından testlerde hatırlama ve kavrama bilişsel seviyelerinde sorular kullanmanın uygun olacağını belirtmiştir. Araştırmada öğrencilerin en fazla çoktan seçmeli, daha sonra doğru yanlış ve eşleştirme, boşluk doldurma, en az açık uçlu soru türünde sınav olmayı istedikleri belirlenmiştir. Yani öğrencilerin en fazla çoktan seçmeli en az da açık uçlu soruyu tercih etmiş olduğu söylenebilir. Yine öğretmenlerin sınavları değerlendirmesinde soruların türüne ve kazanımına çoğunlukla dikkat ettikleri belirlenmiştir. Çok azı bazı istatiksel işlemler

kullanırım şeklinde cevap vermiştir. Öğretmenler soru yazma sürecini kolaylaştırmak sıklıkla ölçme değerlendirme ile ilgili hizmet içi eğitimler almanın, soru inceleme çalışmalarına katılmanın, soru yazma pratikleri yapmanın, uzaktan eğitimlere katılmanın çok etkili olduğunu belirtmiştir.

Kaynaklar

Anıl, D. ve Acar, M. (2008). Sınıf öğretmenlerinin ölçme değerlendirme sürecinde karşılaştıkları sorunlara ilişkin görüşleri. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi, 5*(2), 44-61.

Aslan, C. (2011). Soru sorma becerilerini geliştirmeye dönük öğretim uygulamalarının öğretmen adaylarının soru oluşturma becerilerine etkisi, *Eğitim ve Bilim, 36*(160), 236-249.

Balcı, E. ve Tekkaya C. (2000) Ölçme ve değerlendirme tekniklerine yönelik bir ölçeğin geliştirilmesi. *Hacettepe Üniversite Eğitim Fakültesi Dergisi, 18*, 42 -50.

Baştürk, S. (2014). Ölçme araçlarının taşıması gereken nitelikler. S. Baştürk (Ed.), *Eğitimde ölçme ve değerlendirme* içinde (ss. 21-54). Nobel Akademi.

Çakan, M. (2004). Öğretmenlerin ölçme-değerlendirme uygulamaları ve yeterlik düzeyleri: İlk ve ortaöğretim. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi,* 37 (2), *99-114.*

Doğan, N. ve Atılgan, H. (2004). *Eğitimde ölçme değerlendirme* (5. baskı). Anı Yayıncılık.

Güler, N. (2012). *Eğitimde ölçme ve değerlendirme.* (4. baskı). Pegem Akademi.

Açıkgöz, M. ve Karslı, F. (2015). *A*lternatif ölçme-değerlendirme teknikleri kullanılarak iş ve enerji konusunda geliştirilen başarı testinin geçerlilik ve güvenirlik analizi. *Amasya Üniversitesi Eğitim Fakültesi Dergisi, 4*(1), 1-25.

Kahraman, P. (2014). *Öğretmen adaylarının ölçme-değerlendirme okuryazarlıklarının belirlenmesi ve mikro-öğretim yoluyla geliştirilmesi* (Tez No. 356353) [Doktora Tezi, Çanakkale Onsekiz Mart Üniversitesi Mart Üniversitesi]. Yükseköğretim Kurumu Tez Merkezi.

Kutlu, Ö., Yalçın, S. ve Pehlivan, E. B. (2010). İlköğretim programında yer alan kazanımlara dayalı soru yazma ve puanlama çalışması. *İlköğretim Online*, 9(3), 1201-1215.

Linn, R., and Gronlund, (1995). *Measurement and assesment in teaching* (7th ed.) Prentice-Hall.

Özgenç, M. (2013). Sınıf öğretmenlerinin alternatif ölçme ve değerlendirme bilgi düzeylerinin belirlenmesi. *Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi,* 2, 157-178. https://dergipark.org.tr/tr/pub/zgefd/issue/47941/606525

Popham, W. J. (2002). *Classroom assessment: What teachers need to know?* Allyn and Bacon.

Ülger, K. (2016). Öğrencilerin resim yapma becerilerinde gözlemlenen yaratıcılık ile yaratıcı düşünme becerileri arasındaki ilişki. Abant İzzet Baysal Üniversitesi *Eğitim Fakültesi Dergisi, 16*(4), 2023-2039. https://dergipark.org.tr/tr/pub/aibuefd/issue/28550/304609

Turan-Oluk, N. ve Ekmekci, G. (2017). Alternatif değerlendirme teknikleri ile geleneksel değerlendirme tekniklerinin öğrenci başarısını ölçme açısından karşılaştırılması, *JRES, 4*(2), 172-199.

Turgut, F. (1983). *Eğitimde ölçme değerlendirme* (2. baskı). Saydam Matbaacılık.

Tekindal, S. (2009) . *Okullarda ölçme ve değerlendirme yöntemleri* (2. baskı). Nobel Akademik Yayıncılık.

Tekin, H. (1996). *Eğitimde ölçme ve değerlendirme.* Yargı Yayınları.

Turgut, M. F. ve Baykul, Y. (2012). *Eğitimde ölçme ve değerlendirme* (4. baskı). Pegem Akademi.

Yaman, S. (2016). Çoktan seçmeli madde tipleri ve fen eğitiminde kullanılan örnekleri. *Gazi Eğitim Bilimleri Dergisi, 2*(2), 151-170. https://dergipark.org.tr/tr/pub/gebd/issue/35205/390659

Yıldırım Beyazıt Üniversitesi (2019). *Ölçme değerlendirme el kitabı.* Ankara. https://aybu.edu.tr/GetFile?id=c0bc7317-7fcd-491b-ac67-6753bcc7b22a.pdf

Bireyselleştirilmiş çok aşamalı testlerde farklı yönlendirme yöntemlerinin yetenek kestirimine etkisi

Hasibe Yahşi Sarı ve Hülya Kelecioğlu

Anahtar kelimeler: Bireyselleştirilmiş çok aşamalı testler, çok kategorili maddeler, yönlendirme yöntemleri, panel, modül

Öz

Araştırmanın amacı, bireyselleştirilmiş çok aşamalı test ortamında çok kategorili maddelerden oluşan testlerde bireylerin yetenek kestirimlerinin yönlendirme yöntemlerine göre nasıl değiştiğini incelemektir. Araştırma bir simülasyon çalışmasıdır. Araştırmada üç farklı kategori (3, 4 ve 5 kategorili maddeler), üç test uzunluğu (10, 20 ve 30 madde), iki panel deseni (1-2, 1-2-2) ve iki yönlendirme yöntemi (Maksimum Fisher Bilgisi [MFB] ve Rastgele) olmak üzere 3x3x2x2=36 koşul incelenmiştir. Madde havuzları farklı kategori sayısı için ayrı üretilmiş böylelikle üç farklı madde havuzu oluşturulmuştur. Her bir madde havuzu 200'er çok kategorili maddeden oluşmaktadır. Çok kategorili maddeler genelleştirilmiş kısmı puan modeline (GKPM) göre üretilmiştir. Simülasyon 1000 kişi ve 100 replikasyon olacak şekilde tasarlanmıştır. Analizler R programında yer alan *mstR* paketi ile yapılmıştır. Araştırma sonucunda ortalama mutlak yanlılık, RMSE ve korelasyon değerleri hesaplanmıştır. Araştırma sonucunda kategori sayısı ve test uzunluğu arttıkça ortalama mutlak yanlılık ve RMSE değerleri düşerken, korelasyon değerleri artmaktadır. Yönlendirme yöntemleri açısından, MFB ve rasgele yöntemleri benzer eğilimlerde olmasına rağmen MFI daha iyi sonuçlar vermektedir. 1-2 ve 1-2-2 panel desenleri arasında sonuçlar açısından benzerlik görülmektedir.

Açımlayıcı grafik analizi yönteminin çok kategorili maddelerde incelenmesi: Psikolojik-eğitsel araştırmalarda boyut sayısının kestirilmesi

Ezgi Mor Dirlik

Giriş

Psikolojik ve eğitsel ölçme araçları için gizil faktör ya da boyut sayısını kestirmek uzun zamandır tartışılan bir konu olmuştur ve bu alanlarda ölçme aracı geliştiren ya da geçerlik incelemeleri yapan araştırmacılar için oldukça önemli bir aşama olarak kabul edilmektedir (Timmerman ve Lorenzo-Seva, 2011). Gelişen bilgisayar teknolojileri sayesinde, yakın zamanda, çok sayıda istatistiksel yöntemler önerilmiş, önerilen karmaşık ve esnek modellemeler sayesinde çok boyutluluk, tek boyutluluk incelenmeleri sürdürülmektedir. Farklı yöntemlerin sayısındaki artış, faktör analitik yaklaşımların kullanım sıklığını da geçmiş on yıllara göre belirgin bir şekilde azaltmış olsa da faktör modellerini kullanmak yapı geçerleme sürecinde hala ilk adım olarak görülmektedir (Garrido ve diğ., 2016). Sıklıkla kullanılan açımlayıcı faktör analizinin bunca popülerliğine rağmen, hala araştırmacıların raporlamadan geçtiği ya da irdelemeden kabul ettiği yaklaşımlar bulunmaktadır. Bunlardan biri de kullanılan faktörleştirme tekniğinin belirtilmesidir.

Golino ve Epskamp (2017)'ın Science Direct üzerinden yaptıkları bir incelemede; 1990 ile 2016 yılları arasında sosyal bilimlere ilişkin yapılan 40.132 çalışma kapsamında AFA'nın kullanıldığını belirlemişlerdir. Bu çalışmalardan %22 ile %28'i arasında bir orana sahip çalışma kapsamında ise kullanılan faktörleştirme tekniğinin belirtilmediği tespit edilmiştir. Bu oranın oldukça ciddi bir oran olduğu araştırmacılar tarafından belirtilmiş ve bu çalışmalardan elde edilen sonuçlara şüphe ile yaklaşılmasına yol açtığı ifade edilmiştir. Kullanılan faktörleştirme tekniğine göre faktör sayısı değişiklik gösterebilmekte ve yapıların tanımlarında değişmeler gözlenebilmektedir. Bu nedenle faktörleştirme teknikleri ve hangi tercihin neden seçildiği, hangi durumlarda hangi tekniğin kullanımının daha uygun olduğu oldukça önemli konulardan biridir.

AFA kapsamında, ölçeğin boyut sayısının belirlenmesi için kullanılabilecek çok sayıda faktörleştirme tekniği önerilmiştir ve ancak geleneksel olarak iki temel yaklaşım bulunmaktadır. Özdeğerlere dayalı olarak hesaplanan ve belli durdurma kurallarına göre işleyen yöntemler ilk yaklaşımı oluşturmaktadır. Bu yaklaşımlardan en bilineni Kaiser-Guttman kuralıdır. Bu kurala göre özdeğeri 1'den

yüksek olan tüm boyutlar faktör olarak adlandırılır. Horn'un paralel analizi de bu yaklaşım içinde yer alır. Geleneksel faktör sayısı belirleme yaklaşımlarından bir diğeri ise model uyumuna dayalı olarak faktör sayısının kestirilmesini temel alır. Bayes bilgi kriteri ya da Velicer'in MAP(minimum average partial procedure) yöntemini temel alır. MAP yöntemi, birden çok temel bileşenler analizine bağlı olarak gerçekleşir ve kısmi korelasyonları kullanarak ortak varyansı maksimum yapan faktör sayısının belirlenmesini sağlar. Bayes yaklaşımı ise genişletilmiş Bayes bilgi kriteri (extended Bayesian information criterion- EBIC) ya da standart Bayes bilgi kriteri(Bayesian information criterion -BIC) olarak iki gruba ayrılabilir. Bunlara alternative olarak ise daha az çalışlan ve tercih edilen bir yöntem olan çok basit yapı (very simple structure -VSS) isimli bir yöntemde bulunmakta ve bu yöntem kapsamında orijinal korelasyon matrisinin basitleştirilmiş matris ile yeniden üretilip üretilemeyeceği incelenmektedir. Bunu yaparken de her madde için en yüksek faktör yük değeri korunurken, diğer faktörlerdeki yükleri 0' a eşitlenir (Keith ve diğ., 2016).

Belirtilen bu geleneksel yaklaşımdan farklı olarak grafiksel bir modele dayalı olan Açımlayıcı Grafik analizi (Exploratory Graph Analysis-EGA), Golino ve Epskamp (2017) tarafından geliştirilmiştir. Kümeleme analizine benzer şekilde grafiksel olarak çalışan bu yöntemde, yalnızca faktör sayısına ek olarak hangi maddelerin hangi faktörde oldukları da belirlenir. Yönlendirilmemiş ağ modeline bağlı olarak geliştirilen bu yöntemde veri setindeki değişkenler gruplanarak boyutlara ayrılır. Markov Rastgele Model Alanı olarak adlandırılan yönlendirilmemiş ağ modeline bağlı olan EGA yönteminde, Gaussian Grafik Modeli ve ters kovaryans matrisi kullanılarak modellemeler yapılır. EGA, öncelikli olarak gözlenen değişkenler arasındaki korelasyonu kestirir, LASSO adı verilen bir algoritma ile ters kovaryans matrisini hesaplar, son olarak da faktörler, ortak varyanslar ya da kümeleri kısmi korelasyon katsayılarına bağlı olarak kestirir. Kovaryans matrisine dayalı olarak ağ sistemleri ile çalışan bu yöntemde, faktörlerin netliğinin paralel analizle oldukça benzer düzeyde olduğu belirlenmiştir. Alan yazına yeni kazandırılan bu yöntemle özellikle faktörler arasında güçlü korelasyonların olduğu durumlarda faktör sayısının diğer faktörleştirme tekniklerine göre daha net belirlendiği görülmüştür (Epskamp ve diğ., 2017; Epskamp ve Fried, 2018). Birçok durumda farklı faktörleştirme tekniklerine benzer ya da daha net sonuçlar verdiği gözlenen EGA'nın farklı koşullar için yeni çalışmalarda incelenmesi geliştirenler tarafından da önerilmiştir. O nedenle bu çalışmada hem söz konusu yeni yöntemi kuramsal olarak inceleyip ilgilere tanıtmak hem de çok kategorili puanlanan maddelerden oluşan bir testin faktör sayısının belirlenmesinde EGA yönteminin performansını incelemek amaçlanmıştır. Çalışma kapsamında çok kategorili olarak puanlanan bir ölçeğe ilişkin, en fazla bilinen testlerden biri olan Scree-test ile EGA'nın kestirimlerini incelemek amaçlanmıştır.

Yöntem

Çalışma kapsamında son zamanlarda önerilmiş bir faktörleştirme tekniği olan EGA'nın çok kategorili puanlamanın yapıldığı durumda nasıl sonuçlar verdiği, Scree-test yöntemi ile kıyaslanarak incelenmiştir. Bu genel amaç çerçevesinde çalışmanın temel araştırmalar içinde yer alması uygun bulunmuştur.

Nicel yaklaşım benimsenerek yürütülen bu çalışmanın verileri, çok kategorili maddelerden oluşan ve lisans düzeyindeki öğrencilerin eleştirel düşünme standartlarını kullanma düzeyini öz-belirlemeli bir şekilde ölçmeyi amaçlayan bir ölçme aracı kullanılarak toplanmıştır. Söz konusu ölçme aracının seçilme nedeni, ölçme aracının faktör yapısının biliniyor olması, araştırmacılar geçerlik ve güvenirlik kanıtlarının titiz bir şekilde toplanmış olmasıdır. Faktör yapısı bilinen bir ölçeğin seçilme amacı yöntemler arasında kıyaslama yapmaktır. Çok kategorili puanlanan maddelerin seçilme nedeni, EGA yöntemini geliştirilenlerin, söz konusu yöntem için iki kategorili puanlanan maddelerde ve zekâ gibi maksimum performansa dayalı ölçmelerde bu yöntemi test etmiş olmalarıdır. Araştırmacılar EGA'nın çok kategorili puanlanan maddelerde de çalışması gerektiğini önermişlerdir.

Çalışmada kullanılan veri setinin öncelikli olarak analizlerin varsayımlarını sağlama durumu test edilmiştir. Bu incelemelerin ardından, toplanan veri setinin yapı geçerliğinin belirlenmesi için doğrulayıcı faktör analizi (DFA) yapılmıştır. DFA'nın ardından, EGA ve Scree-test analizlerine geçilmiştir. Veri setinin analizleri için R programında yer alan "*EGAnet"* paketi kullanılmıştır (Golino ve Christensen, 2020). EGA'nın faktör belirlemedeki performansı için çeşitli hata indeksleri hesaplanmıştır. Scree-test ile EGA'nın karşılaştırılması için testin orijinal faktör yapısı ve doğrulayıcı faktör analizinden elde edilen sonuçlar temel alınmıştır.

Sonuçlar

Çalışma kapsamında çok kategorili olarak puanlanan bir beceri testinin faktör yapısı EGA ve Scree-test yöntemi ile incelenecektir. İlk yapılacak analiz olan DFA'nın sonuçları incelenecek ve ölçeğin geliştiricileri tarafından önerilen modele benzer sonuçlar alınıp alınmadığı incelenecektir. Ardından EFA ve Scree-test yönteminden elde edilen sonuçlar karşılaştırılaşacak ve yöntemlerce önerilen faktörler her yönteme göre analiz edilecektir. İki yönteme göre oluşturulan faktörler, faktör yükleri, hesaplanan ortak faktör varyansları hem DFA sonuçları ile hem de birbirleri ile karşılaştırılarak analiz edilecektir. Çalışma sonucunda her iki yöntemden elde edilecek faktör yapılarının benzer olması ancak EGA yönteminin Scree- test metoduna göre daha ayrıntılı sonuçlar vermesi beklenmektedir.

Kaynaklar

Aybek, B, Aslan S., Dinçer, S. ve Coşkun-Arısoy, B. (2015). Öğretmen adaylarına yönelik eleştirel düşünme standartları ölçeği: Geçerlik ve güvenirlik çalışması. *Kuram ve Uygulamada Eğitim Yönetimi Dergisi*, *21*(1), 25-30. https://doi.org/10.14527/kuey.2015.002

Crawford, A. V., Green, S. B., Levy, R., Lo, W. J., Scott, L., Svetina, D., and Thompson, M. S. (2010). Evaluation of parallel analysis methods for determining the number of factors. *Educational and Psychological Measurement*, 70, 885–901. https://doi.org/10.1177/0013164410379332

Epskamp, S., Rhemtulla, M., and Borsboom, D. (2017). Generalized network pschometrics: Combining network and latent variable models. *Psychometrika*, *82*(4), 904–927. https://doi.org/10.1007/s11336-017-9557-x

Epskamp, S., and Fried, E. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods, 23*(4), 617–634. https://doi.org/10.1037/met0000167

Garrido, L. E., Abad, F. J., and Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via Monte Carlo simulation. *Psychological methods*, *21*(1), 93–111. https://doi.org/10.1037/met0000064

Golino, H. F., and Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PLoS One 12*(6), e0174035. https://doi.org/10.1371/journal.pone.0174035

Golino, H., and Christensen, A. P. (2020). EGAnet: Exploratory Graph Analysis -- A framework for estimating the number of dimensions in multivariate data using network psychometrics (version 0.9.4.). https://cran.r-project.org/package=EGAnet

Keith, T. Z., Caemmerer, J. M., and Reynolds, M. R. (2016). Comparison of methods for factor extraction for cognitive test-like data: Which overfactor, which underfactor? *Intelligence*, *54*, 37-54. https://doi.org/10.1016/j.intell.2015.11.003

Timmerman M. E., and Lorenzo-Seva U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16*(2), 209–220. https://doi.org/10.1037/a0023353

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*, 321–327. https://doi.org/10.1007/BF02293557

Velicer, W. F., Eaton, C. A., and Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin, and E. Helmes (Eds.), *Problems and solutions in human* assessment (pp. 41–71). Springer.

Yapı geçerliği çalışmalarında yöntem etkisinin doğrulayıcı faktör analiziyle incelenmesi

Mediha Korkmaz

Anahtar kelimeler: Yöntem etkileri, ilişkili özellik-ilişkili hatalar, ilişkili özellik-ilişkili yöntem, doğrulayıcı faktör analizi

Giriş

Psikoloji alanında olduğu gibi diğer sosyal ve davranış bilimleri alanlarında öz-değerlendirme biçiminde kullanılan ölçme araçları sıklıkla yöntem etkilerinden(method effects) kaynaklanan bazı sorunları barındırmaktadır. Bu tür ölçme araçlarında cevaplayıcıların maddeleri onaylama yönünde tepki yanlılığını azaltmak için ölçek/test maddelerin ifade yönünün pozitif ve negatif (wording items) olması psikometrik olarak önerilmektedir (Anastasi,1982). Ancak bu stratejinin uygulanması da yapı geçerliği ile ilgili çalışmalarda yöntem etkileri gibi bazı sorunlara neden olabilmektedir.

Yöntem etkisi genel anlamda araştırılan yapı ya da özellikten ziyada ölçme prosedürüyle bağlantılı olarak ortaya çıkan varyans olarak tanımlanmaktadır (Fan ve Lance, 2017; Lance ve diğ., 2010; Maul, 2013; Marsh ve diğ., 2010). Bu varyans aynı özellik/yapı hakkında farklı yöntem ya da yöntemleri uygulayarak bilgi edinilmeye çalışıldığında ölçme işleminin temel gerekliliklerinden değil de rastgele ortaya çıkabilmektedir. Yapı geçerliğinde aynı yapının farklı yöntemler aracılığıyla yakınsanması (convergent) çoklu yöntem, farklı özelliklerden ayrışması (disciriminant) çoklu özellik yaklaşımı Campbell ve Fiske'nin 1958 yılında geliştirdiği çoklu özellik- çoklu yöntem (MTMM) olarak bilinmektedir. Günümüzde klasik MTMM yaklaşımı Doğrulayıcı faktör analizi modelleri (DFA) kapsamında ilişkili özellik-ilişkili yöntem (Correlated trait-correlated method-CTCM) ve ilişkili özellik-ilişkili hatalar (correlated trait-correlated uniquness-CTCU) modelleriyle temsil edilmiştir (Marsh ve Bailey, 1991). İlişkili özellik-ilişkili hatalar modelinde yöntem etkisi hatalar arasındaki korelasyonlar ile temsil edilir ve genellikle de yakınsama (convergence) gösteriler (Marsh ve diğ., 2010). İlişkili özellik-ilişkili yöntemde ise yöntem ve özellik örtük değişkenlerinin doğrudan tahmin etmek ve yöntem varyansını hata varyansından ayırmak mümkündür.

Yöntem etkilerinin DFA kapsamında modellenmesinde ilişkili hatalar için 3 model ve ilişkili özellik- ilişkili yöntem için 3 model olmak üzere 6 farklı model söz konusudur. Bu modeller:

Model 1- tek özellik faktörü, pozitif yönlü maddelerin hataları ilişkili

Model 2- tek özellik faktörü, negatif yönlü maddelerin hataları ilişkili

Model 3- tek özellik faktörü, hem pozitif yönlü maddelerin hem de negatif yönlü maddelerin hataları ilişkili

Model 4- tek özellik faktörü ile negatif örtük yöntem faktörü

Model 5- tek özellik faktörü ile pozitif örtük yöntem faktörü

Model 6- tek özellik faktörü ile pozitif ve negatif örtük yöntem faktörleri

Söz konusu modeller inceleme altına alınan herhangi bir psikolojik yapının maddelerin ifade yönüne göre yöntem etkilerini kontrol altında tutmaya olanak sağlamaktadırlar. Model 6 aynı zamanda bifaktorü (iki-yönlü faktör) temsil etmektedir ve her bir madde tek bir özellik faktörüne (genel faktör) ve aynı zamanda madde ifade yönünü içeren iki yöntem faktörüne (grup faktörleri) bağlanmıştır. Model 4 ve model 5 madde ifade yönüne göre (wording items) yöntem etkisi örtük değişken faktöründeki gerçek puan varyansının saptanmasına olanak sağlayacaktır.

Bu çalışmanın temel amacı, yöntem etkilerini ilişkili hatalar ve ilişkili özellik-ilişkili yöntem modelleri kapsamında doğrulayıcı faktör analiziyle test etmektir. Bu doğrultuda madde ifade yönünün negatif ve pozitif olması bir yöntem etkisi olarak ele alınarak, söz konusu 6 model test edilecektir. Yöntem etkilerinin olduğu durumlarda bir ölçme aracından elde edilecek toplam puanların psikometrik olarak temsil ediciliği tartışılacaktır.

Araştırma Rosenberg Benlik Saygısı (RBS) ölçeğinin, pozitif ifade yönü ve negatif ifade yönü ile temsil edilen maddeleri üzerinde yürütülmüştür. Literatürde RBS ölçeğinin yapı geçerliği yöntem etkisi ilişkili hatalar, ilişkili özellik-ilişkili yöntem etkileri üzerinden pek çok araştırmada incelenmiştir (örneğin; Alessandri ve diğ., 2015; Marsh ve diğ., 2010; Tomas ve diğ., 2015).

Yöntem

Araştırmada kolaylıkla bulunabileni örnekleme tekniğiyle toplam 806 kişiye (K=439, E=366, eksik veri=1) ulaşılmıştır. Katılımcıların yaş ranjı 18 ve 75 arasında olup, yaş ortalaması 33.85 (S=12.96) 'tir. Bu çalışmada yöntem etkilerini incelemek üzere Rosenberg Benlik Saygısı-RBS (Rosenbeg Self-Esteem Scale,) ölçeği kullanılmıştır. Rosenberg (1963,1965) tarafından geliştirilen RBS ergen ve yetişkin bireylerde bütüncül benlik saygısının ölçümünde dünya çapında en sıklıkla kullanılan ölçme araçlarından biridir. RBS ölçeği. 5 pozitif ifadeli 5 negatif ifadeli toplam 10 maddeyi kapsamakta olup Likert formatında 4 dereceli (hiç uygun değil; 0-3; tamamen uygun) puanlanmaktadır. RBS tekboyutlu genel-bütüncül bir benlik saygısı özelliğini temsil eder ve tek bir toplam puan elde edilir.

Çalışmanın yöntem etkisi DFA modelleri LISREL 8.8 (Jöreskog ve Sörbom, 2006) programı kullanılarak yapılmıştır. Diğer analizlerde SPSS 17.00 aracılığıyla yürütülmüştür.

DFA modelleri test edilirken kovaryans ve asimptotik kovaryans matrisleri oluşturulmuş, tahmin yöntemi olarak güçlü maksimum olabilirlik (robust maximum likelihood) kullanılmıştır. Asimptotik kovaryans matrisinin de tahmin sürecine ilave edilmesi durumunda Satorra-Bentler χ^2 istatistiğin kullanılması önerilmektedir (Byrne ve Steward, 2006). Böylece model-veri uyumu indeksleri kapsamında Satorra-Bentler χ^2, RMSEA, SRMR, CFI, NFI ve GFI değerleri kriter olarak alınmıştır.

Yöntem etkisini kapsayan ilişkili hatalar (CU) 3 modelle, İlişkili özellik-ilişkili yöntem (CTCM) 3 modelle temsil edilmeden önce ayrıca ilk 3 DFA modeli oluşturulmuştur. Bu 3 model RBS maddelerinin tek bir örtük değişkene bağlandığı tekboyutlu model (hatalar ilişkisiz), pozitif ve negatif ifade yönlü maddelerin ayrı ayrı iki örtük faktöre bağlandığı faktörler arasında kovaryansın olmadığı model ve son olarak pozitif ve negatif ifade yönlü maddelerin ayrı ayrı iki örtük faktöre bağlandığı faktörler arasında kovaryansın olduğu modellerdir. Böylece toplamda 9 farklı model DFA test edilmiştir.

Sonuçlar

Test edilen ilk 3 model yöntem etkisini incelemek üzere seçilen örtük özelliğin(Benlik Saygısı) DFA sonuçlarıdır. Sonuçlarına göre RBS'nın tek boyutlu örtük özellik modeli ve iki örtük değişken ile temsil edilen ancak kovaryansın olmadığı model yetersiz model-veri uyumunu göstermiştir. Buna karşın negatif ve pozitif yönlü maddelerin kendi örtük değişkenlerine bağlandığı ve örtük değişkenler arasında kovaryansın olduğu model oldukça iyi model-veri uyumu göstermiştir(Satorra-Bentler χ^2 = 136.14, sd=34; RMSEA= .061; CFI= .98; NFI= .97; GFI= 0.96; SRMR= .041). Daha sonraki aşamada test edilen madde ifade yönüne göre ilişkili hatalar modellerinden tek bir örtük değişken ile temsil edilen ve pozitif yönlü maddelerin hataların ilişkilendirildiği model (Satorra-Bentler χ^2 = 72.14, sd= 25; RMSEA= .048; CFI= .99; NFI= .98; GFI= .98; SRMR= .027), negatif ifade yönlü maddelerin hatalarının ilişkilendirildiği modelden (Satorra-Bentler χ^2 = 126.05, sd=25; RMSEA= .071; CFI= 0.98; NFI= .97; GFI= .96; SRMR= .038) daha iyi uyum değerleri göstermiştir. Bu sonuçlar madde ifade yönü olarak maddelerin pozitif ve negatif ifade yönlerine sahip olmasını kendi hataları ile ilişkili olduğunu göstermektedir. İlişkili özellik –ilişkili yöntem olarak test edilen hem tek özellik ile pozitif ifade yönlü yöntem faktörü(Satorra-Bentler χ^2= 121.45, sd=30; RMSEA= .062; CFI=.98; NFI= 0.97; GFI= .96; SRMR= .035) hem de negatif ifade yönlü yöntem faktörü oldukça iyi model-veri uyumu değerleri göstermiştir. Bu sonuçlarda madde ifade yönünün yöntem etkisini göstermektedir.

Kaynaklar

Alessandri, G., Vecchione, M., Eisenberg, N., and Laguna, M. (2015). On the factor structure of the Rosenberg (1965) general self-esteem scale. *Psychological Assessment*, *27*(2), 621-635. https://doi.org/ 10.1037/pas0000073

Anastasi, A. (1982). *Psychological testing* (5th ed.). Macmillan.

Byrne, B. M., and Stewart, S. M. (2006). The MACS approach to testing for multigroup ınvariance of a second –order structure: A walk through the process. *Structural Equating Modeling, 13*(2), 287-321. https://doi.org/10.1207/s15328007sem1302_7

Fan, L., and Lance, C.E.(2017). A reformulated correlated trait-correlated method model for multitrait-multimethod data effectively ıncreases converge and admissibility rates. *Educational and Psychological Measuremet, 77*(6), 1048-1063. https://doi.org/ 10.1177/0013164416677144

Jöreskog, K. G., and Sörbom, D. (2006). *LISREL* (version 8.8) [Computer software]. Chicago: Scientific Softare International Inc.

Lance, E. C, Dawson, B., Birkelbach, D, and Hoffman, B. J. (2010). Method effects, measurement error, and substantive conludions. *Organizational Research Methods, 13*(3), 435-455. https://doi.org/10.1177/1094428109352528

Marsh, H. W., Scalas, I. F., and Nagengast, B. (2010). Longitudinal test of competing factor structures fort he rosenberg self esteem scale: Traits, ephemeral artifacts, and stable respeonse styles. *Psychological Assesment, 22*(2), 366-381. https://doi.org/10.1037/a0019225

Marsh, H. W., and Bailey, M. (1991). Confirmatory factor analyses of multitrait-multimethod data: A comparison of alternative models. *Appllied Psychological Measurement, 15*, 47–70. https://doi.org/10.1177/014662169101500106

Maul, A. (2013). Method effects and the meaning of measurement. *Frontiers in Psychology, 4*, Article169. https://doi.org/10.3389/fpsyg.2013.00169

Tomas, J. M., Oliver, A., Hontangas, P. M., Sanch, P. and Galiana, L. (2015). Method effects and gender invariance of the rosenberg self-esteem scale: A study on adolescents. *Acta De Investigación Psicológica, 5*(3), 2194-2203. https://doi.org/10.1016/S2007-4719(16)30009-6

Sıfır atama ve çoklu değer atama yöntemlerinin karakteristik eğri dönüştürme yöntemlerine etkisinin incelenmesi

Gülden Özdemir ve Burcu Atar

Anahtar kelimeler: Kayıp veri, sıfır atama, çoklu değer atama, test eşitleme, karakteristik eğri dönüştürme yöntemleri

Giriş

Sınavlar bireylerin hayatında bazı kritik kararların alınmasında önemli rol oynamaktadır. Bir kuruma personel seçme, görevde yükselme, unvan değişikliği, seviye belirleme, bir üst öğrenime öğrenci seçme vb. amaçla ulusal ya da uluslararası düzeyde gerçekleştirilen bu sınavlar arasında yılda birden çok kez (ALES, YDS, YÖKDİL, TOEFL vb.) ya da belirli döngülerle (TIMSS, PISA, PIRLS vb.) uygulanabilen sınavlar da yer almaktadır. Özellikle aynı amacı taşıyan ve tekrarlanan sınavlarda maddelerin güvenliğini sağlamak için farklı maddelerden oluşan farklı test formları geliştirilmektedir. Farklı test formlarından elde edilen puanların karşılaştırılabilirliğini sağlamak için bu formlar eşitlenmektedir (Cook ve Eignor, 1991). Test eşitleme sonucunda test formları birbirinin yerine kullanılabilir ancak her bir alternatif form için geçerlik kanıtlarının sunulması gerekmektedir. Test geçerliği hakkında soru işareti oluşturabilecek dolayısıyla bireyler hakkında kritik kararların alınmasında önemli bir etken de kayıp verilerdir (Hohensinn ve Kubinger, 2011). Sınavlarda maddelerden bazılarının cevaplanmaması ya da boş bırakılması sonucu kayıp veriler oluşmaktadır. Kayıp veriler, veri setinde daralmaya yol açabileceği gibi yapılacak kestirimlerin gücünün de zayıflamasına neden olacaktır. Ayrıca standart analiz yöntemleri tam veri setine göre hazırlanmakta olup kayıp veri setlerinde uygulanamamaktadır (Rubin, 1987). Bu nedenle kayıp verilerle farklı başa çıkma yöntemleri geliştirilmiştir. Ancak kayıp veriler, kayıp verilerle farklı başa çıkma yöntemlerinin kullanıldığı formlar üzerinde yapılan test eşitleme sonuçlarını da etkilemektedir (Ngudgratoke, 2009; Shin, 2009; Kim, 2015; Ertoprak, 2017). Farklı test formlarının hatasız ya da en az hata ile eşitlenebilmesi için en uygun kayıp veriyle başa çıkma yöntemi kullanılarak veri setinin çözümlenmesi gerekmektedir. Kayıp verilerle başa çıkma yöntemleri ya da test eşitleme üzerine çok sayıda çalışmaya rastlanmış ancak her iki kavramın da birlikte ele alındığı çalışmaların sınırlı olduğu görülmüştür. Bu nedenle bu çalışmada gerçek veri seti üzerinde; sıfır atama ve çoklu değer atama yöntemlerinin, farklı koşullar altında, MTK'ye dayalı test eşitleme yöntemlerinden karakteristik eğri dönüştürme yöntemlerine etkisinin incelenmesi

amaçlanmıştır. Test eşitleme yöntemlerinin performansları, hata kareleri ortalamasının karekköüne (RMSE) göre değerlendirilmiştir. Bu çalışmada ele alınan araştırma soruları şunlardır:

1. Sıfır atama yöntemi uygulanarak elde edilen test formları, karakteristik eğri dönüştürme yöntemlerine göre eşitlendiğinde, RMSE;

a) kayıp verilerin test formları içindeki yeri (her iki test, eşitlenecek test),

b) kayıp veri oranına (%10, %20)

göre nasıl değişmektedir?

2. Çoklu değer atama yöntemi uygulanarak elde edilen test formları, karakteristik eğri dönüştürme yöntemlerine göre eşitlendiğinde, RMSE;

a) kayıp verilerin test formları içindeki yeri (her iki test, eşitlenecek test),

b) kayıp veri oranına (%10, %20)

göre nasıl değişmektedir?

Yöntem

Araştırma verileri, Uluslararası Matematik ve Fen Eğitimleri Araştırması (TIMSS – Trend in International Mathematics and Science Study) 2019 uygulamasındaki veri setinden elde edilmiştir. Çalışma için TIMSS 2019 fen başarı dağılımında, bilgisayar tabanlı uygulama (eTIMSS) yapan, en başarılı ilk on ülke (Singapur, Tayvan, Güney Kore, Rusya, Finlandiya, Litvanya, Macaristan, Amerika Birleşik Devletleri, İsveç, Portekiz) seçilmiştir. İki kategorili puanlanan madde sayısı, ortak madde sayısı ve yanıtlayıcı sayısının en fazla olduğu 7 ve 8 numaralı kitapçıklara verilen yanıtlardan tam veriye sahip rastgele 1000'er kişilik örneklem seçilerek çalışma grupları oluşturulmuştur. Her iki kitapçıkta 13'ü ortak ve 12'si ortak olmayan madde olmak üzere toplam 25 iki kategorili puanlanan madde seçilmiştir. Ardından tamamen rastgele kayıp veri mekanizması altında veri silinerek eşitlenecek test ve her iki testte %10 ve %20 oranında kayıp veri içeren 4 farklı veri seti elde edilmiştir. Bu veri setlerinden sıfır atama ve çoklu değer atama yöntemi kullanılarak kayıp veri problemi çözülmüş 8 farklı çalışma grubu oluşturulmuştur. Oluşturulan çalışma grupları ile, denk olmayan gruplarda ortak test deseni kapsamında MTK'ye dayalı test eşitleme yöntemlerinden karakteristik eğri dönüştürme yöntemleri kullanılarak test eşitleme gerçekleştirilmiştir. Eşitleme sonuçlarının karşılaştırılması için hata kareleri ortalamasının karekökü (RMSE) değeri her koşul için ayrı ayrı hesaplanmıştır. Tüm analizler R yazılımı kullanılarak gerçekleştirilmiştir.

Sonuçlar

Çalışma sonucunda; sıfır atama yöntemi uygulanarak elde edilen test formları, karakteristik eğri dönüştürme yöntemlerine göre eşitlendiğinde, RMSE; kayıp veri oranı %10 olduğunda ve kayıp veriler her iki test formunda da olduğunda en düşük değeri göstermiştir. Çoklu değer atama yöntemi

uygulanarak elde edilen test formları, karakteristik eğri dönüştürme yöntemlerine göre eşitlendiğinde ise RMSE; kayıp veri oranı %10 olduğunda en düşük değeri göstermiştir. Kayıp veri oranı %10 olduğunda ve kayıp veriler her iki test formunda da olduğunda, kayıp veri oranı %20 olduğunda ve kayıp veriler eşitlenecek test formunda olduğunda RMSE'nin en düşük değeri gösterdiği bulgulanmıştır. Araştırmada ele alınan kayıp veri ile başa çıkma yöntemleri arasında genel olarak en düşük RMSE üreten yöntemin çoklu değer atama olduğu tespit edilmiştir. Ancak tam veri setine en yakın RMSE değerleri incelendiğinde; kayıp veri oranı %10 olduğunda sıfır atama yönteminin, kayıp veri oranı %20 olduğunda ise çoklu değer atama yönteminin daha yakın RMSE değerleri ürettiği gözlenmiştir.

Kaynaklar

Cook, L. L., and Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. *Educational measurement: Issues and Practice, 10*(3), 37-45.

Ertoprak, D. G. (2017). *Kayıp verinin test eşitlemeye etkisinin incelenmesi* (Tez No. 470015) [Doktora tezi, Hacettepe Üniversitesi]. Yükseköğretim Kurumu Tez Merkezi.

Hohensinn, C., and Kubinger K. D. (2011). On the impact of missing values on item fit and the model validness of the Rasch model. *Psychological Test and Assesment Modeling, 53*(3), 380-393. http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/07_Hohensinn.pdf

Kim, M. S. (2015). *Linking with planned missing data: Concurrent calibration with multiple imputation.* (Publication No. 10009452) [Doctoral dissertation, The University of Kansas]. ProQuest Dissertations & Theses Global.

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., and Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science.* TIMSS & PIRLS International Study Center.

Ngudgratoke, S. (2009). *An investigatıon of using collateral information to reduce equating biases of the post-stratification equating method* (Publication No. 3381312) [Doctoral dissertation, The University of Kansas]. ProQuest Dissertations & Theses Global.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* John Wiley & Sons.

Shin, S. H. (2009). How to treat omitted responses in Rasch model-based equating. *Practical Assessment Research & Evaluation, 14*(1), 1-8. https://doi.org/10.7275/x9vv-xg85

Bilişsel tanıya dayalı bilgisayar ortamında bireye uyarlanmış testlerde değişken test uzunluğu sonlandırma kuralına göre madde seçim algoritmalarının incelenmesi

Semih Aşiret ve Seçil Ömür Sünbül

Giriş

Birçok modern psikometrist, tekboyutlu örtük nitelikleri ölçecek testlere odaklanmaktadır (Wang, 2013). Bu testler, sıklıkla bireyleri seçme ve yerleştirme, bireylere burs verme gibi amaçlar doğrultusunda yapılan, sonuç odaklı değerlendirmelerde kullanılmaktadır. Eğitimin veya öğretimin sonunda, bireyin genel performansının değerlendirilmesinde kullanılan sonuç odaklı veya değer biçmeye dönük değerlendirmeler, öğrenci veya öğretmene, öğrenme veya öğretme stratejilerini düzeltmeleri için yeterli dönüt vermemektedir. DiBello ve Stout (2007), son yıllarda öğretmen ve yöneticilerin, küçük ve iyi tanımlanmış birçok bilgiyi ölçmek amacıyla test geliştirmeye önem verdiklerini belirtmektedir. Biçimlendirmeye yönelik veya süreç değerlendirme amacıyla yapılan ve uygulanan test sonuçları, öğrenme ve öğretmeyi desteklemek amacıyla kullanılmaktadır. Süreci değerlendirmek ve var olan eksikliklere hızlı çözüm getirmek amacıyla, bireylerin belirli konu alanındaki güçlü ve zayıf yanlarının hızlı bir şekilde ortaya konması gerekmektedir. Bilişsel tanıya dayalı bilgisayar ortamında bireye uyarlanmış test (BT-BOBUT) uygulamaları, bu durum için etkileyici sonuçlar sağlamaktadır. BT-BOBUT, bilişsel tanı (BT) ve bilgisayar ortamında bireye uyarlanmış test (BOBUT) uygulamalarının bir araya gelmesiyle oluşmuştur. BT-BOBUT uygulaması, bireyleri örtük durumlarına göre sınıflamayı ve bu örtük sınıflar üzerinde örtük sınıf modellerini uygulamayı amaçlamaktadır (Cheng, 2009). Daha genel bir ifadeyle, BT-BOBUT uygulamasının amacı, bireylere bireyselleştirilmiş tanısal dönüt vermektir. BOBUT uygulamalarında olduğu gibi, BT-BOBUT uygulamalarında da anahtar durum madde seçim algoritmalarıdır (Cheng, 2009; Zheng, 2015). BOBUT uygulamalarında, çeşitli ihtiyaçlara göre birçok madde seçim algoritması geliştirilmiştir ve bu algoritmalar iyi kurulmuş ve oturmuştur. Ancak gerek BT-BOBUT uygulamasının yeni ortaya çıkması, gerekse sonuçların doğruluğunu etkileyecek alt faktörlerin bulunmasından dolayı, BT-BOBUT uygulamalarıyla ilgili literatürde sınırlı sayıda çalışma bulunmaktadır. BT-BOBUT uygulamalarında kullanılan madde seçim algoritmalarının öncelikli ilgisi yüksek ölçme doğruluğu elde etmektir (Zheng, 2015). Bu amaç doğrultusunda, BT-BOBUT uygulamalarında, hangi koşullarda hangi madde seçim algoritmasının kullanılması gerektiğinin belirlenmesi önem taşımaktadır. Bununla birlikte, BT-BOBUT ile ilgili yapılan çalışmalarda farklı

madde seçim algoritmaları geliştirilmiştir. Ancak, literatürde ölçme doğruluğunu sağlamak amacıyla geliştirilmiş madde seçim algoritmalarının performanslarını, benzer koşullarda inceleyen bu kadar geniş kapsamlı çalışma yer almamaktadır. BT-BOBUT uygulamalarının daha çok biçimlendirmeye yönelik değerlendirmelerde ve ders içi ölçmelerde, öğrenciye hızlı dönüt vermek amacı taşıdığından, öğretmenin, öğretimin sonunda kısa sürede testi uygulayarak, öğrenciye tanısal dönütte bulunması gerekmektedir. Bu açıdan bakıldığında, kullanılacak algoritmaların hesaplama süreleri önem taşımaktadır.

Bu araştırmada, BT-BOBUT uygulamalarında madde seçim algoritmalarının değişken test uzunluğu sonlandırma kuralında nitelik sayısına ve madde kalitesine göre nitelik ve örüntü koruma oranları, hesaplama süresi, ve ortalama test uzunlukları açısından incelenmesi amaçlanmıştır.

Literatürde, bazı madde seçim algoritması (MPWKL, GDI, PWCDI, PWKL) dışındaki algoritmaların farklı koşullardaki hesaplama sürelerine ait bir çalışma yer almamaktadır. Bu çalışma, değişken test uzunluğu sonlandırma kuralına göre madde seçim algoritmaları ölçme doğruluğuyla birlikte hesaplama süresi açısından da değerlendirerek pratikte yapılacak çalışmalar için en doğru ve hızlı ölçme yapan algoritmaların tespit edilmesine katkı sağlayacaktır.

Bu çalışmayla, belirlenen koşullarda madde seçim algoritmalarının performansları incelenerek, çalışma kapsamında kullanılan madde seçim algoritmaları, değişimlenen faktörler ve bu faktörlerin düzeyleri çeşitlendirilerek alana özgül değer katacağı öngörülmektedir. Ayrıca, bu çalışmadan elde edilen bulguların pratikteki uygulamalar için araştırmacılara referans olacağı öngörülmektedir.

Yöntem

Bu çalışma, BT-BOBUT uygulamasında, madde seçim algoritmalarını çeşitli faktörlere göre inceleyen simülatif bir çalışmadır. Çalışmada yer alan faktörler aşağıda belirtilmiştir:

Çalışmada, nitelik sayısı 5 ve 6 olarak değişimlenmiştir. PWKL ve HKL (Cheng, 2009), MI (Wang, 2013), GDI ve MPWKL (Kaplan ve diğ., 2015), PWCDI ve PWACDI (Zheng ve Chang, 2016) ve JSD (Minchen ve de la Torre, 2016) madde seçim algoritmaları kullanılmıştır. Madde parametreleri düşük ayırtedicilik-düşük varyans için U(0,15, 0,25), düşük ayırtedicilik-yüksek varyans için U(0,10, 0,30), yüksek ayırtedicilik-düşük varyans için U(0,05, 0,15), yüksek ayırtedicilik-yüksek varyans için U(0,00, 0,20) olacak şekilde tekbiçimli dağılımdan üretilmiştir. Çalışmada, birinci en yüksek sonsal olasılık eşik değeri 0,80, ikinci en yüksek sonsal olasılık eşik değeri 0,10 olarak ele alınmıştır. Ayrıca maksimum test uzunluğu 40 olarak ele alınmıştır.

Çalışma kapsamında, değişimlenen faktörler ve bu faktörlerin düzeylerine göre veri üretimi ve verilerin analizi *R.3.6.1.* (R Core Team, 2019) programıyla gerçekleştirilmiştir. Çalışmada, 5 ve 6 nitelik sayısında 480 maddeden oluşan iki ayrı madde bankası üretilmiştir. Q matrisi, madde madde ve nitelik nitelik üretilmiştir. Q matrisindeki her madde ve nitelik için 0-1 arasında tekbiçimli dağılımdan bir sayı üretilmiş; üretilen sayı .30'dan küçükse 1, değilse 0 olarak ilgili Q matrisindeki hücreye kodlanmıştır. Ayrıca, her madde en az bir nitelik ölçecek şekilde sınırlandırma yapılmıştır. Her bireyin, her niteliği

%50 başarma şansına sahip olacak şekilde 3000 birey üretilmiştir. Belirlenen madde parametreleri ve Q matrisine göre 3000 bireye ait madde tepkileri ve DINA modele göre her bireyin her maddeyi doğru cevaplama olasılıkları hesaplanmıştır. Çalışmada, DINA model kullanılmıştır.

Çalışma kapsamında da ilk madde seçimi seçkisiz olarak yapılarak, diğer algoritmalarda sabit tutulmuştur. Bireylerin bilişsel örüntüleri MAP kestirim yöntemi kullanılarak kestirilmiştir. Madde seçim algoritmalarının değerlendirilmesinde, örüntü koruma oranı, ortalama hesaplama süresi ve ortalama test uzunluğu ölçütleri kullanılmıştır.

Sonuçlar

Madde ayırtediciliği ve madde varyansı arttıkça, maddelerin ortalama test uzunluğunun azaldığı, nitelik sayısı arttığında ise, ortalama test uzunluklarının arttığı sonucuna ulaşılmıştır. Madde seçim algoritmaları, ortalama hesaplama süreleri açısından değerlendirildiğinde, GDI algoritmasının ortalama hesaplama süresi sabit testlerde olduğu gibi, değişken test uzunluğu sonlandırma kuralında da en düşük olduğu sonucuna ulaşılmıştır. Nitelik sayısının artmasıyla, ortalama test uzunluğu ve ortalama hesaplama süresi arttığından, çok uzun nitelik sayısından kaçınılması veya maddenin ölçeceği maksimum nitelik sayısına sınırlandırma getirilmesi önerilmektedir. Yüksek ayırtedicilik-yüksek varyans madde kalite düzeyinde, algoritmaların ortalama test uzunlukları 5 nitelik koşulunda, 5-7 ve 6 nitelik koşulunda 6-8 aralığındadır. Bu sebeple, sınıf içi ölçmelerin yapılacağı çalışmalarda madde kalitesinin yüksek olması, BT-BOBUT uygulamasının etkili kullanılmasını sağlayacağı düşünülmektedir. Bununla birlikte, tüm koşullarda JSD algoritmasının ortalama test uzunluğu en kısadır. Ancak, madde kalitesi düşük olduğunda, JSD algoritmasının ortalama hesaplama süresi MPWKL algoritması dışındaki diğer algoritmalara göre yüksektir. Bu açıdan, madde kalitesi yüksek olduğu durumlarda JSD algoritmasının, madde kalitesi düşük olduğunda ise, hesaplama süreleri de dikkate alındığında JSD algoritmasının yanı sıra, GDI ve MI algoritmalarının da kullanılması önerilmektedir.

Kaynaklar

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*, 619–632. https://doi.org/10.1007/s11336-009-9123-2

DiBello, L., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao ve S. Sinharay (Eds.), *Handbook of Statistics* (Volume 26, pp. 979-1030). Elsevier.

Kaplan, M., de la Torre, J., and Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement, 39*, 167–188. https://doi.org/10.1177/0146621614554650

Minchen, N. D., & de la Torre, J. (2016, July 11-15). *The continuous G-DINA model and the Jensen-Shannon divergence* [Paper presentation]. 81st International Meeting of the Psychometric Society (IMPS), Asheville, NC, USA.

R Core Team (2019). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement, 73*(6), 1017–1035. https://doi.org/10.1177/0013164413498256

Yigit, H. D., Sorrel, M. A., and de la Torre, J. (2019). Computerized adaptive testing for cognitively based multiple-choice data. *Applied Psychological Measurement, 43*(5), 388–401. https://doi.org/10.1177/0146621618798665

Zheng, C. (2015). *Some practical item selection algorithms in cognitive diagnostic computerized adaptive testing—Smart diagnosis for smart learning* (Publication No. 3738011) [Doctoral Dissertation, University of Illinois at Urbana]. ProQuest Dissertations & Theses Global.

Zheng, C., and Chang, H. H. (2016). High-efficiency response distribution–based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement, 40*, 608-624. https://doi.org/10.1177/0146621616665196

PISA 2018 okuma becerileri alt testinin Mantel-haenszel, SIBTEST ve lojistik regresyon yöntemleri ile değişen madde fonksiyonu açısından incelenmesi

Özge Erdoğan ve Hakan Yavuz Atar

Anahtar kelimeler: PISA 2018, değişen madde fonksiyonu, Mantel-Haenszel, lojistik regresyon, SIBTEST

Giriş

Uluslararası düzeyde yapılan uygulamalardan biri olarak değerlendirilen PISA, Ekonomik İşbirliği ve Kalkınma Örgütü (OECD) tarafından hazırlanıp üç yıl ara ile yapılan, 15 yaş grubuna dâhil olan öğrencilerin kazandıkları bilgi ve becerileri değerlendiren uluslararası düzeyde gerçekleştirilen en büyük eğitim araştırmalarından biridir (MEB, 2019). Uluslararası çalışmalar, ülkelerin eğitim sistemlerini karşılaştırmak, değerlendirmek ve geliştirmek için geniş bir bakış açısı sağlamaktadır. Bu nedenle, ülkeler bu sınavlara önem vererek mevcut durumlarının ortaya koymakta ve uluslararası rekabet ortamında konumlarını diğer ülkelere göre kolayca kıyaslayabilmektedirler (Gök ve diğ., 2014).

Uluslararası düzeyde düzenlenen uygulamalarda, uygulamaya katılan bireyler çeşitli özellikler bakımından birbirlerinden ayrılmaktadır. Sınav dilinin farklı olması, farklı demografik, sosyoekonomik, kültürel özelliklere sahip olma, cinsiyet gibi özellikler uygulamaya katılan bireylerin performanslarını etkilemektedir (Asil ve Gelbal, 2012; Ercikan, 1998). Farklı kültür ve dil gruplarını katılımcı olarak barındıran geniş ölçekli bir uygulamanın sonuçlarına dayanılarak, ülkelerin eğitim sistemleri hakkında doğru yorumlar ve ülkeler arası karşılaştırmalar yapabilmek için ölçülen yapının gruptan bağımsız olması gerekir. Gruptan bağımsız olmak, aynı yetenek seviyesindeki katılımcıların farklı alt gruplardan gelseler dahi test puanlarının eşit olması anlamını ifade eder (Osterlind, 1983; Hambleton ve Rogers,1996; Zumbo, 1999). PISA uygulamasında ölçülecek yapının gruptan bağımsız olması koşulu sağlanamazsa yapılacak karşılaştırmalarda ortaya çıkan farklılık gerçek durumdan mı yoksa yapının gruptan gruba farklılık göstermesinden mi kaynaklandığı bilenemeyecektir (Somer ve diğ., 2009). Bu nedenle farklı kültür ve dile uyarlanan PISA uygulamasında ölçme aracının geçerliği, sonuçlara karışan hata miktarı, maddelerin yanlılığı gibi konular önem kazanmaktadır (Gök ve diğ., 2014).

Bir testin geçerliğini olumsuz yönde etkileyen unsurların başında yanlılık gelir (Clauser & Mazor, 1998). Yanlılık, aynı yetenek seviyesinde yer alan fakat farklı alt gruplardan gelen bireylerin bir maddeyi doğru yanıtlama olasılıklarının farklı olması olarak tanımlanmaktadır (Osterlind, 1983; Hambleton ve Rogers,1996; Zumbo, 1999; Raju ve Ellis, 2002).

Testi alan bireylerin eşit fırsatlara sahip olup olmadıklarını belirlemenin ilk adımı, değişen madde fonksiyonunu (DMF) ortaya çıkmaktadır (Camilli ve Shepard, 1994). Farklı gruplara ait puanların karşılaştırılabilir olması çoğunlukla DMF analizleri ile değerlendirilir (Ercikan ve diğ., 2004). DMF çalışmaları, yanlılığın istatistiksel bir göstergesi olarak düşünülebilir (Angoff, 1993; Zumbo ve Gelin, 2005). Yanlılık araştırmalarında yapılacak ilk iş, testte yer alan maddelere istatiksel analizler uygulayarak DMF içerip içermediklerinin tespit edilmesidir. Bir maddenin yanlı olduğunu ileri sürebilmek için öncelikle o maddenin DMF içermesi beklenir. İstatiksel olarak DMF gösteren maddeler belirlendikten sonra bu maddelerin farklılık gösterme kaynakları belirlenmelidir. DMF'nin varlığı alt gruplar arasındaki gerçek farklılıktan ya da madde yanlılığından meydana geliyor olabilir (Camilli ve Shepard, 1994).

Bir testte yer alan maddelerin DMF içerip içermediğini ortaya çıkarmayı sağlayan birçok yöntem bulunmaktadır. Ancak bu yöntemler, maddeleri DMF'li olarak belirlerken izledikleri farklı matematiksel süreçler, kullandıkları farklı algoritmalar ve kesme noktaları nedeniyle tam bir uyum içinde değillerdir (Doğan ve Öğretmen, 2008; Gök ve diğ., 2010; Alatlı ve Şenel, 2020; Bakan-Kalaycıoğlu ve Berberoğlu, 2010). Bu nedenle, DMF'yi belirlemek için güçlü bir istatistiksel yöntem üzerinde ortak bir fikre varılamamıştır. Bu nedenle yanlılık incelemelerinde birden fazla DMF belirleme yöntemlerinin bir arada kullanılması ve elde edilen sonuçların karşılaştırılarak karar verilmesi önerilmektedir (Hambleton, 2006).

Bu çalışmada, uluslararası eğitimsel değerlendirme çalışmalarından biri olan PISA 2018 uygulamasında yer alan Okuma Becerileri alt testinin farklı kültür ve dile göre değişen madde fonksiyonu gösterip göstermediğinin üç farklı yöntemle incelenmesi amaçlanmıştır. PISA uygulamasında ölçme araçları İngilizce ve Fransızca olmak üzere orijinal olarak iki dilde hazırlanmakta ve daha sonra katılımcı ülkelere kendi ulusal dillerine çevirmeleri için gönderilmektedir (OECD, 2007). Bu nedenle bu çalışmada, kıyaslama yapabilmek adına PISA uygulamasını kendi ulusal dilinde alan Birleşik Krallık ve Fransa örneklemleri ile kendi ulusal dilinde bir çeviriye dönüştürerek uygulayan Türkiye örneklemi tercih edilmiştir. MH, LR ve SIBTEST yöntemleri kullanılarak yapılan analizlerde, DMF ile ilgili elde edilen sonuçların birbiriyle uyumu incelenmiştir.

Yöntem

Bu araştırmada, PISA 2018 okuma becerileri alt testinde bulunan maddelerin farklı kültür ve dil değişkenine göre DMF gösterip göstermediği incelenmektedir. PISA 2018 okuma becerileri alt testi verileri üç farklı yöntemle (SIBTEST, Lojistik Regresyon, Mantel-Haenszel) DMF analizine tabi tutulmuş ve sonuçlar betimlenmiştir. Bu amaç doğrultusunda betimsel yöntemden yararlanılmıştır.

PISA 2018 uygulamasında bireyselleştirilmiş test deseni uygulanmıştır. Bu nedenle bu çalışmada yapılan analizlerde, okuma becerileri alt testinde Temel bölüm, 1.Aşama ve 2.Aşama bölümlerinde aynı

demette yer alan madde paketleri ele alınmıştır. Temel bölüm için 1.paket (Core RC1), 1.Aşama için 1.paket (Stage 1 - R11H), 2. Aşama için 1. paket (Stage 2 - R21H) seçilmiştir. Seçilen madde paketlerinde toplamda 39 madde yer almaktadır. Ancak madde paketlerinde 1-0 şeklinde iki kategorili olarak puanlanmayan 4 madde analize dahil edilmeyecekleri için veri setinden çıkarılmış ve toplam madde sayısı 35 olarak belirlenmiştir.

Çalışma grubu olarak ise PISA uygulamasını kendi ulusal dilinde alan Birleşik Krallık ve Fransa örneklemleri ile kendi ulusal dilinde bir çeviriye dönüştürerek uygulayan Türkiye örneklemi tercih edilmiştir. Yukarıda belirtilen madde paketlerini yanıtlayan Türkiye'den 125, Birleşik Krallık'tan 326 ve Fransa'dan 143 öğrenci olmak üzere toplamda 594 öğrenci çalışma grubuna dâhil edilmiştir. Çalışma grubunda yer alan ülkeler; Türkiye – Birleşik Krallık, Türkiye – Fransa ve Fransa – Birleşik Krallık şeklinde ikişerli gruplarda analiz edilmiştir.

Verilerin analizine geçilmeden önce veri seti, kayıp veri ve uç değerler bakımından incelenmiştir. Yapılan kayıp veri analizi sonucunda, Little ve Rubin (2002) tarafından yapılan sınıflandırmaya göre veri seti içerisinde yer alan kayıp verilerin rastlantısal olarak dağılmadığı görülmüştür. Bu noktada, kayıp verileri gidermek amacıyla, kayıp verilerle baş etme yöntemlerinden çoklu atama (multiple imputation) yöntemi kullanılarak her bir örneklem için tam bir veri seti elde edilmiştir. Tek değişkenli uç değer analizleri sonucunda ise, ±3 aralığını -3,02 z puanı ile aşan tek bir değere rastlanılmıştır. Birleşik Krallık örnekleminde yer alan bu değer, veri setinden çıkarılmayarak yapılacak olan analizlere dâhil edilmiştir.

Sonuçlar

PISA 2018 okuma becerileri alt testinde yer alan maddelerin Türkiye - Birleşik Krallık, Türkiye - Fransa, Fransa - Birleşik Krallık gruplarında MH, LR ve SIBTEST yöntemleriyle DMF analizleri gerçekleştirilmiştir. PISA 2018 uygulamasını kendi ulusal dilinde bir çeviriye dönüştürerek uygulayan taraf olması nedeniyle Türkiye dezavantajlı grup olarak kabul edilmiştir. Bu nedenle analizlerde Türkiye odak grup, Fransa ve Birleşik Krallık ise referans grup olarak belirlenmiştir.

PISA 2018 okuma becerileri alt testinde yer alan maddelerin Türkiye ve Birleşik Krallık örnekleminde MH yöntemiyle analiz edildiğinde 14, LR yöntemiyle analiz edildiğinde 18, SIBTEST yöntemiyle analiz edildiğinde ise 11 maddede DMF tespit edilmiştir. Türkiye ve Fransa örneklemi, MH yöntemiyle analiz edildiğinde 6, LR yöntemiyle analiz edildiğinde 8, SIBTEST yöntemiyle analiz edildiğinde ise 5 maddede DMF tespit edilmiştir. Fransa ve Birleşik Krallık örneklemi MH yöntemiyle analiz edildiğinde 9, LR yöntemiyle analiz edildiğinde 13, SIBTEST yöntemiyle analiz edildiğinde ise 8 maddede DMF tespit edilmiştir.

Kaynaklar

Alatlı, B. A. ve Şenel, S. (2020). Değişen madde fonksiyonunun belirlenmesinde "difR" r paketinin kullanımı: Ortaöğretime geçiş sınavı fen alt testi. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, *53*(3), 865-901. https://doi.org/10.30964/auebfd.684727

Angoff, W. (1993). Perspective on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–24). Lawrence Erlbaum Associates.

Asil, M. ve Gelbal, S. (2012). PISA öğrenci anketinin kültürler arası eşdeğerliği. *Eğitim ve Bilim*, *37*(166), 236-249. http://egitimvebilim.ted.org.tr/index.php/EB/article/view/1501/462

Bakan-Kalaycıoğlu, D., and Berberoğlu, G. (2010). Differential item functioning analysis of the science and mathematics items in the university entrance examinations in Turkey. *Journal of Psychoeducational Assessment, 29*(5), 467-478. https://doi.org/10.1177/0734282910391623

Camilli, G., and Shepard, L.A. (1994). *Methods for identifying biased test items.* Sage.

Clauser, B. E., and Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31-47.

Doğan, N., and Öğretmen, T. (2008). Değişen madde fonksiyonunu belirlemede Mantel-Haenszel, ki-kare ve lojistik regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim*, *33*(148), 100-112.

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research*, *29*(1998), 543-553.

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., and Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*(3), 301-321.

Gök, B., Atalay Kabasakal, K. ve Kelecioğlu, H. (2014). PISA 2009 öğrenci anketi tutum maddelerinin kültüre göre değişen madde fonksiyonu açısından incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *5*(1), 72-87.

Gök, B., Kelecioğlu, H. ve Doğan, N. (2010). Değişen madde fonksiyonunu belirlemede Mantel-Haenzsel ve lojistik regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim, 35*(156), 3-16.

Hambleton, R. K. (1996, April 9-11). *Guidelines for adapting educational and psychological tests* [Conference presentation]. the Annual Meeting of the National Council on Measurement in Education, New York.

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care*, *44*(3), 182-188.

Milli Eğitim Bakanlığı (2019). *PISA 2018 Türkiye ön raporu.* http://www.meb.gov.tr/meb_iys_dosyalar/2019_12/03105347_PISA_2018_Turkiye_On_Raporu.pdf sayfasından erişilmiştir.

OECD (2007). *PISA 2006: Science competencies for tomorrow's world: Volume 1 analysis,* Paris: OECD Publications. https://doi.org/10.1787/9789264040014-en

Osterlind, S. J. (1983). *Test item bias* (No. 30). Sage.

Raju, N. S., and Ellis, B. B. (2002). Differential item and test functioning. In F. Drasgow, and N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 156–188). Jossey-Bass.

Somer, O., Korkmaz, M., Dural, S., ve Can, S. (2009). Ölçme eşdeğerliğinin yapısal eşitlik modellemesi ve madde cevap kuramı kapsamında incelenmesi. *Türk Psikoloji Dergisi*, *24*(64), 61-75.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): ogistic Regression Modeling as a Unitary Framework for Binaryand Likert-Type (Ordinal) Item Scores.* Directorate of Human ResourcesResearch and Evaluation.

Zumbo, B. D., and Gelin, M. N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological/community moderated (or mediated) test and item bias. *Journal of Educational Research and Policy* Studies*, 5*(1), 1-23.

Aktif öğrenmeye dayalı test hazırlama etkinliklerinin öğretmenlerin ölçme ve değerlendirmenin programdaki durumuna yönelik tutumlarına etkisi

Merve Yıldırım-Seheryeli, Burcu Gürkan ve Ufuk Akbaş

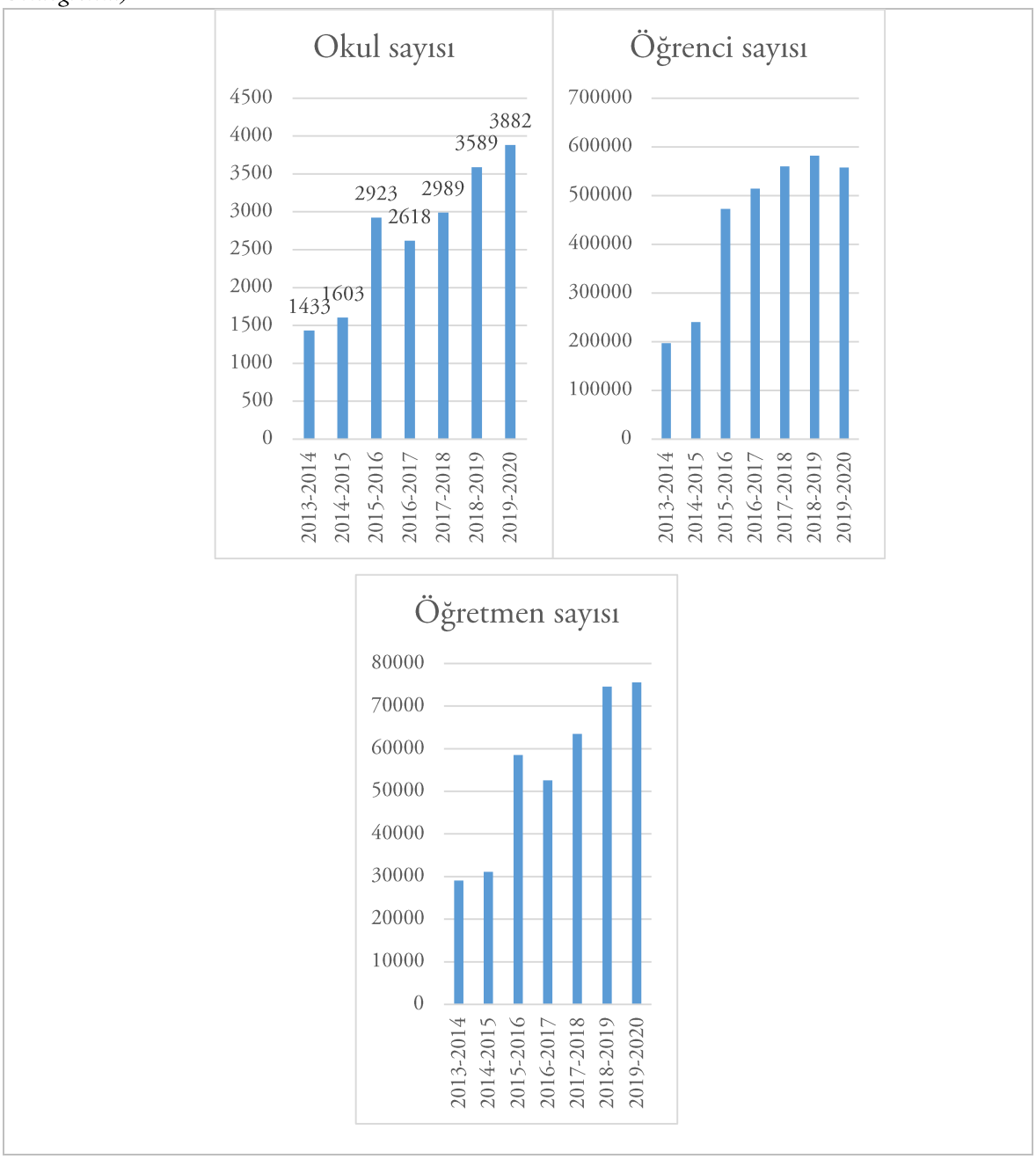
Anahtar kelimeler: Aktif öğrenme, test hazırlama, ölçme ve değerlendirmenin programdaki durumuna yönelik tutum

Giriş

Diğer ülkelerde olduğu gibi ülkemizde de dershanelerin varlığı, özellikle fırsat eşitsizliği yaratması (Garipağaoğlu, 2016; Özoğlu, 2011) sınav odaklı bir sistemde dershanelerin okulların yerini alması ve okulların temel işlevini yerine getirmesini zorlaştırması (Garipağaoğlu, 2016; Gümüş, 2014) gibi nedenlerle uzun yıllar tartışma konusu olmuştur. Nitekim 2014 yılında "Milli Eğitim Bakanlığı, Özel Öğretim Kurumları Yönetmeliğinde Değişiklik Yapılmasına Dair Yönetmelik" (Resmi Gazete, 2014) kapsamında dershanelerin belirli kurallar çerçevesinde kapatılarak en geç 2015 Eylül ayına kadar temel liselere dönüştürülmesi kararı verilmiştir. MEB Eğitim istatistikleri (2020) raporuna göre 2013-2020 yılları arasında ülkemizdeki özel öğretim kurumlarının sayıları Şekil 1'de verilmiştir. Beklendiği gibi okul sayısındaki büyük artış 2015-2016 eğitim öğretim yılında gerçekleşmiştir.

Şekil 1

Eğitim Öğretim Yıllarına Göre Özel Öğretim Kurumları Okul, Öğrenci vee Öğretmen Sayıları (Lise / Ortaöğretim)



MEB'e bağlı eğitim kurumlarına öğretmen olarak atanacakların "Yükseköğretim kurumlarında genel kültür, özel alan ve pedagojik formasyon eğitimi almış olma (ya da adaylık döneminde tamamlama)" şartını sağlamaları gerekmektedir (MEB, 2021). Özel öğretim kurumlarında ise eğitim personeli öğretmen, uzman öğretici ve usta öğretici olarak farklılaşmaktadır. Usta öğretici *"En az ortaöğretimden mezun olmuş ve alanında öğrenim gördüğünü belgelendiren veya alanında sertifika sahibi*

olan öğreticiler", uzman öğretici ise *"Görevlendirileceği alanda yükseköğrenim görmüş öğreticiler"* şeklinde tanımlanmıştır. Öğretmen olarak görevlendirileceklerden ise eğitim fakültesi mezunu olmayanlardan "*ortaöğretim alan öğretmenliği tezsiz yüksek lisans ya da pedagojik formasyon programı başarı belgesi örneği*" istenmektedir. Bu durum usta ve uzman öğretici için geçerli değildir (Özel Öğretim Kurumları Yönetmeliği, 2008).

Özel dershanelerden resmi okullara atanan fen bilimleri öğretmenleri ile yapılan bir çalışmada (Çetin ve Sadık, 2020) öğretmenlerin mesleki yeterlik algılarının yüksek olduğunu fakat bu algının dersi planlama, kullanılan yöntem-teknik, materyal kullanma ve hazırlama konularında yaptıkları etkinliklerde örtüşmediğini, öğretmenlerin öğrencileri sınavda başarılı olacak şekilde yetiştirmeleri gerektiği ile ilgili kaygı yaşadıkları için soru çözerek ders anlatmayı tercih ettikleri görülmüştür.

Ceylan ve diğ. (2020) tarafından hazırlanan TALIS 2018 raporuna göre, katılımcılardan "alan eğitimi sırasındaki öğretmen eğitimi" ile öğretmenlik yeterliğini kazandığını belirten öğretmenlerin oranı OECD ortalamasının üstünde iken "öğretmenlik eğitim sürecinin kapsamı"na geldiğinde ise "öğretilen konuların/derslerin içeriği ve sınıf içi uygulamasının eğitim sürecinde yer aldığını belirtenlerin oranı OECD ortalamasının altındadır. Yine bu rapora göre öğretmenler, mesleki gelişim faaliyetlerinde en az "öğrenci değerlendirmesinin analizi ve kullanımı" en çok ise "müfredat bilgisi" konularına yer verildiğini belirtmişlerdir. Buna rağmen ihtiyaç duydukları mesleki gelişim konularında "müfredat bilgisi" diğer "alan bilgisi" ve "pedagojik yeterlikler" başlıkları ile benzer düzeyde ihtiyaç olduğunu belirtmişlerdir. Beklendiği gibi en fazla ihtiyaç yine "öğrenci değerlendirmesinin analizi ve kullanımı" başlığındadır. Bu çalışmada "öğrenci değerlendirmesinin analizi ve kullanımı" konusunda mesleki gelişimi sağlamak için müdahale programı olarak "test hazırlama etkinlikleri" kullanılmıştır.

Alanyazın incelendiğinde öğretmenlerin mesleki gelişimleri için ölçme ve değerlendirmeden alan bilgisine, müfredat bilgisinden eğitim durumlarının planlanmasına birçok konuda hizmet içi eğitimlerin verilmesini öneren çok sayıda çalışmaya rastlanmıştır (Budak ve Demirel, 2003; Çepni ve Çoruhlu, 2010; Çetin ve Sadık, 2020; Yıldırım-Seheryeli ve Gelbal, 2020; Gelbal ve Kelecioğlu, 2007). Fakat bu çalışmalarda öğretmenlerin görüşleri, algıları incelenmiş, verilen hizmet içi eğitimlerin sınıf içi yansımaları değerlendirilmiştir. Öğretmenlerin mesleki gelişimleri için herhangi bir kalıcı çözüm önerisine rastlanmamıştır. Bu nedenle bu çalışmada müdahale programı aktif öğrenmeye dayalı olarak hazırlanmıştır.

Aktif öğrenme, öğrencilerin öğrenme sürecinin sorumluluğunu araştırarak, sorgulayarak, tartışarak, paylaşarak, rol oynayarak vb. eylemlerle aldıkları uygulamalardır. Pasif alıcı konumdan çıkan öğrenciler, kendi öğrenme akışlarını ve grup çalışmalarını yürütme ve sürdürme becerilerini geliştirirler. Aynı zamanda aktif öğrenme öğrencileri öğretmen ve akranlarıyla iş birliği yapmaya, karar almaya, üst düzey düşünmeye, disiplinler arası çalışabilmeye, karmaşık içerikleri gerçek yaşam bağlamlarında yapılandırmaya ve anlamaya cesaretlendirir (Grabinger ve Dunlop, 1995). Aktif öğrenme uygulamaları her yaş öğrenenin öğrenme etkinliklerinde kullanılabilmektedir. Örnek olay, proje, benzetişim, grup çalışmaları vb. birçok sürecin büyük yaş grubu öğrencilerde yürütülebildiği anlaşılmaktadır (Silberman,

1996). Araştırmalar aktif öğrenme yöntem ve tekniklerinin öğretmen adaylarının da mesleki yeterliklerinin geliştirilmesinde güçlü bir etkiye sahip olduğunu göstermektedir (Niemi ve diğ., 2016).

Bu çalışmanın amacı, araştırmacılar tarafından hazırlanan aktif öğrenmeye dayalı test hazırlama etkinliklerinin bir özel okuldaki 25 lise öğretmeninin ölçme ve değerlendirmenin programdaki durumuna yönelik tutumlarına etkisini incelemektir.

Yöntem

Araştırma seçkisiz olmayan deneysel desende yürütülmüştür (Büyüköztürk ve diğ., 2020). Gaziantep'teki bir vakıf kurumunda çalışan lise öğretmenleri seçkisiz olmayan örnekleme yöntemlerinden biri olan uygun örnekleme yöntemi ile belirlenmiş tüm öğretmenler çalışmaya dâhil edilmiştir. Bu nedenle araştırmada kontrol grubu yer almamakta, aynı deneysel işlemi alan sayısal ve sözel zümre öğretmenlerinden oluşan iki farklı deney grubu bulunmaktadır. Sayısal zümrede yedi matematik, üç fizik, üç kimya ve üç biyoloji öğretmeni olmak üzere 17 öğretmen; sözel zümrede dört Türk dili ve edebiyatı, iki tarih ve iki coğrafya öğretmeni olmak üzere sekiz öğretmen vardır. Öğretmenlerden, üç alt boyuttan oluşan Ölçme ve Değerlendirmeye İlişkin Tutum Ölçeğinin (Çalışkan ve Yazıcı, 2013) "Programdaki Duruma İlişkin Tutumlar" alt boyutu kullanılarak ön, son ve izleme ölçümleri elde edilmiştir. İzleme testinde Covid-19 salgını nedeniyle sayısal gruptaki altı öğretmene ulaşılamamıştır.

Aktif öğrenmeye dayalı test hazırlama etkinlikleri 14 hafta olacak şekilde zümre çalışmaları şeklinde planlanmıştır. Öğretmenler, aynı okulda görev yapan bir ölçme ve değerlendirme uzmanı ve bir araştırmacı her hafta 40 dakikalık etkinliklerde buluşmuşlardır. Etkinlikler, ölçme ve değerlendirmenin önemi, taksonomiler, kazanım ve taksonomiye uygun soru yazma çalışmaları, test ve madde istatistikleri ve geri bildirim gibi konu başlıklarında grup çalışmasında beyin fırtınası yaparak, sorgulayarak ve tartışarak sürekli geribildirim ile tamamlanmıştır. Ölçme ve değerlendirme uzmanı tüm haftalarda katılımcı gözlemci olarak bulunmuş ve akışı takip etmiştir. Aynı zamanda öğretmenlere doküman sağlama, soru havuzu programını kullanma, testi uygulama konularında öğretmenlere ve araştırmacılara destek olmuştur.

Veri analizinde öncelikle normallik testleri yapılmış ve verilerin normal dağılımdan çok büyük sapma göstermediğinin kanıtı olarak alınmıştır. Uygulanan işlemin deney öncesinden sonrasına fark yaratıp yaratmadığına ilişkin öncelikle öntest-sontest puanlarına ilişkin 2x2 ANOVA testi, daha sonra öntest-sontest-izleme testi puanlarına ilişkin 2x3 ANOVA testi kullanılmıştır.

Sonuçlar

Araştırmanın her iki testi sonucunda da ortak etki anlamlı bulunmamıştır, bu nedenle temel etkiler yorumlanmıştır. Öntestten sonteste olan değişim incelendiğinde grup temel etkisinde, ölçümler fark etmeksizin farklı gruplarda olan öğretmenlerin puan ortalamaları arasında anlamlı farklılık olduğu görülmüştür ($F(1,23)= 6.31$; $p< .05$; kısmî $\eta^2= .22$). Sayısal ve sözel zümre öğretmenlerinin grup fark etmeksizin öntest-sontest puan ortalamalarının müdahale öncesinden sonrasına anlamlı farklılık

gösterdiği bulunmuştur ($F(1,23)= 16.02$; $p< .05$; kısmî $\eta^2= .41$). Öntest puan ortalamaları yüksek olan sözel zümrenin sontest puan ortalamaları da yüksektir. Öntest-sontest ve izleme testinde olan değişim incelendiğinde de ölçüm fark etmeksizin gruplar arasında anlamlı fark olduğu ($F(1,17)= 6.28$; $p< .05$; kısmî $\eta^2= .27$) ; yine grup fark etmeksizin ölçümler arasındaki fark anlamlı bulunmuştur ($F(1,17)= 9.11$; $p< .05$; kısmî $\eta^2= .35$). Bonferroni ile test edilen ikili karşılaştırmalar incelendiğinde ise yalnız sontest ile izleme testi arasında anlamlı farklılık görülmüştür. Bu bulgu müdahaleden sonra artan tutum puanlarının kalıcılığının devam ettiği anlamına gelmektedir. Bu durumda aktif öğrenmeye dayalı test hazırlama etkinliklerinin sayısal ve sözel zümre öğretmenlerinin ölçme ve değerlendirmenin programdaki durumuna yönelik tutumlarını arttırmada etkili olduğu söylenebilir.

Kaynaklar

Demirel, Ö. ve Budak, Y. (2003). Öğretmenlerin hizmet içi eğitim ihtiyacı. *Kuram ve Uygulamada Eğitim Yönetimi*, *33*(33), 62-81.

Büyüköztürk, Ş. (2020). *Sosyal bilimler için veri analizi el kitabı.* Pegem Akademi.

Ceylan, E., Özdogan-Özbal, E., Sever, M. ve Boyacı, A. (2020). *Türkiye'deki öğretmen ve okul yöneticilerinin görüşleri, öğretim koşulları: TALIS 2018 öğretmen ve okul yöneticileri yanıtları analizi.* Millî Eğitim Bakanlığı Yayınları.

Çalışkan, H. ve Yazıcı, K. (2013). Ölçme ve değerlendirmeye yönelik tutum ölçeğinin geliştirilmesi ve sosyal bilgiler öğretmenlerinin tutum düzeylerinin çeşitli değişkenlere göre incelenmesi. *Journal of Human Sciences, 10*(1), 398–415. Retrieved from https://j-humansciences.com/ojs/index.php/IJHS/article/view/2535.

Çepni, S. ve Çoruhlu, T. Ş. (2010). Alternatif ölçme ve değerlendirme tekniklerine yönelik hazırlanan hizmet içi eğitim kursundan öğretime yansımalar. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, *28*(28), 117-128.

Çetin, A. ve Sadık, F. (2020). Özel dershanelerden resmi okullara atanan fen bilimleri öğretmenlerinin mesleki yeterlik algıları ve davranışlarının incelenmesi. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi*, *21*(1), 456-503.

Garipağaoğlu, B. Ç. (2016). Özel dershanelerden özel okullara dönüşüm projesi. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, *16*(1), 140-162.

Gelbal, S. ve Kelecioğlu, H. (2007). Teachers' proficiency perceptions of about the measurement and evaluation techniques and the problems they confront. *Hacettepe University Journal of Education, 33*, 135-145.

Grabinger, R.S., ve Dunlap, J.C. (1995). Rich environments for active learning: a definition, *ALT-J, 3*(2), 5-34, https://doi.org/10.1080/0968776950030202

Gümüş, A. (2014). Dershane düzenlenmesi tartışmalarına eğitsel bir bakış. *İlmi Etüdler Derneği*, 3, 1-8. https://www.ilem.org.tr/images/IPN_3_arife_gumus.pdf

Millî Eğitim Bakanlığı Özel Öğretim Kurumları Yönetmeliği (2008, 8 Mart). *Resmî Gazete* (Sayı: 26810). Erişim adresi: https://www.resmigazete.gov.tr/eskiler/2008/03/20080308-6.htm

Millî Eğitim Bakanlığı (2020, 4 Eylül). *Resmî istatistikler*. https://sgb.meb.gov.tr/www/resmi-istatistikler/icerik/64

Millî Eğitim Bakanlığı (2021). *Öğretmenlik alanları, atama ve ders okutma esasları*. https://ttkb.meb.gov.tr/www/ogretmenlik-alanlari/icerik/201

Nartgün, Ş. S., Altundağ, Ö. Ü. ve Özen, R. (2012). Öğrencilerin sosyal ve ekonomik yaşamlarına dershanelerin etkisi. *Journal of Educational and Instructional Studies in the World*, *2*(1), 54-61.

Niemi, H., Nevgi, A., and Aksit, F. (2016). Active learning promoting student teachers' professional competences in Finland and Turkey, *European Journal of Teacher Education*, *39*(4), 471-490. https://doi.org/10.1080/02619768.2016.1212835

Öner, G. (2007). *Özel dershanelerin ilköğretim matematik öğretimindeki yeri ve önemi* (Tez No. 201948)[Yüksek Lisans Tezi, Eskişehir Osmangazi Üniversitesi]. Yükseköğretim Kurumu Tez Merkezi.

Özoğlu, M. (2011). Özel dershaneler: Gölge eğitim sistemiyle yüzleşmek. *Siyaset, Ekonomi ve Toplum Araştırmaları Vakfı,* 36, 1-28.

Silberman, M. (1996). *Active Learning: 101 Strategies to Teach Any Subject*. Massachusetts: Allyn and Bacon.

Yıldırım-Seheryeli, M., and Gelbal, S. (2020). Practices and opinions of teachers working at public, private and International Baccalaureate schools on measurement and evaluation. *International Journal of Curriculum and Instructional Studies, 10*(1), 221-260. https://doi.org/10.31704/ijocis.2020.008

Investigation of Type-I-error and power of similarity indices by using two-stage analysis via person-fit statistics

Arzu Uçar ve Celal Deha Doğan

Introduction

Validity is the degree to which a test can accurately measure what it intends to measure. Scores other than the purpose of the test; other variables such as bias in scoring, giving points to handwriting, cheating (answer copying) of persons may also be involved. It is important that the measurement results used when making decisions about persons are obtained from a valid test or that the measurements are valid. Because while making decisions or feedback about persons, the measurement results obtained about the person are used. The decisions taken affect the life of the person. In such important situations, the motivation of the person to cheat (answer copying) increases. The cheating situation is seen as compromising the validity of inferences about the competence, skills, and competencies of individuals. Many statistical methods have been developed and continue to be developed to detect suspicious persons (copier(s), suspicious copier(s)) individuals. Although the purpose of the statistical methods used is to detect the suspicious person, these methods sometimes detect person who does not cheat as a suspicious person. This inaccurate detection of statistics is called type I error. Although studies based on type I and power of statistics to detect test fraud are common in the literature, it is seen that there is no single statistic that gives the best performance (Armstrong and Shi, 2009; Belov and Armstrong, 2010; Belov, 2013, 2014, 2016; Belov et al., 2007; Karabatsos, 2003; Krimpen-Stoop and Meijer, 2001; Shu, 2010; Sotaridona and Meijer, 2002, 2003; Sunbul and Yormaz, 2018; Wollack, 1997; Wollack and Cohen, 1998; Wollack, 2006; Yormaz and Sunbul, 2017; Yormaz, 2019; Zopluoglu and Davenport, 2012; Zopluoglu, 2016). Therefore, while detecting suspicious person(s), the results of a single statistic are not adhered to and more evidence is given importance. In this study, the performance of the methods used to detect suspicious person(s) in different scenarios was examined, taking into account the situations that can be encountered frequently in the real world.

Method

The number of replications (100), the ability level of the suspicious person (very low (-3, -1.51) and low (-1.5, 0)) and the average difficulty level of the test ((easy (-2.5), 0) and difficult (0.01,2.5)) 400 data were simulated for type-I-error. For each replication, ability levels of 1500 persons were generated

from the normal distribution. For power analyses, the number of replication (100), the copy ratio (0.1, 0.4, 0.6), the ability level of the suspicious personl ((-3, -1.51)-(-1.5, 0)) and the average difficulty level of the test ((-2.5, 0)-(0,01,2.5)) 1200 data were generated. In order to determine the type-I-error rates of copy and similarity indexes (ω and GBT), in each replication with data known not to be manipulated the "CopyDetect" package in the R program (Zopluoglu, 2018) is used. The probability that were less than and equal to α level (the decision "there is test fraud due to answer copying") got 1, and the ones that were larger ("there is no test fraud due to answer copying") got 0, then a 1-0 matrix was created. The ratio of the sum of the "1" in the matrix to the total number of selected pairs was calculated. In the two-stage analysis, firstly, suspicious persons were detected by using person-fit/KL statistics. One of the persons who were seen as the remaining potential source according to where the suspicious persons sat in the classroom was selected and included in the analysis with the copiers. Probability of copy and similarity indexes were calculated. The ratio of the sum of the "1" values in the matrix to the number of suspicious person-source pairs obtained after using the person-fit/KL index in the two-stage analysis was calculated. The steps, performed to calculate the type-I-error rate, were repeated on the manipulated data, and the power ratios were calculated.

Results

In the light of the interaction effect results of the variables in this study, the type I error rate values obtained when the ω and GBT indexes were used no stage analysis was higher than the type I error rate values obtained when they were used in the two stage analysis. The reults show that the power performances of the ω index are generally lower than the power ratio of the ω and GBT index, which is used with two-stage analysis, and the power ratio of the GBT index, which was used no stage analysis. As a result, type I error and power ratio performances obtained by two-stage analysis of copy and similarity indexes used in the study showed better results compared to the performances obtained with no stage analysis.

References

Armstrong, R., & Shi, M. (2009). Model-free CUSUM methods for person fit. *Journal of Educational Measurement, 46*, 408-428. https://doi.org/10.1111/j.1745-3984.2009.00090.x

Belov, D. I., & Armstrong, R. D. (2010). Automatic detection of answer copying via Kullback–Leibler divergence and K-index. *Applied Psychological Measurement, 34*, 379–392. https://doi.org/10.1177/0146621610370453

Belov, D. I. (2013). Detection of test collusion via Kullback–Leibler divergence. *Journal of Educational Measurement*, *50*, 141–163. https://doi.org/10.1111/jedm.12008

Belov, D. (2014). *Detection of aberrant answer changes via Kullback– Leibler divergence. Journal of Educational Measurement, 50*(2), 141-163. https://doi.org/10.1111/jedm.12008

Belov, D. (2016). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement*, *40 (2)*, 83-9. https://doi.org/10.1177/0146621615603327

Belov, D. I., Pashley, P. J., Lewis, C., & Armstrong, R. D. (2007). Detecting aberrant responses with Kullback–Leibler distance. In K. Shigemasu, A. Okada, T. Imaizumi and T. Hoshino (Eds.), *New trends in psychometrics* (pp. 7-14). Universal Academy Press.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277–298. https://doi.org/10.1207/S15324818AME1604_2

Krimpen-Stoop, E. M. L. A. and Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics*, *26*(2), 199– 217. https://doi.org/10.3102/10769986026002199

Shu, Z. (2010*). Detecting test cheating using a deterministic, gated item response theory model* (Publication No. 3434164) [Doctoral dissertation, The University of North Carolina]. ProQuest Dissertations & Theses Global.

Sotaridona, L. S., and Meijer, R. R. (2002). Statistical properties of the K-index for detecting answer copying in a multiple-choice test. *Journal of Educational Measurement, 39(2)*, 115–132. https://www.jstor.org/stable/1435251

Sotaridona, L. S. and Meijer, R. R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement, 40*(1), 53–70. https://www.jstor.org/stable/1435054

Sunbul O., and Yormaz, S. (2018a). Effects of test level discrimination and difficulty on answercopying indices. *International Journal of Evaluation and Research in Education*, *7(1)*, 32- 38. http://doi.org/10.11591/ijere.v7i1.11488

Wollack, J. A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement*, *21*(4), 307–320. https://doi.org/10.1177/01466216970214002

Wollack, J. A. (2006). Simultaneous use of multiple answer copying indexes to improve detection rates. *Applied Measurement in Education*, *19*(4), 265–288. https://doi.org/10.1207/s15324818ame1904_3

Wollack, J. A., and Cohen, A. S. (1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement, 22*(2), 144–152. https://doi.org/10.1177/01466216980222004

Yormaz, S., and Sunbul, O. (2017). Determination of type I error rates and power of answer copying ındices under various conditions. *Educatıonal Sciences: Theory & Practıce*, *17*(1), 5-26. https://doi.org/10.12738/estp.2017.1.0105

Yormaz, S. (2019). *Test guwenligi açısından bireyler arasındaki olası is birliğinin incelenmesi* (Tez No. 608727) [Doktora tezi, Mersin Üniversitesi]. Yükseköğretim Kurumu Tez Merkezi.

Zopluoglu, C. (2016). Classification performance of answer-copying indices under different types of IRT models. *Applied Psychological Measurement 40*(8), 592–607. https://doi.org/10.1177/0146621616664724

Zopluoglu, C. (2018). Computing response similarity ındices for multiple-choice tests (CopyDetect) (Version 1.3) [Computer software manual]. https://cran.rproject.org/web/packages/CopyDetect/CopyDetect.pdf

Zopluoglu, C., and Davenport, E. C., Jr. (2012). The empirical power and type I error rates of the GBT and ω indices in detecting answer copying on multiple-choice tests. *Educational and Psychological Measurement*, *72(6)*, 975–1000. https://doi.org/10.1177/0013164412442941

Kayıp verinin Mantel-haenszel, MIMIC ve olabilirlik oranı DMF belirleme yöntemlerinin performanslarına etkisinin incelenmesi

Rabia Akcan ve Kübra Atalay Kabasakal

Anahtar kelimeler: Değişen madde fonksiyonu, Mantel Haenszel, MIMIC, olabilirlik oranı, kayıp veri

Giriş

Kayıp veri nicel araştırmalarda sıklıkla karşılaşılan bir problemdir. Standart istatistiksel yöntemler tam veri setleri için geliştirildiğinden kayıp veriler araştırmacı için önemli bir sorun haline gelmektedir. Araştırmacılar kayıp veri problemiyle başa çıkmak amacıyla analizden önce veriyi düzeltmek için geliştirilen çeşitli yöntemler kullanmaktadırlar. Eksik veri içeren bireylerin veri setinden çıkartılması (liste bazında silme vb.) bunlardan biridir. Bir diğer yöntem ise kayıp verilere değişken ortalamalarının atanmasıdır. Ancak bu geleneksel yöntemler örneklem istatistiklerinin ciddi derecede yanlı olmasına neden olabilmektedir (Peugh and Enders, 2004).

Kayıp veriler, test geliştirmenin önemli bir adımı olan madde yanlılığı çalışmalarına ilişkin gerçekleştirilen değişen madde fonksiyonu (DMF) analizlerinde de önemli bir yere sahiptir. Yaygın olarak kullanılan Mantel Haenszel (MH), SIBTEST ve lojistik regresyon gibi DMF belirleme yöntemleri kayıp veriyle baş edecek şekilde tasarlanmamıştır. Bu analizlerde tercih edilen kayıp veriyle baş etme yöntemleri yanlılık sebebi olabilmektedir. Tercih edilen yöntem, maddede gerçekte var olan DMF'nin yok gibi gözükmesine veya maddede DMF olmadığı halde maddenin DMF göstermesine neden olabilir (Banks, 2015).

Kayıp veri problemini çözmek için başvurulacak yönteme, kayıp veri oranı, kayıp veri mekanizması ve kayıp verinin yapısı dikkate alınarak karar verilmelidir. Kayıp veri oranı doğrudan istatistiksel olarak yapılacak çıkarımların niteliğiyle ilişkilidir. Ancak literatürde geçerli istatistiksel sonuçlar elde etmek için belirlenmiş kabul edilebilir bir kayıp veri oranı ölçütü bulunmamaktadır (Dong and Peng, 2013). Rubin'in (1976) sınıflamasına göre kayıp veri üç farklı şekilde tanımlanmaktadır: Tamamen rastgele kayıp (TRK), rastgele kayıp (RK) ve rastgele olmayan kayıp (ROK). Madde cevapları bağlamında TRK, cevaplayıcıların maddeyi sistematik bir mekanizma olmadan tamamen rastgele olarak boş bıraktığı durumu ifade eder. RK kayıp veri içeren bir gözlemin olasılığının ölçülebilen bir başka

değişkenle doğrudan ilişkisi olması durumunda meydana gelir. Erkek öğrencilerin bir maddeyi boş bırakma olasılığının kız öğrencilere göre daha yüksek olması RK mekanizmasına örnek verilebilir. ROK ise kayıp veri olma olasılığının değişkenin kendi değeriyle ilişkili olması durumudur. Bu durumda öğrenci cevabı bilmediği için soruyu boş bırakmış olabilir (Finch, 2011b).

Literatür incelendiğinde kayıp veriyle baş etme yöntemlerinin DMF analizleri üzerindeki etkisini farklı koşullar altında inceleyen simülasyon çalışmalarına rastlanmıştır (Banks & Walker, 2006; Finch, 2011a; Finch, 2011b; Robitzsch ve Rupp, 2009). Bu çalışmalarda farklı kayıp veri mekanizmalarının birer koşul olarak ele alındığı görülürken; gerçek veri setleri üzerinden yapılan araştırmalarda farklı kayıp veri mekanizmaları üzerinde çalışan çok fazla araştırmaya rastlanmamıştır. Ayrıca hem simülasyon hem de gerçek veri setleriyle yapılan çalışmalarda klasik DMF belirleme yöntemlerinden MH, lojistik regresyon ve SIBTEST yöntemlerinin sıklıkla kullanıldığı görülmüştür. Bazı çalışmalarda ise Klasik Test Kuramı ve Madde Tepki Kuramına dayalı yöntemlerin karşılaştırıldığı belirlenmiştir. Doğrulayıcı Faktör Analizi'ne (DFA) dayalı bir yöntem olan çoklu nedenler çoklu göstergeler (MIMIC) yöntemi de son yıllarda DMF belirleme çalışmalarında yaygın olarak kullanılmaya başlanmıştır (Akın-Arıkan ve diğ., 2016; Finch, 2005; Shih and Wang, 2009). Kayıp veri DFA da dahil olmak üzere her türlü veri analizini etkileyebilir (Harrington, 2009). Bu nedenle kayıp veri olması durumunda MIMIC yönteminin performansının nasıl etkilendiği değerlendirilmelidir.

Yöntem

Bu çalışmada ikili puanlanan maddelerde farklı kayıp veri oranı ve kayıp veri mekanizmaları altında oluşturulan veri setlerine kayıp veriyle başa etme yöntemleri uygulanarak MH, Madde Tepki Kuramı Olabilirlik Oranı (MTK-OO) ve MIMIC yöntemlerinin bu koşullar altında DMF belirleme performanslarını belirlemek amaçlanmıştır. Bu bakımdan bu çalışma nicel araştırmalar kapsamında betimsel bir çalışmadır.

Araştırmada kullanılan gerçek veri seti PISA 2015 (Program for International Student Assessment) uygulamasının fen bilimleri testindeki S12 madde kümesinde yer alan 17 maddeden oluşmaktadır. Finlandiya veri setindeki 17 maddenin tümüne yanıt veren 1099 öğrenci çalışmanın örneklemini oluşturmaktadır.

Veri setinde yer alan 16 madde ikili puanlanan maddelerdir. Kısmi puanlanan bir madde (CS637Q02S) ise 1-0 şeklinde kodlanmış ve analizler 17 madde üzerinden yürütülmüştür. Veri setinden kayıp veriler atılmış; 550 kız ve 549 erkek öğrenci olmak üzere 1099 kişilik tam veri seti elde edilmiştir. Bu veri seti üzerinde daha sonra referans olarak kullanılmak üzere MH, MTK-OO ve MIMIC yöntemleriyle cinsiyete göre DMF analizleri gerçekleştirilmiş ve sonuçlar kaydedilmiştir. Tam veri setinden TRK, RK ve ROK mekanizmaları altında %10 ve %30 oranında veri silinerek kayıp veri içeren veri setleri elde edilmiştir. Bu veri setlerinin kayıp veri problemi liste bazında silme (LBS), sıfır atama (SA) ve Beklenti maksimizasyonu (BM) yöntemleri kullanılarak çözülmüştür. Kayıp veri problemi çözülen veri setlerinin belirlenen üç yöntemle DMF analizleri gerçekleştirilmiştir. Analizlerden elde

edilen sonuçlar tam veri setinden elde edilen referans sonuçlarla karşılaştırılarak I.tip hata ve güç oranları hesaplanmıştır.

Sonuçlar

Tam veri seti üzerinde MH analizleri için R yazılımında "difR" paketi (Magis ve diğ., 2020) kullanılmıştır. MIMIC analizleri "MplusAutomation" paketi (Hallquist and Wiley, 2021) kullanılarak Mplus üzerinden gerçekleştirilmiştir. MTK-OO analizleri için "mirt" paketinden (Chalmers, 2021) yararlanılmıştır. MTK-OO analizlerinde ön analiz yapılarak her bir maddenin DMF analizi için tüm maddeler ankor madde olarak kullanılmış, bu ön analiz sonuçları ve diğer DMF yöntemlerinin sonuçları göz önüne alınarak beş madde (m12-m16) son analiz için ankor madde olarak belirlenmiştir. MH, MTK-OO ve MIMIC yöntemleriyle gerçekleştirilen ve referans olarak kullanılan DMF analizlerinin sonuçları Tablo 1'de verilmiştir.

Tablo 1

Tam Veri Setinden Elde Edilen DMF Sonuçları

	MH		MTK-OO		MIMIC	
	Δ_{MH}	Düzey	G^2	p	Beta	p
M1	0.071	-	0.421	.517	0.038	.704
M2	-1.422	B	11.698	.001	-0.297	.000
M3	-0.226	-	2.982	.084	0.027	.685
M4	0.178	-	0.703	.402	0.047	.476
M5	-0.815	-	0.132	.717	-0.267	.019
M6	-0.641	-	7.742	.005	-0.110	.123
M7	0.491	-	0.465	.495	0.089	.229
M8	0.332	-	0.033	.856	0.153	.030
M9	0.326	-	0.536	.464	0.097	.205
M10	-1.313	B	6.981	.008	-0.218	.001
M11	-0.750	A	8.962	.003	-0.178	.010
M12	0.489	-			0.068	.335
M13	0.221	-			0.126	.056
M14	0.664	-			0.098	.221
M15	0.732	-			0.113	.138
M16	0.175	-			0.031	.768
M17	1.367	B	0.819	.366	0.220	.005

Tablo 1 incelendiğinde MH yöntemiyle üç maddenin orta düzeyde, bir maddenin düşük düzeyde DMF gösterdiği belirlenmiştir. MIMIC yönteminde altı madde; MTK-OO yönteminde ise dört madde DMF göstermiştir. Çalışma kapsamında farklı kayıp veri oranı ve kayıp veri mekanizmaları altında oluşturulan veri setlerine kayıp veriyle başa etme yöntemleri uygulanarak DMF analizleri gerçekleştirilmiştir. Nihai sonuçlar tam veri seti ve oluşturulan tüm veri setlerinden elde edilen sonuçlar

kıyaslanarak her bir DMF yönteminin I.tip hata ve güç oranın hesaplanması, hangi koşulda hangi yöntemin daha iyi çalıştığının tespit edilmesiyle elde edilecektir.

Kaynaklar

Akın Arıkan, Ç., Uğurlu, S. ve Atar, B. (2016). MIMIC, SIBTEST, Lojistik Regresyon ve Mantel-Haenszel yöntemleriyle gerçekleştirilen DMF ve yanlılık çalışması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi , 31*(1), 34-52.

Banks, K. (2015). An introduction to missing data in the context of differential item functioning. *Practical Assessment, Research & Evaluation , 20* (12), 1-10.

Banks, K., and Walker, C. (2006). *Performance of SIBTEST when focal group examinees have missing data*. National Council of Measurement in Education.

Chalmers, R. P. (2021). mirt: Multidimensional item response theory (version 1.34). https://cran.r-project.org/web/packages/mirt/index.html

Dong, Y., and Peng, C. Y. (2013). Principled missing data methods for researchers. *SpringerPlus, 2*, 222. https://doi.org/10.1186/2193-1801-2-222

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*(4), 278-295.

Finch, H. (2011a). The use of multiple imputation for missing data in uniform DIF analysis: Power and type I error rates. *Applied Measurement in Education*, 24, 281-301.

Finch, H. (2011b). The impact of missing data on the detection of nonuniform differential item functioning. *Educational and Psychological Measurement , 71*(4), 663-683.

Hallquist, M., and Wiley, J. (2021). MplusAutomation, version 1.0.0: An R package for facilitating large-scale latent variable analyses in Mplus. https://cran.r-project.org/web/packages/MplusAutomation/index.html adresinden alınmıştır

Harrington, D. (2009). *Confirmatory Factor Analysis.* Oxford University Press.

Magis, D., Beland, S., & Raiche, G. (2020). difR: Collection of methods to detect dichotomous differential item functioning (DIF) (version 5.1). https://cran.r-project.org/web/packages/difR/difR.pdf adresinden alınmıştır

Peugh, J. L., and Enders, C. K. (2004). Missing Data in Educational Research:A Review of Reporting Practices and Suggestions for Improvement. *Review of Educational Research , 74*(4), 525-556.

Robitzsch, A., & Rupp, A. A. (2009). Impact of missing data on the detection of differential ıtem functioning the case of mantel-haenszel and logistic regression analysis. *Educational and Psychological Measurement , 69*(1), 18-34.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592.

Shih, C. L., and Wang, W. C. (2009). Differential item functioning detection using multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement , 33*(3), 184-199. https://doi.org/10.1177/0146621608321758

Examination of alfa, omega and AVE reliability coefficients according to number of items, number of response category and sample size: A simulation study

Ahmet Salih Şimşek

Anahtar kelimeler: Cronbach Alfa, McDonald Omega, AVE, güvenirlik, örneklem büyüklüğü, simülasyon

Giriş

Ölçek geliştirme çalışmalarında yaygın olarak Cronbach Alfa güvenirlik katsayısı rapor edilmektedir. Bazı çalışmalar, Cronbach Alfa güvenirlik katsayısı yerine McDonald Omega güvenirlik katsayısının tercih edilmesi gerektiğine vurgu yapmaktadır (Hayes ve Coutts, 2020). Matematiksel altyapıları birbirinden farklı olmasına karşın Alfa ve Omega katsayıları birbirine yakın değerler üreten güvenirlik katsayılarıdır. Literatürde, Omega katsayısının mutlak üstünlüğünü gösteren bir kanıt bulunmamasına karşın bazı akademik dergilerde Alfa yerine Omega katsayının rapor edilmesi için yönlendirme yapıldığı görülmektedir. Alanyazında, Alfa ve Omega katsayıları arasındaki farkın incelendiği çalışmalara ihtiyaç duyulduğu belirtilmektedir (Deng ve Chan, 2016). Ölçek geliştirme çalışmalarında Alfa yerine Omega katsayısının rapor edilmesine yönelik vurgunun yerindeliği bu araştırmanın problemini oluşturmaktadır. Bu araştırmada, Alfa, Omega ve AVE güvenirlik katsayılarının farklı koşullar altında ürettikleri değerler betimsel ve ilişkisel olarak incelenmiştir.

Yöntem

Bu araştırma, bir simülasyon (benzetim) çalışmasıdır. Araştırma kapsamında, belirlenen koşullara uygun olarak R ortamında veriler üretilmiş ve analiz edilmiştir. Araştırma verilerinin üretilmesi için psych (2.1.6; Revelle, 2011) paketi kullanılmıştır. Araştırma kapsamında dört farklı madde sayısı (3, 6, 12, 24), dört farklı örneklem büyüklüğü (50,200,500,1000) ve üç farklı yanıt kategori sayısı (3, 5, 7) kullanılarak 48 farklı deneysel koşul oluşturulmuştur. Her koşul için 100 farklı veri seti üretilerek toplam 4800 veri seti üretilmiştir. Araştırma normal dağılıma sahip veri üretilmesi için sim.poly.npn() fonksiyonu kullanılmıştır.

Veri setleri için öncelikle paralel analiz yöntemi kullanılarak boyut sayısı incelenmiştir. Tek boyutluluğun incelenmesi için en yüksek özdeğere sahip iki komşu faktör için özdeğer oranları

hesaplanmıştır. Komşu faktörler özdeğer oranı $\frac{\lambda_1}{\lambda_2} > 4$ olması tek boyutlu yapıya işaret etmektedir (Slocum-Gori ve Zumbo, 2011). Araştırmada üretilen veri setleri için komşu faktörler özdeğer oranı incelendiğinde tüm veri setleri için tek boyutlu yapı elde edildiği belirlenmiştir. Bu nedenle tüm veri setleri için tek boyutlu yapı tanımlanarak lavaan (0.6-9; Rosseel, 2012) paketi cfa() fonksiyonu ile DFA analizi yapılmıştır. Oluşturulan DFA modelleri için semTools (0.5-5; Jorgensen ve diğ., 2021) paketi reliability() fonksiyonu ile Cronbach Alfa (α), McDonald Omega (ω) ve Ortalama Açıklanan Varyans (AVE) güvenirlik katsayıları elde edilmiştir. Araştırmada madde sayısı koşulu i_1=3 veri setleri için elde edilen komşu faktörler özdeğer oranları için anormal değerler (IQR=1001e+06, σ= 9746e+07) elde edilmiştir. Diğer madde sayısı koşulları (6, 12, 24) için hesaplanan değerlere (IQR= 0.93, σ= 1.13) göre aşırı ve anormal değerler ürettiği için 3 madde sayısı koşulu korelasyon ve varyans analizlerinde analiz dışında bırakılmıştır.

Sonuçlar

Betimsel olarak incelendiğinde madde sayısı arttığında ve yanıt kategori sayısı arttığında daha tutarlı sonuçlar elde edildiği belirlenmiştir. Örneklem büyüklüğünün ise madde sayısı ve yanıt kategori sayısı parametrelerine görece güvenirlik katsayılarındaki değişimde etkili olmadığı belirlenmiştir. Madde sayısı koşuluna göre elde edilen bulgular literatür ile benzerlik göstermesine karşın örneklem büyüklüğü koşulu için elde edilen bulgular literatürle farklılık göstermektedir (Ercan ve diğ., 2007). Tüm koşullar için elde edilen Alfa, Omega ve AVE katsayıları arasındaki korelasyonlar incelenmiştir. AVE ile hem Alfa hem de Omega güvenirlik katsayıları arasında anlamlı orta düzeyde ($r_{AVE-\alpha}$= .54, $r_{AVE-\omega}$= .61) ilişki bulunmuştur. Alfa ve Omega güvenirlik katsayıları arasında ise anlamlı çok yüksek ($r_{\alpha-\omega}$= .99) ilişki elde edilmiştir. Tüm koşullar için hesaplanan Alfa ve Omega katsayıları arasında Omega katsayısı lehine .02 anlamlı fark bulunmuştur. Etki büyüklüğünün değerlendirilmesi için hesaplanan Cohen d katsayısı incelendiğinde düşük etkiye (Cohen's *d*=.15) sahip olduğu belirlenmiştir. Alfa ve Omega katsayıları arasındaki farka ilişkin bulgular gerçek veriler ile yürütülen çalışmalar ile desteklenmektedir (Deng ve Chan, 2016). Sonuçlar Omega güvenirlik katsayısının Alfa katsayısına göre daha yüksek sonuçlar verdiğini ancak oluşan bu farkın düşük bir etkiye sahip olduğunu göstermiştir.

Kaynaklar

Deng, L., and Chan, W. (2017). Testing the difference between reliability coefficients alpha and omega. *Educational and psychological measurement*, *77*(2), 185-203. https://doi.org/10.1177/0013164416658325

Ercan, I., Yazici, B., Sigirli, D., Ediz, B., and Kan, I. (2007). Examining Cronbach alpha, theta, Omega reliability coefficients according to sample size. *Journal of Modern Applied Statistical Methods*, *6*(1), 27. https://doi.org/10.22237/jmasm/1177993560

Hayes, A. F., and Coutts, J. J. (2020). Use Omega rather than Cronbach's alpha for estimating reliability. *Communication Methods and Measures*, *14*(1), 1-24. https://doi.org/10.1080/19312458.2020.1718629

Hedges, L. and Olkin, I. (1985). *Statistical methods in meta-analysis.* Elsevier Inc.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., and Rosseel, Y. (2021). semTools: Useful tools for structural equation modeling (version 0.5-5). https://cran.r-project.org/package=semTools

Revelle W (2021). *psych: Procedures for Psychological, Psychometric, and Personality Research* (version 2.1.6). https://cran.r-project.org/package=psych.

Rosseel Y (2012). lavaan: An r package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://www.jstatsoft.org/v48/i02/.

Slocum-Gori, S. L., and Zumbo, B. D. (2011). Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. *Social Indicators Research*, *102*(3), 443-461. https://doi.org/10.1007/s11205-010-9682-8

Madde konum etkisinin TIMSS 2015 uygulamasında farklı başarı düzeylerine sahip ülkelerde incelenmesi

Sinem Demirkol ve Hülya Kelecioğlu

Anahtar kelimeler: Madde konum etkisi, açıklayıcı madde tepki kuramı, TIMSS

Giriş

Geniş ölçekli uygulanan sınavlarda farklı kitapçıkların kullanılması, öğrenciler arasında kopya çekme davranışını önlemeye veya test kapsamın genişletilmesi için farklı maddelerin uygulamasına yardımcı olmaktadır. Bir maddenin doğru cevaplanma olasılığı sadece madde özelliklerine (örneğin zorluğu) ve öğrencilerin yeterliliklerine bağlı olmalıdır. Ancak gerçekleştirilen test uygulamalarında çoğu zaman bu amaca ulaşılamamaktadır (Asseburg ve Frey, 2013). Bir maddenin testin başında veya sonunda yer alması maddenin güçlüğünü dolayısıyla bireylerin performans düzeylerini etkileyebilir. Maddenin farklı kitapçıklarda farklı konumlarda (sıralarda) yer alması madde karakteristik özelliklerinde, dolayısıyla testi alan bireylerin performanslarında istenmeyen değişikliklere yol açabilir (Leary ve Dorans, 1985). Bu durumda, maddelere verilen yanıtlar ölçülmek istenilen yapının (ör. Bireylerin yetenek düzeyleri) dışında, yapıyla alakasız olan bağlam (context) etkilerine de bağlı olabilir (Messick, 1995). Bağlam etkisi, geniş ölçekli uygulanan sınavların analizde sıklıkla ortaya çıkan madde tepki kuramının (MTK) temel varsayımı olan yerel bağımsızlık sayıltısını ihlal etmekte ve madde ve birey parametre tahminlerini saptırarak elde edilen sonuçların geçersiz olmasına yol açmaktadır (AERA, APA ve NCME, 2014).

Bağlam etkilerinin en çok bilinen alt başlıklarında biri madde konum etkisidir. Madde konum etkisi, aynı maddenin farklı sıralarda sunulmasının madde parametreleri üzerindeki etkisini ifade etmektedir (Mollenkopf, 1950; Leary ve Dorans, 1985). Madde konum etkisi pozitif veya negatif olabilir. Pozitif madde konum etkisi, bir maddenin testin sonraki kısımlarında uygulanmasının maddeyi daha kolay hale getirmesi olarak ifade edilebilir. Bu durum, bireylerin test uygulaması sırasında test formuna daha aşina hale gelmeleri, madde tipini daha iyi anlamaları veya testin başında bireylerin heyecan düzeyinden kaynaklanan olumsuz durumların daha sonra azalmasından kaynaklanabilir. Bu yüzden bu etkiye öğrenme veya uygulama etkisi de denilmektedir. Literatür incelendiğinde, Kingston ve Dorans (1984), Lisansüstü Kayıt Sınavı (Graduate Record Examination) verilerini kullanarak gerçekleştirdikleri araştırmalarında farklı konumda yer alan aynı maddelerin güçlük parametrelerindeki ortalama farkları karşılaştırmış ve pozitif madde konum etkisi saptamışlardır. Verguts ve De Boeck (2000), zeka testine ait

verileri kullanarak yürüttükleri çalışmalarında öğrenme etkisi olduğunu belirtmişlerdir. Schweizer ve diğ., (2009), doğrulayıcı faktör analizini kullanarak yürüttükleri çalışmalarında Gelişmiş Aşamalı Matrislere (Advanced Progressive Matrices) ait verileri kullanmış ve öğrenmeyi temsil edebilecek pozitif madde konum etkisinin olduğunu belirtmişlerdir.

Negatif madde konum etkisi, bir maddenin testin sonraki kısımlarında uygulanmasının maddeyi daha zor hale getirmesi olarak ifade edilebilir. Bu etki yorgunluk etkisi olarak da adlandırılır. Kingston ve Dorans (1984), GRE testinde okuduğunu anlama maddelerinin daha sonraki konumlarda yer almasının madde güçlüklerini orta derecede arttırdığını saptamışlardır. Hohensinn ve diğ., (2008), 4. Sınıf matematik testine ait verileri kullanarak madde konum etkisini araştırmışlar ve negatif madde konum etkileri olduğunu saptamışlardır. Albano (2013) çok seviyeli madde tepki kuramına ait modelleri kullanarak madde konum etkisini araştırmış ve okuma testinde negatif madde konum etkileri olduğunu belirtmiştir. Hartig ve Buchholz (2012), PISA 2006 fen verilerini kullanarak on farklı ülke üzerinden gerçekleştirdiği çalışmasında tüm ülkelere ait verilerde negatif madde konum etkisini olduğunu saptamışlardır.

Bireylerin maddelere verdiği cevapların test formundan bağımsız olduğu varsayılır. Bu varsayımın ihlali, bağlam etkileri olarak adlandırılır ve en sık karşılaşılan bağlam etkilerinden biri maddenin farklı konumlarda yer almasından kaynaklanan madde konum etkisidir. Madde konum etkisi ölçülmek istenen yapının dışında, istenmeyen varyansın olası kaynaklarıdır. Bu varyansı göz ardı etmek bireylerin test tarafından ölçülen özelliklerinin yanı sıra madde özellikleri (örneğin madde zorluğu ve madde ayırt ediciliği) ve test özellikleri (örneğin güvenilirlik ve geçerlilik) hakkında yanlı çıkarımlara yol açabilir. Dolayısıyla bu etkilerin belirlenmesi ve uygun şekilde ölçme modeline dahil edilmesi daha geçerli sonuçların elde edilmesine yardımcı olacaktır. Bu bilgiler ışığında bu çalışmanın amacı madde konum etkisinin TIMSS 2015 fen bilimleri alanında farklı başarı düzeyine sahip ülkeler üzerinde incelenmesidir.

Yöntem

Uluslararası Matematik ve Fen Eğilimleri Araştırması (TIMSS -Trends in International Mathematics and Science Study), Uluslararası Eğitim Başarılarını Değerlendirme Kuruluşu (IEA-International Association for the Evaluation of Educational Assessment) tarafından yürütülen ve dört yılda bir gerçekleşen tarama çalışmasıdır. Araştırmanın çalışma grubunu Singapur, Türkiye ve Fas ülkelerinde TIMSS 2015 sınavına giren ve fen bilgisi kümelerinden en az birisine cevap veren 8. Sınıf öğrencileri oluşturmaktadır. Analize dahil edilen öğrenci sayısı Fas örnekleminde 13027 (% 47.7 kadın, % 52.3 erkek), Türkiye örnekleminde 6079 (%48.4 kadın, %51.6 erkek), Singapur örnekleminde ise 6116 (%48.7 kadın, %51.3 erkek)'dir.

TIMSS 2015 uygulamasında çoktan seçmeli, açık uçlu gibi farklı madde formatları mevcuttur. Araştırma kapsamında, madde konum etkisi çoktan seçmeli fen bilimleri maddelerine ait veriler kullanılarak incelenmiştir. TIMSS 2015 kitapçıkları dört farklı kümenin (2 matematik, 2 fen bilimleri olacak şekilde) farklı kombinasyonları ile oluşturulmuştur. Bu kümelerde yer alan maddelerin konumları

sabittir. Bu yüzden bu çalışmada ele alınan konum değişkeni madde kümelerin konumlarıdır. Araştırmanın analizleri açıklayıcı madde tepki kuramı çerçevesinde ele alınmıştır. Açıklayıcı madde tepki kuramına ait modeller, geleneksel madde tepki kuramı modellerinden farklı olarak, madde ve bireyler arasında görülen farklılıkları incelemek için madde ve birey düzeyinde açıklayıcı değişkenlerin ele alınmasına yardımcı olur (De Boeck ve Wilson, 2004). Bu kapsamda madde özelliği olarak kabul edilen madde konumu, maddelerin güçlükleri arasındaki farklılıkları belirlemek için modele dahil edilebilir (Debeer ve Janssen, 2013). Açıklayıcı madde tepki kuramı modellerinde madde özelliklerinin eklenmesiyle oluşturulan model Lineer Lojistik Test Modeli (LLTM) olarak adlandırılırmıştır (De Boeck ve Wilson, 2004). Verilerin analizi, R paketlerinden biri olan ve açıklayıcı madde tepki kuramı modellerinin analizine uygun olan *eirm* paketi ile gerçekleştirilmiştir (R Core Team, 2013; Bulut, 2021).

Sonuçlar

Analizlere madde ve birey düzeyinde herhangi bir açıklayıcı değişken içermeyen temel model (M0) kurularak başlanmıştır. Temel model ile birey ve maddelere ait rastgele etkiler kestirilmiştir. Daha sonra temel modele madde konumu sabit etkisi dahil edilerek M1 modeli ve madde konumunun bireyler içerisindeki rastgele etkisini ele alan M2 modeli ile devam edilmiştir. Analizler üç ülke için ayrı ayrı gerçekleştirilmiştir. Kurulan modellere ilişkin model uyum indeksleri ve ki-kare fark testi sonuçları karşılaştırılmıştır. Elde edilen sonuçlar incelendiğinde, tüm ülkelerde ortaya çıkan madde konum etkisinin negatif olduğu, başka bir ifade ile fen bilimleri alanında yer alan herhangi bir maddenin bir küme daha sonra yer alması madde güçlüğünü arttırdığı görülmüştür. Bu etki Türkiye ve Fas için istatistiksel olarak anlamalı iken, Singapur'a ait verilerde ortaya çıkan etkinin oldukça küçük ve istatistiksel olarak anlamlı olmadığı saptanmıştır. Ayrıca kurulan modellere ilişkin model uyum indeksleri incelendiğinde ise, Fas'ta ortaya çıkan madde konum etkisinin bireyler arasında farklılaştığı fakat bu etkinin Türkiye ve Singapur'da bulunan bireyler arasında sabit olduğu saptanmıştır.

Kaynaklar

Albano, A. D. (2013). Multilevel modeling of item position effects. *Journal of Educational Measurement*, *50*(4), 408–426. https://doi.org/10.1111/jedm.12026

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, & NCME]. (2014). *Standards for educational and psychological testing.* American Psychological Association.

Asseburg, R., and Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling,* 55, 92–104. https://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2013_20130326/06_Asseburg.pdf

Bulut, O. (2021). *eirm: Explanatory item response modeling for dichotomous and polytomous item responses, R package version 0.3.0* [Computer software]. http://cran.r-project.org/package=eirm.

Debeer, D., and Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, *50*(2), 164-185. https://doi.org/10.1111/jedm.12009

De Boeck, P., and Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* Springer.

Core Team, R. (2016). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Hartig, J., and Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling*, *54*(4), 418–431. https://www.proquest.com/docview/1355923397/fulltextPDF/BB039678286B4924PQ/1?accountid=8319

Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L., and Frebort, M. (2008). Examining item-position effects in large-scale assessment using the linear logistic test model. *Psychology Science Quarterly*, *50*(3), 391–402. http://www.psychologie-aktuell.com/fileadmin/download/PschologyScience/3-2008/06_Hohensinn.pdf

Kingston, N. M., and Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, 8, 147–154. https://doi.org/10.1177/014662168400800202

Leary, L. F., and Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research*, 55, 387–413. https://doi.org/10.3102/00346543055003387

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Schweizer, K., Schreiner, M., and Gold, A. (2009). The confirmatory investigation ofAPM items with loadings as a function of the position and easiness of items: A two-dimensional model of APM. *Psychology Science*, *51*(1), 47–64. https://www.psychologie-aktuell.com/fileadmin/download/PschologyScience/1-2009/03_Schweizer.pdf

Verguts, T., and De Boeck, P. (2000). A Rasch model for detecting learning while solving an intelligence test. *Applied Psychological Measurement,* *24*(2), 151–162. https://doi.org/10.1177/01466210022031589

Öğretmenlerin ölçme davranışlarının geçerliğinin incelenmesi

İbrahim Hakkı Tezci, Bayram Bıçak ve Derya Çobanoğlu Aktan

Anahtar kelimeler: Geçerlik, geçerli ölçmeler, ölçme davranışları, ölçek geliştirme.

Giriş

Okul ve sınıf içinde yapılan ölçme ve değerlendirme faaliyetleri, öğrenci başarısının belirlenmesi, geçerli ve güvenilir bir temele dayanan bilgilerin üretilmesi ve bu bilgilerden yola çıkarak bireyler hakkında önemli kararların alınması bakımından çok önemli bir role sahiptir (Atılgan, 2014). Eğitim ortamlarının en önemli yapı taşı olan öğretmenler, öğrencilerin başarısından ve bu bilgilerin üretilmesi için gerekli olan ölçme ve değerlendirme faaliyetlerinin uygulanmasından sorumlu olan bireylerdir. Sözü geçen faaliyetler bütünü ölçme işlemi ile başlayıp elde edilen özellik hakkında bazı kararların alınması ile devam etmektedir. Eğitim alanında yapılan ölçmelerde ölçülecek özellik ve grup farklılıkları ölçme araçlarını belirlemektedir. Bu farklılıklardan dolayı yapılacak değerlendirmenin amacına uygun bir şekilde gerçekleşebilmesi için ona uygun bir ölçme aracı seçilmelidir. Seçilen ölçme araçlarının doğru sonuçlar verebilmesi adına bazı özelliklere sahip olması gerekmektedir. Bu özelliklerin başında ise bir ölçme aracının ölçmek istediği niteliği amaçlanan özelliği ölçebilecek nitelikte olması anlamına gelen geçerlik kavramı bulunmaktadır (Baykul, 2010). Geçerlik kavramı her ne kadar ölçme aracına ait bir özellik gibi gözükse de aslında bahsedilen ölçme aracından elde edilen puanların geçerliğidir. Ölçme aracından elde edilen puanlar, ölçülmek istenen özelliği yansıttıkları oranda geçerli kabul edilirler. Amacı dışında ve standart dışı elde edilmiş puanlar geçerli sonuçlar üretemez. Geçerliği düşük puanlar ise dönem sonu verilecek notların da geçerliğinden şüphe edilmesine neden olur.

Mevcut öğrenmelerin gerçek bir ölçüsünü temsil etmediği durumlarda notların geçerliği sorgulanır. Geçerliği sağlamak için notların öğretim amaç ve hedefleriyle eşleşmesi önerilir (Brookhart, 2009). McMillan (2008) notlandırmanın profesyonel bir sürece dayandığına ve uygulamanın aynı okuldaki öğretmenler arasında bile büyük farklılıklar gösterdiğini öne sürmektedir. Eğitim ortamlarının üç bileşeni ölçme, notlandırma ve değerlendirme doğrusal olmadığında, ölçüm tavsiyelerine uymayan bir değerlendirme durumu ortaya çıkar ve not verme keyfi hale gelebilir. Bu durum, keyfi ve standart dışı not verme uygulamalarının kullanılmasının öğrencinin notunun geçerliğini sorguladığına inanan Guskey (2006) tarafından da desteklenmektedir. Gallagher (1998), ölçme ve notlandırma ile ilgili en temel ilkelerin geçerlik ve güvenilirlik olduğuna inanmaktadır. Ölçme uzmanları tarafından belirlenen önerilere

herkesin uyum göstermesinin eğitim ortamında sağlıklı bir yapının oluşumuna katkı sağlayacağı düşünülmektedir. Ancak öğretmenler arasında tutarlı bir şekilde uygulanmayan bu öneriler, öncelikle öğrenci başarısına dayanan notların kullanılmasıyla beraber çaba, katılım, ilgi, tutum gibi diğer akademik olmayan faktörlerin de notlandırma için sıklıkla kullanılan kriterler olduğunu göstermektedir.

McMillan (2008) öğretmenlerin aynı not verme ölçeğini ve ilkelerini kullanma durumlarında bile, okullar arasındaki notlandırmalar arasında çok az tutarlılık olduğunu belirtmektedir. Dockery (1995), not verme uygulamalarının sıklıkla keyfi olduğuna ve öğretmenden öğretmene farklılık gösterdiğine inanmaktadır. Bu durumun oluşmasında öğretmenlerin profesyonel kararlar vermelerini gerektiren sınıf ortamında sürekli olarak ortaya çıkan ve öngörülemeyen benzersiz durumların varlığı, farklı faktörlerin öğretmenler tarafından notlandırmada farklı ağırlıklandırılmasının sebep olduğu söylenebilir. Ölçme uzmanları çeşitli not verme uygulamalarını önermelerine rağmen, öğretmenler akademik ve akademik olmayan faktörlere verdikleri önem temelinde kendi uygulamalarını geliştirirler.

Brookhart (1991, 1993) önerilen uygulamalar ile öğretmenlerin gerçek puanlama uygulamaları arasındaki uyuşmazlığın bir geçerlik probleminin belirtisi olduğunu düşünmektedir. Bu çalışma kapsamında öncelikle öğretmenlerin geçerli ölçme yapıp yapmadıklarını belirlemek amacıyla geçerli ölçmeler ölçeği geliştirilmiştir. Daha sonrasında geliştirilen ölçek yardımıyla öğretmenlerin ölçme davranışları cinsiyet, branş, kıdem ve ölçme bilgisi değişkenleri açısından karşılaştırılmıştır.

Yöntem

Öğretmenlerin yaptıkları ölçme işlemlerinin geçerli olup olmadığını ve söz konusu bu davranışların cinsiyet, branş, kıdem ve ölçme bilgisi değişkenleri açısından farklılaşıp farklılaşmadığını belirlemeyi amaçlayan bu çalışma nicel araştırma yöntemlerinden betimsel tarama ile dizayn edilmiştir. Betimsel tarama araştırma türleri bir örneklem veya araştırma grubunun çalışmalar neticesinde tutum, eğilim, davranış ve görüşlerinin belirlenmesi ve bunun nicelleştirilmesini ifade etmektedir (Creswell, 2013). Karakaya (2012), betimsel araştırmanın geniş bir grup üzerinde herhangi bir olay ya da olgu ile ilgili olarak tutumlarının alındığı ve betimlendiği çalışmalar olarak görmektedir. Araştırmanın verileri ise devlet okullarında görev yapan farklı demografik özelliklere sahip ortaokul öğretmenlerinden toplanmıştır. Araştırmanın veri toplama aracını araştırma sırasında geliştirilen "Geçerli Ölçmeler Ölçeği" oluşturmaktadır. Söz konusu ölçek geliştirme sürecinde 589 öğretmenden elde edilen veriler kullanılmıştır. 20 madde 4 alt faktörden oluşan ölçek tüm ölçek gelitirme süreçlerine uygun olarak psikometrik nitelikleri kanıtlanmıştır. Ölçek geliştirme süreci sonucu gerçekleştirilen ortalama karşılaştırma testlerinde 287 öğretmenden elde edilen veriler kullanılmıştır. Söz konusu karşılaştırmalara geçilmeden önce tüm alt gruplarda ilgili alt faktör puanlarını dağılımları incelenmiştir. Bir puan dağılının normalliğine karar vermede birden fazla yöntemden elde edilen sonuçlar kullanışmıştır. Normalliğe karar vermede Kolmogorov-Smirnov testi, basıklık-çarpıklık katsayıları ve Z puanları kullanılmıştır. İlgili ölçütlerden çoğunluğu sağlayan puanlar normal kabul edilmiştir. Araştırmada tüm alt düzeyleri normal kabul edilen değişkenlerin alt düzey sayısı iki ise bağımsız gruplar t testi, ikiden fazla alt düzeye sahipse

tek yönlü varyans analizi (Anova), en az bir alt düzeyi normallik göstermeyen (çarpık) değişkenlerin alt düzey sayısı iki ise Mann Whitney U testi, ikiden fazla alt düzeye sahipse Kruskal Wallis H testi uygulanmıştır. Bu araştırma, eğitim ortamlarının en önemli ögelerinden olan ölçme sürecinin öğretmenler tarafından nasıl uygulandığının ve algılandığının belirlenmesi, geçerli notlandırmalar yapıp yapmadığı konusunda bilgilerin elde edilmesi konusunda önemli görülmektedir.

Sonuçlar

Araştırma kapsamında öncelikle öğretmenlerin ölçme davranışlarının geçerliğini inceleyen "Geçerli Ölçmeler Ölçeği" adı altında psikometrik nitelikleri kanıtlanmış bir ölçek geliştirilmiştir. Söz konusu ölçek 4 faktörden oluşmakta olup Notlandırmaya Diğer Değişkenlerin Karışımı (NDK), Doğru Puanlama İşlemi (DPİ), Gerçek Performansın Gösterimi (GPG) ve Hatırlamanın Nota Karışması (HNK) olarak adlandırılmıştır. Toplam 20 maddeden oluşan ölçeğin, toplam varyansın %54.37'sini açıklayan dört faktörlü yapının kuramsal temele uygun olduğu bulunmuştur. Ölçeğin doğrulayıcı faktör analizi sonucu elde edilen uyum değerlerinin, yakınsak-ayrışım geçerlik ve güvenirlik değerlerinin iyi olduğu belirlenmiştir. Geliştirilen ölçek kullanılarak öğretmenlerin geçerli ölçme davranışları cinsiyet, branş, kıdem ve ölçme bilgisi değişkenleri açısından karşılaştırılmıştır. Buna göre branş alt gruplarında yapılan karşılaştırmalarda NDK ve HNK alt faktörlerinde anlamlı farklılıklar bulunurken DPİ ve GPG alt faktörlerinde benzer puanlara sahip oldukları gözlenmiştir. Cinsiyet alt gruplarında yapılan incelemelerde kadın ve erkekler arasında NDK, HNK ve DPİ alt faktörlerinde anlamlı farklılaşma bulunurken GPG alt faktöründe böyle bir farklılık görülmemiştir. Öğretmenlerin kıdem alt boyutlarında yapılan karşılaştırmalar incelendiğinde NDK, HNK ve DPİ alt faktörlerinde yüksek kıdem düzeyinin geçerli ölçmeler alt faktörlerinde daha avantajlı hale geldiği ancak GPG alt faktöründe ise herhangi bir farklılığa neden olmadığı bulunmuştur. Son olarak öğretmenlerin yeterli ölçme bilgisine sahip olup olmama durumunun geçerli ölçmeler alt boyutlarına olan etkisi incelenmiş ve tüm alt boyutlarda ölçme bilgisi iyi olan öğetmenlerin iyi olmayanlara göre daha geçerli ölçmeler yaptığı test edilmiştir.

Kaynaklar

Atılgan, H. (2014). *Eğitimde ölçme ve değerlendirme.* Anı Yayıncılık.

Baykul, Y. (2010). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması.* Pegem Akademi Yayıncılık.

Brookhart, S. M. (1991). Letter: Grading practices and validity. *Educational Measurement: Issues and Practice*, *10*(1), 35-36. https://doi.org/10.1111/j.1745-3992.1991.tb00182.x

Brookhart, S. M. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement*, *30*, 123-142.

Brookhart, S. M. (1994). Teachers' grading: Practice and theory. *Applied Measurement in Education*, *7*(4), 279–301.

Brookhart, S. M. (2009). *Exploring Formative Assessment. The Professional Learning Community Series.* Association for Supervision and Curriculum Development. 1703 North Beauregard Street, Alexandria, VA 22311-1714.

Creswell, J. W. (2017). *Araştırma deseni: Nitel, nicel ve karma yöntem yaklaşımları* [Research Design: Qualitative, Quantitative, and Mixed Methods approaches]. (S. B. Demir, Çev. Ed.). Eğiten Kitap (2013).

Dockery, E. R. (1995). Better grading practices. *The Education Digest, 60*(5), 34-37.

Gallagher, J. D. (1998). *Classroom assessment for teachers.* Prentice Hall.

Guskey, T. R. (2006). Making high school grades meaningful. *Phi Delta Kappan*, *87*(9), 670-675. https://journals.sagepub.com/doi/pdf/10.1177/003172170608700910?casa_token=Xf8-WV5jC8QAAAAA:59LG5MO0bPpNYtu2N-ZXy-3Pt9eiKxQsYF-7VpMNX_5xOP6nrh2wIhchERiKNyBE47GIWvUncKg

Karakaya, İ. (2012). Bilimsel araştırma yöntemleri. A. Tanrıöğen (Ed.) *Bilimsel araştırma yöntemleri* içinde. Anı Yayıncılık

McMillan, J. H. (2008). *Assessment essentials for standards-based education.* Corwin Press.

Öğretmenlerin 21. Yüzyıl becerileri ve STEM etkinliklerini ölçme değerlendirmeye yönelik görüşleri

Çağrı Avan ve Bahattin Aydınlı

Anahtar kelimeler: STEM, 21. yüzyıl becerileri, öğretmen görüşleri

Giriş

Eğitim sistemleri zaman içerisinde kendini yenilemektedir. Bu süreç yeni yaklaşımları ve buna bağlı olarak değişiklikleri de beraberinde getirmektedir. Özellikle değişen dünyanın ve bu dünyada etkin bir vatandaş olacak bireylerin gereksinimlerini sağlamak için süreç temelli ve üretime yönelik yaklaşımlar geliştirilmektedir.

STEM eğitim yaklaşımı son dönemlerde öğretmenler tarafından sıklıkla kullanılmaya başlanmıştır. Bu yaklaşım, içerisinde farklı becerileri barındırmakta ve özellikle bireyin bilimsel sorgulama sürecini hayat problemleri üzerinde kullanmasını gerekli kılmaktadır. Bu noktada öğretmenlerin disiplinler arası bir bakış açısıyla öğrencilerde bulunması istenen günümüz yetkinliklerine (21. yüzyıl becerileri, yaşam becerileri vb.) odaklanmasını gerekli kılmaktadır.

Öğretmenler STEM eğitimi konusunda formal ve informal olarak kendilerini geliştirmeye çalışsa da özellikle beceri temelli ölçümler konusunda yeterli bilgiye sahip değillerdir. Bu çalışma kapsamında öğretmenlerin 21. yüzyıl becerileri ve STEM etkinliklerini ölçme ve değerlendirme konularındaki görüşlerini belirlemek amaçlanmıştır.

Yöntem

Çalışma kapsamında nitel bir yöntem kullanılmıştır. Verilerin toplanması sürecinde yapılandırılmış bir görüşme formu oluşturulmuş ve Türkiye'nin farklı yerlerinde görev yapan 379 öğretmene illerde bulunan ölçme değerlendirme merkezleri yoluyla ulaşılmıştır. Yapılandırılmış form google form aracılığıyla uygulanmıştır. Form dâhilinde öğretmenlere bilimsel süreç becerilerini, STEM temelli uygulamaları, 21. yüzyıl becerilerini nasıl ölçtükleri ve değerlendirdikleri konularında sorular sorulmuştur. Toplanan veriler kod, kategori ve temalar şeklinde gruplandırılmış ve demografik bilgiler ile birlikte analiz edilmiştir.

Sonuçlar

Öğretmen yanıtları incelendiğinde öğrenme alanına göre STEM'in en iyi uygulandığı alanın fizik sonrasında kimya ve üçüncü olarak da astronomi olduğu belirlenmiştir. Öğretmenlerin ölçme değerlendirme uygulamalarında önceliği bilişsel alana verdikleri görülmektedir. İkinci olarak ise psikomotor alan gelmektedir. Duyuşsal alana ise çok az odaklanılmaktadır.

Beceriler boyutunda bilimsel süreç becerilerinin STEM ile doğrudan ilişkili olduğu belirlenmiştir. Öğretmenlerin yirmi birinci yüzyıl becerilerinin ise STEM ile doğrudan ilişkili olmadığı düşüncesinde oldukları belirlenmiştir. Özellikle okuma ve dil becerileri, Medya ile ICT okuryazarlığının STEM becerileri için çok fazla gerekli olmadığını düşündükleri anlaşılmaktadır.

Kaynaklar

Aşık, G., Doğanca Küçük, Z. ve Çorlu, S. (2017). STEM-FETEMM eğitiminde ölçme değerlendime yaklaşımı. S. Çorlu, *STEM kuram ve uygulamaları* içinde (s. 21-36). Pusula.

Baykal, M. (2017). Türkiye yeterlilikler çerçevesi (Tyç)'nin öğrenci değerlendirme programı (PISA) açısından değerlendirilmesi. *Edu 7: Yeditepe Üniversitesi Eğitim Fakültesi Dergisi*, *6*(8), 69-79.

Constantinou, C. P., Tsivitanidou, O. E., and Rybska, E. (2018). What is inquiry-based science teaching and learning? In C. P. Constantinou, O. E. Tsivitanidou, and E. Rybska, *Professional development for ınquiry-based science teaching and learning* (pp. 1-23). Springer.

Çorlu, S. (2017). STEM: Bütünleşik öğretmenlik çerçevesi. S. Çorlu (Ed.), *STEM kuram ve uygulamaları* içinde (ss. 1-10). Pusula.

Gökbayrak, S. ve Karışan, D. (2017). Stem etkinliklerinin fen bilgisi öğretmen adaylarinin bilimsel süreç becerilerine etkisi, *Batı Anadolu Eğitim Bilimleri Dergisi*, *8*(2), 63-84.

Guzey, S. S., Moore, T. J., Harwell, M., and Moreno, M. (2016). STEM integration in middle school life science: Student learning and attitudes. *Journal of Science Education and Technology*, *25*(4), 550-560.

İdin, Ş. (2017). STEM yaklaşımı ve eğitime yansımaları. E. Karademir (Ed.), *Örnek ve uygulama destekli fen öğretiminde disiplinlerarası beceri etkileşimi* içinde (ss. 255-286). Pegem Akademi.

Kruger, C. J., Scogin, S. C., and Jekkals, R. E. (2019). The STREAM program: Project-based learning inan outdoor contex. *Kappa Delta Pi Record*, *55*(2), 85-88. https://doi.org/10.1080/00228958.2019.1580987

Investigation of item pre-knowledge cheating using joint hierarchical modeling of responses and response times under different conditions

Ebru Balta and Celal Deha Doğan

Introduction

Aberrant testing behaviors on educational and psychological tests has been known to compromise the accuracy of results on assessments of student achievement and thus influence the inferences drawn from these scores (Cizek and Wollack, 2017; Meijer, 1997; van der Linden and Guo, 2008; van Krimpen-Stoop and Meijer, 2001). There are several common types of aberrant testing behaviors to detect, including answer-copying, pre-knowledge cheating, creative thinking, lucky guessing, and random responding (Cizek and Wollack, 2017; Haberman and Lee, 2017; Karabatsos, 2003; Kingston and Clark, 2014; Lee and Haberman; 2016; Sijtsma and Meijer, 1992; Sinharay, 2017b). Item pre-knowledge causes the emergence of aberrance response patterns. It occurs due to illegal access to some exam items in the item pools before the exam and thus memorizing the items by violating the exam security measures. Computer-based large-scale applications become widespread with the global COVID 19 pandemic period. With the rise of CBT, increasingly more attention has been given to response time (RT) in assessment practice and psychometric research. It is seen that various RT models have been developed as methods to identify both examinees who may have benefitted from item pre-knowledge and the items which have prior exposure. Several response time models have been development such as loglinear response time model (van der Linden et al., 1999), effective response time model (Meijer and Sotaridona, 2006), bayesian lognormal response time model (van der Linden, 2006), model that combines a response time model with an IRT model for purposes of simultaneously modeling item responses and RTs (van der Linden, 2007), and mixture model (Schnipke and Scrams, 1997; von Davier and Rost, 2007; Wang and Xu, 2015). van der Linden (2007), proposed a joint model of an IRT and a response time model within a hierarchical framework, called as H-IRTRT in this study. In this joint modeling approach, response accuracy (RA) and RT data have been jointly modeled using a hierarchical latent variable model. The connection of RT patterns with patterns of RA will certainly increase the power of detecting aberrant behavior (van der Linden & Guo, 2008). Person-fit statistics are proposed for joint models to detect aberrant RA and RT patterns (Marianti et al., 2014; Man et al., 2018; Fox & Marianti, 2017).In this study, the log-likelihood of the responses and RTs is referred to as a person-fit statistic, denoted as $L(l_s{}^y)$ and $L(l^t)$ and Kullback-Leibler divergence measure are used in order to

evaluate the fit of RA and RT patterns. In this study, it is aimed to reveal in the detection of preknowledge cheating behaviors the rate of Type I error and the rate of power of response time based Kullback-Leibler divergence measure and *L* person fit statistic under different conditions via modeling patterns of response accuracy and response times using a joint hierarchial model.

Method

In line with the purpose of the study, 200 data sets were generated with 50 replication that item responses were modeled with three parameter logistic model and response times were modeled with the Bayesian log-normal response time model under the conditions of test length (15, 50) and difficulty level of the compromised items (medium difficulty, difficult) for obtained means of Type I error rate of methods. In order to obtain the mean power rate of the methods, 1800 data sets were generated with 50 replication under the conditions of test length, the difficulty level of the compromised items, the ratio of compromised items (20%, 40%, 60%), changing in response time of the compromised items (from a uniform distribution in the range of 10 to 15 seconds, fix to 20 seconds, fix to 30 seconds). Gibbs sampling algorithm is used as Monte Carlo Markov Chain (MCMC) approach in estimating patterns of response accuracy and response times using a joint hierarchial model parameters for each data set. In the analysis of the data, in the cheating scenario, who have item preknowledge individuals were selected from individuals with low ability level, and the rate of fraudulent data was created as 5% of the 1000 sample size, were examined of the methods of performance of pre-knowledge cheating. R.4.0.1 software was used for data generation and analyses.

Results

The performance of each person-fit statistic and *Kullback-Leibler* divergence for item responses and RTs were evaluated under each condition separately. As a result of the study, it was found that the *Kullback-Leibler* divergence measure was stronger than the *L* person fit statistics in detecting aberrant patterns of item response and response times patterns under the conditions of test length difficulty level of the compromised items, the ratio of compromised items, changing in response time of the compromised items; however, it was found to have a high Type I error rate. Thus, it is seen that there is a need for studies to reduce the Type 1 error rate of the Kullback Leibler divergence measure.

References

Cizek, G., and Wollack, J. (2017). Identification of item preknowledge by the methods of information theory and combinatorial optimization. In G. Cizek and J. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests,* (pp.217-233). Routledge.

Fox, J.-P., and Marianti, S. (2017). Person-fit statistics for joint models for accuracy and speed. *Journal of Educational Measurement*, *54*(2), 243–262. https:doi.org/10.1111/jedm.12143

Haberman, S., and Lee, Y. (2017). *A statistical procedure for testing unusually frequent exactly matching responses and nearly matching responses.* (Research Report No: RR-17-23). EducationalTestingService. https://doi.org/10.1002/ets2.12150

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*(4), 277–298. https://doi.org/https://doi.org/10.1207/s15324818ame1604_2

Kingston, N., and Clark, A. (2014). *Test fraud: Statistical detection and methodology.* Routledge.

Lee, Y., and Haberman, S. (2016). Investigating test-taking behaviors using timing and process data. *International Journal of Testing, 16*(3), 240–267. https://doi.org/10.1080/15305058.2015.1085385

Man, K., Harring. J., Quayang.Y., and Thomas, S. (2018). Response time based nonparametric Kullback-Leibler divergence measure for detecting aberrant test taking behavior. *International Journal of Testing, 18*(2), *155-177.* https://doi.org/10.1080/15305058.2018.1429446

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., and Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, *39*(6), 426–451. https://doi.org/10.3102/1076998614559412

Meijer, R. R. (1997). Person-Fit research: An introduction. *Applied Measurement in Education, 9*(1), 3-8. https://doi.org/10.1207/s15324818ame0901_2

Meijer, R. R., and Sotaridona, L. S. (2006). *Detection of advance item knowledge using response times in computer adaptive testing.* Law School Admission Council. *(LSAC Research Report 03-03).* https://ris.utwente.nl/ws/portalfiles/portal/5129730/LSAC_CT-03-03.pdf

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*(3), 213-232. http://www.jstor.org/stable/1435443

Sijtsma, K., and Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's non-parametric IRT model. *Applied Psychological Measurement, 16*(2), 149–157. doi:10.1177/014662169201600204

Sinharay, S. (2017b). Which statistic should be used to detect item preknowledge when the set of compromised items is known? *Applied Psychological Measurement, 41*(6), 403 421. https://doi.org/10.1177/0146621617698453

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72,* 287–308. https://doi.org/10.1007/s11336-006-1478-z

van der Linden, W. J., and Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73,* 365–384. https://doi.org/10.1007/s11336-007-9046-8.

van der Linden, W. J., Scrams, D. J., and Schnipke, D. L. (1999). Using response-time constraints to control for speededness in computerized adaptive testing. *Applied Psychological Measurement, 23*(3), 195-210. https://doi.org/10.1177/01466219922031329

van Krimpen-Stoop, E. M. L. A., and Meijer, R. R (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics*, *26*(2), 199–217. https://doi.org/10.3102/10769986026002199

von Davier, M., and Rost, J. (2007). Mixture distribution item response models. In C. R. Rao, and S. Sinharay (Eds.) *Handbook of statistics,* (pp. 643-661). Elsevier.

Wang, C., and Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456-477. https://doi.org/10.1111/bmsp.12054

4. Sınıf öğrencilerinin matematiksel akıl yürütme becerisine ilişkin puanlarının karar ağaçları ve sinir ağları algoritmalarıyla sınıflandırılması: TIMMS 2019 Türkiye örneği

Fatma Gül Uzuner ve Tuncay Öğretmen

Giriş

Düşünme, simgesel olarak uyarı ile tepki arasındaki boşluğun doldurulmasıdır. Belli kurallara göre yapılan düşünme ise muhakemedir (Morgan, 2017). İngilizce'de reasoning kelimesinin Türkçe karşılığı muhakeme ya da akıl yürütme şeklinde ifade edilmektedir (Yavuzsoy-Köse, 2016). Akıl yürütme becerisi; matematik eğitiminde matematik yaparken, matematiği öğrenirken veya öğretirken kullanılan temel becerilerden biridir. Aynı zamanda bu beceri, bireyin yaşamının erken dönemlerinde dikkat edilmesi gereken ve gittikçe her seviyede değişim gösteren bir beceridir (Olkun ve Toluk-Uçar, 2006). Dolayısıyla bireylerin erken yaşlarda temel matematiksel düşüncelerin gelişmesi noktasında nesne, dil, sembol ve resimler şeklinde ifade edilen dört temel yaşantıdan geçmeleri yararlı olacaktır (Haylock ve Cockburn, 2014). Bu bağlamda, özellikle okul öncesi eğitim veya erken çocukluk eğitimi kapsamında bireyin sonraki eğitim yaşantılarının temeli oluşturulmalı ve çocukların bütünsel olarak gelişimine katkı sağlanmalıdır (Günsoy, 2009). Nitekim erken çocukluktaki yaşantıların beyin gelişimi üzerinde oldukça önemli bir etkisi olduğu bilinmekle birlikte; bu yaşantılar aynı zamanda bireyin öğrenmesini, sağlığını, davranışını ve gelirini bile etkileyebilmektedir (The World Bank, 2021). Bu noktada, geniş kitleler üzerinde erken öğrenme sürecinin nasıl etkileri olduğunu daha görebilmek için farklı araştırma yaklaşımlarından yararlanılabilir. Veri madenciliği de bu yaklaşımlardan biri olduğu düşünülebilir.

Veri madenciliği, basitçe geniş ölçekli bir veriden bilginin çıkarılması ya da madencilik (mining) anlamına gelmektedir. Literatürde veri madenciliği anlamına gelen (bilgi madenciliği, bilgi çıkarma, veri/kalıp analizi, veri arkeolojisi ve veri tarama vb.) farklı terimler de kullanılmaktadır (Han ve diğ., 2012). Veri madenciliği klasik istatistiki yöntemlerdeki varsayımları gerektirmeyen, verilerin içindeki gizli bağlantıları ortaya koyan, bunlardan hareketle geleceğe yönelik kestirimde bulunmaya olanak sağlayan devimsel bir süreçtir (Altaş ve Gülpınar, 2012). Bir bilgi keşfi olarak veri madenciliği bazı adımların tekrarlı dizilerinden oluşur. İlk olarak gürültülü ya da alakasız veriler kaldırılarak veri temizleme işlemi yapılır. Temizleme işleminin ardından, birbiriyle entegre olabilecek veriler birleştirilir. Daha sonra veri seçimi yapılarak veriler madencilik için uygun hale getirilir. Bunun ardından, verilerin özüne inilerek verilere ilişkin örüntüler ortaya çıkarılmaya çalışılır ve çıkarılan örüntülerin

değerlendirmesi yapılır. Bu sürecin sonunda da ulaşılan bilgiler sunulur (Han ve diğ., 2012). Bu süreçte, farklı tekniklerin kullandığı söylenebilir.

Veri madenciliği sürecinde kullanılan tekniklerden biri de karar ağacı (decision tree) tekniğidir. Bu teknik basit bir yapıya, görselleştirme mekanizmasına ve farklı değişkenlere karşı sağlam algoritmalara sahip olması nedeniyle veri madenciliğinde sıkça kullanılmaktadır (Altaş ve Gülpınar, 2012). Karar ağaçları bilginin keşfinde ve veri madenciliğinde en güçlü yaklaşımlardan biridir (Bhargava ve diğ., 2013). Bu ağaçlar, ağacın kökünden başlayıp yapraklarına kadar inen yapılardır (Quinlan, 1986). Karar ağacı, kök (root) adı verilerin düğümlerden (node) oluşan örnek uzayın öz yinelemeli bir bölümü olarak ifade edilen bir sınıflandırıcıdır. Karar ağacındaki gelen kenarları olmayan düğüm, karar ağacının köküdür. Giden kenarları olan düğümlere test düğümleri (test node) ya da iç düğümler (internal node) adı verilir. Bunların dışında son düğüm (terminal node) ya da karar düğümü (decision node) olarak da bilinen düğümlere yaprak adı verilir. Düğümler test ettiği niteliğe göre etiketlenirken düğümden çıkan dallar da karşılık gelen değerlerle etiketlenir. Buna göre, bir analist karar ağacına bakara tahminlerde bulunabilir ve ilgili popülasyonun davranış özelliklerini anlayabilir (Rokach ve Maimon, 2010). Bu bağlamda, karar ağaçlarına ilişkin olarak tek değişkenli ve çok değişkenli iki yaklaşımın olduğu görülmektedir. Tek değişkenli karar ağacında bölme işlemi iç düğümlerdeki bir iç özelliğe göre yapılır (Bhargava ve diğ., 2013). Tek değişkenli karar ağaçlarının oluşturulmasında bazı temel adımlar vardır. Öncelikle tüm durumların aynı sınıfa ait olup olmadığı kontrol edilmeli, her bir özellik için bilgi ve bilgi kazancı hesaplanmalı ve en iyi ayırıcı özellik bulunmalıdır (Korting, 2013 akt. Bhargava ve diğ., 2013). Bilgi kazancı girdiler ve çıktılar arasındaki ilişkinin ölçülmesinde kullanılır ve bilgi kazancının hesaplanması sürecinde veri düzensizliğinin bir ölçüsü olan entropi (entropy) esas alınır. Bu noktada, örneğin küçük ve verimli bir ağaca ulaşmak için, bölme en yüksek kazanca göre yapılmalıdır. Ayrıca veri setinin iyi tanımlanmamış küçük alt kümeleri içerebilmesine ilişkin kullanılan budama tekniği de karar ağaçlarının oluşturulmasında kullanılan en önemli tekniklerdendir. Tek değişkenli karar ağacı tekniklerinden birisi de J48'dir (Bhargava ve diğ., 2013).

J48 (C4.5) sınıflandırıcı bir karar ağacı algoritmasıdır. Aynı zamanda J48 tüm karar ağaçları sisteminin adını aldığı bir programdır. Bu programda yaprak (leaf) bir sınıfı belirtir. Karar düğümü (decision node) ise testin her olası sonucu için bir dal ve alt ağaçla tek bir öz nitelik üzerinde gerçekleştirilecek testleri belirtir. J48 (C4.5) bu düğümün temelinde en fazla frekansı olan sınıfı kullanır (Quinlan, 1993). Karar ağaçlarından biri olan rastgele ağaç (random tree) tekniğinde öznitelikler rastgele bir şekilde seçilir ve herhangi bir budama tekniği içermeyerek hatayı en aza indirmeye çalışır. Rastgele ağaç algoritması, sınıflandırma için sınıf olasılıklarını tahmin etme seçeneğine sahiptir (Hamsagayathri ve Sampath, 2017). Bu algoritma, tüm verileri sınıflandırır ve yaprak olmayan düğümlerde özyinelemeli olarak çalışır (Kalmegh, 2015). Kısaca ilk önce veri olmadan rastgele bir şekilde karar ağaçlarının yapısı oluşturulur. Karar ağaçlarının tüm yapısı oluşturulurken test sürecinden geçirilmemiş bir özellik rastgele bir şekilde seçilir. Bunun akabinde, eğitim örnekleri tek tek incelenerek ağacın düğümlerine ilişkin değerler güncellenir (Fan ve diğ., 2003). Rastgele ağaçta budama süreci yoktur ve veriye bağlı olarak nasıl

sınıfların olabileceğine imkân veren bir yapısı vardır (Akçetin ve Çelik, 2014). Bazı öğrenme algoritmalarında da zaten yerleşik bir rastgele bileşen bulunur. Örneğin, geri yayılım algoritmasını kullanarak çok katmanlı algılayıcıları (multilayer perceptron) öğrenirken ağ ağırlıkları rastgele seçilen küçük değerlere ayarlanır (Ian ve Eibe, 2005) ve bu şekilde örnekleri sınıflandırır (Kara ve Şamlı, 2021).

Algılayıcı (perceptron) en eski denetimli (supervised) algoritmalardan biri olup yapay sinir ağının nöronudur. Derin sinir ağlarının örüntü tanıma noktasında çok iyi olduğu kanıtlandığı için algılayıcı kavramının altında biyolojik sinir ağlarının mantığı yatmaktadır. Tek katmanlı algılayıcılar doğrusal olmayan ayrılabilen fonksiyonları kullanarak modellenememe problemi vardır. Ancak çok katmanlı algılayıcılarda (multilayer perceptron) veya diğer adıyla ileri beslemeli sinir ağlarında (feedforward neural network) böyle bir problem yoktur. Çok katmanlı algılayıcı giriş, çıkış ve bir veya daha fazla gizli katmandan oluşan yapay bir sinir ağıdır (Kaluza, 2016). Giriş düğümlerindeki nöronlara bağımsız değişken, çıkış düğümlerindeki nöronlarına ise bağımlı değişken denebilir. Giriş katmanındaki nöronlar sadece veri modellerini alır ve bunları sonraki katmana iletir. Gizli katmandaki nöronlar ise giriş ve çıkış katmanları arasındaki bağlantıyı öğrenmek ve bu bağlantının haritasının çıkarılmasında etkin rol oynar. Bu bağlamda, çok katmanlı ağlarda birden fazla gizli katman olabilir ancak genellikle bir katman kullanır (Zhang, 2010). Buradaki en önemli hususlardan biri yapay sinir ağlarının deneyimlerden öğrenebilmesi ve bilgiler arasında ilişkiler kurabilmesidir (Uğur ve Kınacı, 2006). Her bir katmandaki algılayıcılar bir sonraki katmandaki algılayıcılarla doğrudan bağlantılıdır ve bu bağlantıların ağırlıkları algılayıcıların ağırlıklarına benzer. Çok katmanlı algılayıcıları eğitmek için çıkış değerlerinin doğru değerlerle karşılaştırılmasına olanak sağlayan geri yayılım yaklaşımı popüler bir yaklaşımdır. Hata düzeyi belli bir eşiğin altına inene kadar eğitim döngüsü kapsamında her bağlantının ağırlıkları ağ üzerinden geri beslenir. Bu şekilde daha doğru modellere ulaşılabilir (Kaluza, 2016).

Karar ağaçları ve yapay sinir ağlarına ilişkin literatürde farklı araştırmaların olduğu ancak eğitim alanında ilgili araştırmaların çok fazla olmadığı söylenebilir. Bununla ilgili olarak, Doğan (2017) veri madenciliği ile ilgili üretilen lisansüstü tezlerin incelenmesine yönelik yaptığı araştırmada, tezlerin ağırlıklı olarak işletme, sayısal yöntemler ve yönetim bilişim sistemleri alanlarıyla ilgili olduğunu tespit etmiştir. Ayrıca ilgili araştırma dikkate alındığında eğitim alanıyla ilgili çok fazla tezin olmadığı görülmektedir. Bu bağlamda, Akgün ve Özek (2020) yaptıkları araştırmada, eğitsel veri madenciliği ile ilgili 102 çalışmayı incelemiştir. Buna göre konu bazında yapılan araştırmaların en fazla başarının tahmin edilmesi üzerinde yoğunlaştığı görülürken uygulama bazındaki çalışmalarda ise ağırlıklı olarak karar ağaçları ve sinir ağları üzerinde çalışıldığı görülmüştür. Bununla birlikte, WEKA'nın eğitsel veri madenciliği noktasında en fazla kullanılan analiz programlarından biri olduğu sonucuna ulaşılmıştır. Ayrıca okul öncesi ve ilköğretim düzeyinde örneklem seçme sıklığının çok fazla olmadığı da görülmektedir. Bu araştırmada, ilkokul öğrencilerine ve matematiksel akıl yürütme becerisine ilişkin verilerin incelenmesi söz konusu olduğu için bu yönüyle literatüre katkı sağlayacağı düşünülmüş ve aşağıdaki sorulara cevap aranmaya çalışılmıştır:

1. Matematik öğrenmeye ilişkin ilgi ve erken öğrenme süreçlerine göre (ilkokuldan önce ve ilkokul 1. sınıfta) ilkokul 4. sınıf öğrencileri matematiksel akıl yürütme becerileri bakımından nasıl sınıflandırılmaktadır?

2. Matematiksel akıl yürütme becerisi bakımından ilkokul 4. sınıf öğrencilerinin sınıflandırılmasında kullanılan bağımsız değişkenlerin önem dereceleri nasıldır?

3. Matematiksel akıl yürütme becerisi bakımından ilkokul 4. sınıf öğrencilerinin sınıflandırılmasında kullanılan algoritmaların hangisinin sınıflandırma derecesi en yüksektir?

Yöntem

Evrenden seçilen bir örneklem üzerinden araştırılan unsura yönelik betimleme (Creswell, 2014), genel tarama modelleri olarak ifade edilebilir (Karasar, 2015). Sosyal araştırmalarda büyük kitlelere ilişkin bir konu hakkında bilgi edinebilmek amacıyla tarama (betimsel) çalışmalar yapılabilmektedir (Can, 2018). Bu araştırmada da ilkokul 4. sınıf öğrencilerinin matematik başarısının sınıflandırılmasında matematik öğrenmeye ilişkin ilgisinin ve erken öğrenme sürecinin matematiksel akıl yürütme becerisi üzerinde ne derece etkili olduğunun belirlenmesi amaçlanmıştır. Bu amaçla yapılan araştırmada genel tarama modeli benimsenmiştir. Bu model kapsamında, araştırmanın örneklemini TIMMS 2019'a katılan ilkokul 4. sınıfta öğrenim gören 4028 öğrenci oluşturmaktadır.

TIMSS 2019 kapsamında ilkokul 4. sınıf öğrencilerinin matematiksel akıl yürütme becerilerinin sınıflandırılmasında hangi özelliklerin etkili olduğuna dair fikir edinebilmek amacıyla matematik öğrenmeye ilişkin ilgi, erken öğrenme ve öğrencilerin matematiksel akıl yürütme puanları ile ilgili olan TİMSS 2019 verileri incelenmiştir. TIMMS verilerinin incelenmesi sonucunda araştırma kapsamında ele alınan özellikler ile bu özelliklere ilişkin bilgiler Tablo 1'de verilmiştir.

Tablo 1

Matematikte Akıl Yürütme Becerisinin Sınıflandırılmasına Ilişkin Özellikler ve Maddeleri

Özellikler	Madde Sayısı	Maddeler
Matematik öğrenmeye ilişkin ilgi ölçeği	9	ASBM02A, ASBM02B, ASBM02C, ASBM02D, ASBM02E, ASBM0, ASBM02G, ASBM02H, ASBM02I
Liket tipi	4	Kesinlikle katılıyorum=1, Katılıyorum=2, Katılmıyorum=3 ve Kesinlikle katılmıyorum=4 (TIMMS sınıflandırması)
Erken öğrenme1 (İlkokuldan önce)	18	ASBH01A, ASBH01B, ASBH01C, ASBH01D, ASBH01E, ASBH01F, ASBH01G, ASBH01H, ASBH01I, ASBH01J, ASBH01K, ASBH01L, ASBH01M, ASBH01N, ASBH01O, ASBH01P, ASBH01Q, ASBH01R
Likert tipi	4	Sıklıkla=1, Bazen=2, Hiçbir zaman=3 (TİMMS sınıflandırması)
Erken öğrenme2 (İlkokul 1. sınıfta)	7	ASBH06A, ASBH06B, ASBH06C, ASBH06D, ASBH06E, ASBH06F, ASBH06G
Likert tipi	4	Çok iyi=1, Orta düzeyde iyi=2, Çok iyi değil=3, Hiç iyi değil=4 (TIMMS sınıflandırması)

(devam)

Tablo 1 *(devam)*

Özellikler	Madde Sayısı	Maddeler
Erken öğrenme3 (İlkokul 1. sınıfta)	3	ASBH07A, ASBH07B, ASBH07C
Seçenekler	4	100 veya daha fazlasına kadar sayar=4, 20'ye kadar sayar=3, 10'a kadar sayar=2, Sayamaz=1 (TİMMS sınıflandırması)
Akıl yürütme becerisi puanı	5	ASMREA01, ASMREA02, ASMREA03, ASMREA04, ASMREA05 Ortalaması kesme noktası alınarak 2 düzeyli kategorik değişken (yüksek=1, düşük=0) haline getirilmiştir.

Tablo 1'de verilen özelliklere ilişkin detaylı bilgiler şu şekildedir: Matematik Öğrenmeye İlişkin İlgi Ölçeği (MÖİİÖ): İlgi kavramı, bireyin neyi sevip sevmediği veya neyi tercih edip etmediği ile ilgilidir (Morgan, 2017). Araştırmada katılımcıların matematik öğrenmeye ilişkin ilgi seviyelerini tespit edebilmek amacıyla TIMSS öğrenci anketinde (Student Questionnaire) yer alan 9 madde faktör analizine (doğrulayıcı ve açımlayıcı) tabi tutulmuştur. MÖİİÖ'nün Cronbach's alfa güvenirlik katsayısı .88, KMO değeri .92 ve Barlett testi sonucunun istatistiksel olarak anlamlı olduğu görülmüştür (χ^2= 17516.71, sd= 36, p< .01). Açımlayıcı faktör analizi kapsamında MÖİİÖ'nün 18 maddeden oluşan tek boyutlu bir yapısı olduğu ve bu maddelerin toplam varyansın %53.81'ini açıkladığı tespit edilmiştir. Bu sürecin ardından da MÖİİÖ için doğrulayıcı faktör analizi yapılmıştır. Buna göre tek faktörlü ölçeğin model uyum indekslerine göre (χ^2/sd= 12.67, RMSEA= .05, NFI= 1, CFI= 1) geçerli olduğuna dair bilgilere ulaşıldığı söylenebilir (Schumacher ve Lomax, 2004).

Erken Öğrenme1 (EÖ1): Araştırmada katılımcıların ilkokuldan önce erken öğrenme seviyesini tespit edebilmek amacıyla TIMSS ev anketinde (Home Questionnaire-Early Learning Survey) yer alan erken öğrenme (ilkokuldan önce) ile ilgili 18 maddeye faktör analizine (doğrulayıcı ve açımlayıcı) tabi tutulmuştur. EÖ1'in Cronbach's alfa güvenirlik katsayısı .94, KMO değeri .96 ve Barlett testi sonucunun istatistiksel olarak anlamlı olduğu görülmüştür (χ^2= 38598,84, sd= 153, p< .01). Açımlayıcı faktör analizi kapsamında EÖ1'in 18 maddeden oluşan tek boyutlu bir yapısı olduğu ve bu maddelerin toplam varyansın %49.13'ünü açıkladığı tespit edilmiştir. Bu sürecin ardından da EÖ1 için doğrulayıcı faktör analizi yapılmıştır. Buna göre tek faktörlü ölçeğin model uyum indekslerine göre (χ^2/sd =10.40, RMSEA= .05, NFI= .99, CFI= .99) geçerli olduğuna dair bilgilere ulaşıldığı söylenebilir (Schumacher ve Lomax, 2004).

Erken Öğrenme2 (EÖ2): Araştırmada katılımcıların ilkokula başladığında (birinci sınıf) erken öğrenme seviyesini tespit edebilmek amacıyla TIMSS ev anketinde (Home Questionnaire-Early Learning Survey) yer alan erken öğrenme ile ilgili 7 madde faktör analizine (doğrulayıcı ve açımlayıcı) tabi tutulmuştur. EÖ2'nin Cronbach's alfa güvenirlik katsayısı .96, KMO değeri .91 ve Barlett testi sonucunun istatistiksel olarak anlamlı olduğu görülmüştür (χ^2=31528.43, sd=21, p< .01). Açımlayıcı faktör analizi kapsamında EÖ2'nin 7 maddeden oluşan tek boyutlu bir yapısı olduğu ve bu maddelerin toplam varyansın %79.06'sını açıkladığı tespit edilmiştir. Bu sürecin ardından da EÖ2 için doğrulayıcı faktör analizi

yapılmıştır. Buna göre tek faktörlü ölçeğin model uyum indekslerine göre (χ^2/sd =219.91, RMSEA= .02, NFI= .98, CFI= .98) geçerli olduğuna dair bilgilerin olduğu söylenebilir (Schumacher ve Lomax, 2004).

Erken Öğrenme3 (EÖ3): Araştırmada katılımcıların ilkokula başladığında (birinci sınıf) erken öğrenme seviyesini tespit edebilmek amacıyla TİMSS ev anketinde (Home Questionnaire-Early Learning Survey) yer alan erken öğrenme ile ilgili 3 madde faktör analizine (doğrulayıcı ve açımlayıcı) tabi tutulmuştur. EÖ3'ün Cronbach's alfa güvenirlik katsayısı .91, KMO değeri .73 ve Barlett testi sonucunun istatistiksel olarak anlamlı olduğu görülmüştür (χ^2= 8465.19, sd= 3, p< .01). Açımlayıcı faktör analizi kapsamında EÖ3'ün 3 maddeden oluşan tek boyutlu bir yapısı olduğu ve bu maddelerin toplam varyansın %84.20'sini açıkladığı tespit edilmiştir. Bu sürecin ardından da EÖ3 için doğrulayıcı faktör analizi yapılmıştır. Buna göre tek faktörlü ölçeğin model uyum indeksine göre (χ^2/sd= 0) modelin oldukça iyi bir uyum verdiği söylenebilir (Schumacher ve Lomax, 2004).

Matematiksel Akıl Yürütme Becerisi (Reasoning): Matematiksel akıl yürütme, bireyin yaşamda karşılaştığı matematiksel veya gerçek yaşam problemlerini mantıksal ve sistematik düşünmeyi içerir (Trends in International Mathematics and Science Study [TIMSS], 2019). Bu araştırmada, TİMMS kapsamında katılımcıların akıl yürütmeye becerisine ilişkin puanları bağımlı değişken olarak ele alınmıştır.

Bu araştırmanın amacı doğrultusunda, Türkiye'ye ilişkin ilgili verilere TIMMS'in resmi internet sitesinden ulaşılmış ve SPSS formatında verilere ulaşılmıştır. TIMMS verileri ilgili literatür bağlamında incelenerek ilkokul öğrencilerinin matematiksel akıl yürütme becerilerine etki edebilecek olan erken öğrenme ve matematik öğrenmeyle ilgili değişkenler incelenmiştir. İncelenen değişkenler ayrı bir şekilde dosyalanmıştır. Bu süreçte, ilgili dosyaya TIMMS'in açık veritabanından ulaşılması ve bu dosyaların kişisel herhangi bir bilgiyi içermemesinden dolayı araştırmanın etik bir problem içermediği ifade edilebilir.

Verilerin hazırlanmasından analizine kadar yapılan işlemler sırasıyla verilmiştir.

1) Veri analizine başlamadan önce verilerde hata kontrolü yapılmalı ve hatalı noktalarda gereken düzeltmelere yer verilmelidir (Pallant, 2017). Bu bağlamda, ilk aşamada araştırmaya dahil edilen tüm veriler herhangi bir olası hataya karşın kontrol edilmiştir. Veri setinde bir hata olmadığı tespit edilmiştir.

2) İkinci aşamada kayıp veri incelemesi yapılmıştır. Nitekim kayıp veriler araştırmanın sonuçlarını etkileyebilmektedir (Akbaş ve Koğar, 2020). Bu bağlamda, Akbaş ve Koğar'ın (2020) önerdiği biçimde, kayıp verilerin incelenmesi hususunda öncelikle veriler betimsel olarak ön incelemeye tabi tutulmuştur. Daha sonra kayıp veri içeren ve içermeyen satırların ortalama puanları karşılaştırılmıştır. Karşılaştırma sonucunda aralarında anlamlı farklılıkların olduğu bazı değişkenlere rastlanmıştır, p< .01. Bu durumda kayıp verilerin tamamen seçkisiz kayıp olmadığı sonucuna varılabilir. Bu sonucun ardından, başka bir kayıp veri inceleme yolu olan Little's MCAR testi kullanılmıştır. Little's MCAR testinin manidar bir sonuç verdiği görülmüştür, χ^2= 6717.837, sd=6131, p= .000. Bu durumda da verilerin tamamen seçkisiz

kayıp olmadığı sonucunda varılabilir. Bu bağlamda, veri setinde %15'ten az kayıp değerler olduğunda kayıp değerlerin yerine değer atanabilir (Çokluk ve diğ., Büyüköztürk, 2018) çözümünden yararlanılmıştır. Bu yöntemde, kayıp değerlerin yerine mevcut verilerden hesaplanan ortalama değerler atanır (Tabachnick ve Fidell, 2007). Bu şekilde, araştırmanın veri setindeki kayıp değerler yerine ortalama değerler atanmıştır.

3) Veri setindeki uç değerlerin incelenmesi noktasında normal dağılım eğrisine sahip histogram grafikleri incelenmiştir (Akbaş ve Koğar, 2020). Bu inceleme sonucunda, verilerin normal dağılım eğrisi altında ve dağılımlarının normal olduğu görülmüştür.

4) Matematik Öğrenmeye İlişkin İlgi Ölçeği (MÖİİÖ), Erken Öğrenme1 (EÖ1), Erken Öğrenme2 (EÖ2), Erken Öğrenme3 (EÖ3) ölçeklerinin geçerlik ve güvenirlik çalışmaları yapılmıştır. Bu kapsamda ölçülmek istenen yapının belirtilen ölçeklerle ne derece gerçekleştiğini belirleyebilmek amacıyla faktör analizi yapılmıştır (Tavşancıl, 2014). Faktör analizi kapsamında aynı kümede yer alan maddelerin temelinde yer alan yapıyı ortaya koyabilmek amacıyla açıklayıcı faktör analizi, bu yapıyı doğrulayabilmek amacıyla da doğrulayıcı faktör analizi yapılmıştır (DeVellis, 2014). Güvenirlik katsayısı olarak Cronbach Alfa güvenirlik katsayısı kullanılmış ve ölçeklerin güvenirlik düzeylerinin .90'nın üzerinde olduğu görülmüştür. Bu durumda, .90 ve üzeri bir Cronbach's alfa katsayısına sahip bir ölçeğin çok yüksek düzeyde güvenirliğe sahip olduğu söylenebilir (Özdamar, 2017).

5) TİMMS raporunda verilen beş faklı olası akıl yürütme puanları toplanmış ve ortalaması alınmıştır. Ortalama puan kesme noktası olarak alınmış ve ilgili puan kategorik bir değişkene dönüştürülmüştür. Buna göre ortalamanın üzerinde olanlar "yüksek=1" ve ortalamanın altında olanlar "düşük=0" olacak şekilde sınıflandırılmıştır. Literatürde puanların ortalamasının kesme noktası olarak alınmasına yönelik çalışmalar da mevcuttur (Aksu ve Güzeller, 2016; Kutlu ve diğ., 2011; Üstün ve diğ., 2019).

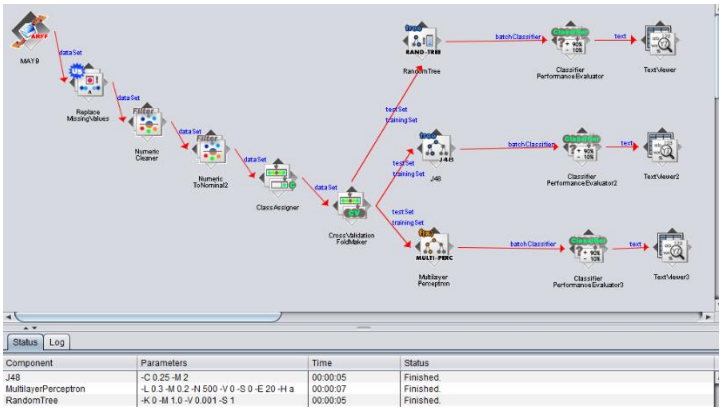
6) Bu çalışmada WEKA paket programının J48 (C4.5), rastgele ağaç ve çok katmanlı algılayıcı algoritmaları kullanılmıştır. Bu süreçte test seçenekleri noktasında verilerin istenilen parça sayısına bölünüp bir parçası test gerisi eğitim verisi olarak kullanılan çapraz-doğrulama yöntemi (cross-validation folds) kullanılmıştır (Aydemir, 2018). Buna göre, makine öğrenmesi bağlamında öncelikle veri ve problemin tanımlanması yapılmış, sonrasında veriler toplanarak ön işlemlerden geçirilmiş, modeller oluşturulmuş ve oluşturulan modeller değerlendirilmiştir (Kaluza, 2016). Veri madenciliği noktasında veri ön işleme süreçleri de dikkate alınmıştır (Tarkan ve diğ., 2011).

7) Bilgi akışları (knowledge flow), veri işleme süreci için görsel bir kullanıcı ara birimi uygular (Kaluza, 2016). Buna göre, yapılan araştırmanın bilgi akışı Şekil 1'de verilmiştir.

8) Verilerin analize hazırlanmasından analizine kadar bütün süreç ilk yedi maddede verilmiştir. Bu şekilde veri analiz süreci tamamlanmıştır.

Şekil 1

Çalışmanın Bilgi Akışı (Knowledge Flow)



Bu araştırmada, WEKA programı aracılığı ile erken öğrenme (EÖ1, EÖ2, EÖ3) ve matematik öğrenmeye ilişkin ilgi (MÖİİÖ) değişkenlerinin ilkokul 4. sınıf öğrencilerinin matematiksel akıl yürütme becerilerinin sınıflandırılmasında ne derece etkili olduklarını belirleyebilmek amacıyla J48, rastgele ağaç ve çok katmanlı algılayıcı algoritmaları kullanılmıştır. Kullanılan algoritmaların sınıflandırma sonuçları Tablo 2'de verilmiştir.

Tablo 2

Algoritmaların Öğrencilerin Matematiksel Akıl Yürütme Becerisine İlişkin Doğru Sınıflandırma Oranları

Algoritma Türü		Düşük=0	Yüksek=1	Başarı yüzdesi (%)	Algoritmanın sınıflandırma oranı
J48 (C4.5)	Düşük=0	1018	881	%53.61	%65.47
	Yüksek=1	510	1619	%76.05	
	Toplam (%)	%37.93	%62.97		
	Modelin oluşturulma süresi: 0.02 sn				
Rastgele Ağaç (Random Tree)	Düşük=0	1066	833	%56.13	%65.17
	Yüksek=1	570	1559	%73.23	
	Toplam (%)	%40.62	%59.38		
	Modelin oluşturulma süresi: 0 sn				
Çok Katmanlı Algılayıcı (Multilayer Perceptron)	Düşük=0	989	910	%52.08	%65.44
	Yüksek=1	482	1647	%77.32	
	Toplam (%)	%36.52	%63.48		
	Modelin oluşturulma süresi: 1.36				

Tablo 2'ye bakıldığında, *J48 algoritmasına göre* matematik akıl yürütme becerisi puanlarının düşük olduğu 1899 katılımcıdan 1018'inin (%53.61) doğru sınıflandırıldığı fakat 881 (%46.39) katılımcının matematik akıl yürütme becerisi puanlarının yüksek olmasına rağmen düşük olarak sınıflandırıldığı anlaşılmaktadır. Bununla birlikte, J48 algoritmasına göre, matematik akıl yürütme becerisi puanlarının yüksek olduğu 2129 katılımcıdan 1619'unun (%76.05) doğru sınıflandırıldığı fakat 510 (%23.95) katılımcının matematik akıl yürütme becerisi puanlarının düşük olmasına rağmen yüksek olarak

sınıflandırıldığı anlaşılmaktadır. Bu bağlamda, WEKA programı aracılığı ile J48 algoritmasının doğru sınıflandırma oranının %65.47 ve modelin oluşturulma süresinin 0.02 sn olduğu görülmektedir. Rastgele ağaç algoritmasına göre matematik akıl yürütme becerisi puanlarının düşük olduğu 1899 katılımcıdan 1066'sının (%56.13) doğru sınıflandırıldığı fakat 833 (%43.87) kişinin matematik akıl yürütme becerisi puanlarının yüksek olmasına rağmen düşük olarak sınıflandırıldığı anlaşılmaktadır. Ayrıca aynı algoritmaya göre, matematik akıl yürütme becerisi puanlarının yüksek olduğu 2129 katılımcıdan 1559'unun (%73.23) doğru sınıflandırıldığı fakat 570 (%26.77) katılımcının matematik akıl yürütme becerisi puanlarının düşük olmasına rağmen yüksek olarak sınıflandırıldığı anlaşılmaktadır. Bu bağlamda, WEKA programı aracılığı ile rastgele ağaç algoritmasının doğru sınıflandırma oranının %65.17 ve modelin oluşturulma süresinin 0 sn olduğu görülmektedir. Çok katmanlı algılayıcı algoritmasına göre matematik akıl yürütme becerisi puanlarının düşük olduğu 1899 katılımcıdan 989'unun (%52.08) doğru sınıflandırıldığı fakat 910 (%47.92) kişinin matematik akıl yürütme becerisi puanlarının yüksek olmasına rağmen düşük olarak sınıflandırıldığı anlaşılmaktadır. Ayrıca aynı algoritmaya göre, matematik akıl yürütme becerisi puanlarının yüksek olduğu 2129 katılımcıdan 1647'sinin (%77.32) doğru sınıflandırıldığı fakat 482 (%22.77) katılımcının matematik akıl yürütme becerisi puanlarının düşük olmasına rağmen yüksek olarak sınıflandırıldığı anlaşılmaktadır. Bu bağlamda, WEKA programı aracılığı ile çok katmanlı algılayıcı algoritmasının doğru sınıflandırma oranının %65.44 ve modelin oluşturulma süresinin 1.36 sn olduğu görülmektedir. Katılımcıların matematiksel akıl yürütme becerisinin sınıflandırılmasına ilişkin WEKA programı aracılığı ile kullanılan algoritmaların performans ve sınıflandırma ölçümleri Tablo 3'te verilmiştir.

Tablo 3

Algoritmaların Öğrencilerin Matematiksel Akıl Yürütme Becerisine İlişkin Performans ve Sınıflandırma Ölçümleri

Algoritma türü	Performans ölçüleri					Sınıflandırma doğruluğunun detayları: ağırlıklı ortalama değerleri (Weighted Avg.)		
	1. Kappa istatistiği	2. Ortalama mutlak hata	3. Ortalama hata karekök	4. Göreli mutlak hata %	5. Göreli hata karekök %	6. TP oranı	7. FP oranı	8. F ölçütü
J48 (C4.5)	.2997	.4423	.4708	88.759	94.3072	.655	.358	.650
Rastgele ağaç	.2958	.4352	.4676	87.323	93.6774	.652	.359	.649
Çok katmanlı algılayıcı	.298	.4262	.4737	85.519	94.9035	.654	.360	.648
Performans ölçütleri ve sınıflandırma doğruluğunun detayları	1. Kappa İstatistiği (Kappa Statistic)					Değerlendirme kriteri (Landis ve Koch, 1977, s. 165): < 0: Düşük (poor) 0.01 - 0.20: Önemsiz (slight) 0.21 - 0.40: Zayıf (fair) 0.41 - 0.60 Orta (moderate) 0.61 - 0.80: İyi (subtantial) 0.81 - 1.00: Çok iyi (almost perfect)		

(devam ediyor)

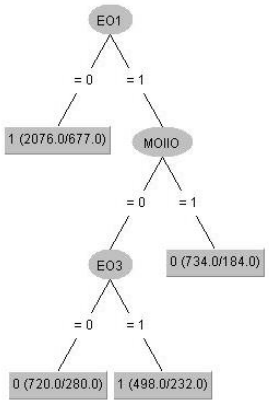
Tablo 3 *(devam)*

Performans ölçüleri					Sınıflandırma doğruluğunun detayları: ağırlıklı ortalama değerleri (Weighted Avg.)		
1. Kappa istatistiği	2. Ortalama mutlak hata	3. Ortalama hata karekök	4. Göreli mutlak hata %	5. Göreli hata karekök %	6. TP oranı	7. FP oranı	8. F ölçütü
2. Ortalama mutlak hata (Mean Absolute Error, MAE) 3. Ortalama hata karekök (Root Mean Square Error, RMSE) 4. Göreli mutlak hata (Relative Absolute Error, RAE) 5. Göreli hata karekök (Root Relative Squared Error, RRSE)					Değerlendirme kriteri (Graczy ve diğ., 2009): Minimum değer: 0, Maksimum Değer: ∞, İstenen Sonuç: Minimum olması		
6. Doğru pozitif oranı (True Positive Rate) 7. Yanlış pozitif oranı (False Positive Rate) 8. F-Ölçütü (F-Measure)					Doğru ve yanlış pozitif oranı, akıl yürütme becerisi bakımından yüksek (1) ve düşük (0) sınıflandırılmalarında doğru ve yanlış sınıflandırma oranlarının ortalama değerleridir. F-Ölçütü: "Kesinlik ve duyarlılığın harmonik ortalamasıdır." (Coşkun ve Baykal, 2011, s. 4). Bilgi alma noktasında algoritmaların performansını değerlendirir (Wood, 2019).		

Tablo 3'e bakıldığında, algoritmaların öğrencilerin matematiksel akıl yürütme becerisine ilişkin performans ve sınıflandırma ölçümleri görülmektedir. Öncelikle üç algoritma için Kappa istatistiği değerlerinin birbirine yakın olduğu ve bu değerler göz önüne alındığında genel olarak öğrencilerin akıl yürütme becerisine göre yüksek ve düşük sınıflandırma noktasındaki uyumun Landis ve Koch'a (1977) göre, zayıf olduğu ifade edilebilir. Kappa istatistiği noktasında, en yüksek uyum değerinin J48 algoritmasında olduğu görülmektedir. Ortalama mutlak hata ve ortalama hata karekök değerlerinin sıfıra kısmen yakın olduğu söylenebilir. Göreli mutlak hata değerinin çok katmalı algılayıcı algoritmasında en küçük olduğu ve göreli hata karekök değerinin de rastgele ağaç modelinde en küçük olduğu görülmektedir. Tablo 3'teki doğru ve yanlış pozitif oranları incelendiğinde ise katılımcıların akıl yürütme becerisi bakımından yüksek ve düşük sınıflandırılmalarında doğru sınıflandırma oranının ortalama değerlerinin .655 değeri ile J48 algoritmasında olduğu; yanlış sınıflandırma oranının ortalama değerlerinin .360 değeri ile çok katmanlı algılayıcı algoritmasında en yüksek olduğu görülmektedir. Algoritmaların performansı açısından da F-ölçütü değerleri incelendiğinde, en yüksek değerin .650 ile J48 algoritmasına ait olduğu görülmektedir. Katılımcıların matematiksel akıl yürütme becerisinin sınıflandırılmasına ilişkin WEKA programı aracılığı ile kullanılan J48 algoritmasının görseli Şekil 2'de, rastgele ağaç algoritmasının görseli Şekil 3'te ve çok katmanlı algılayıcı algoritmasının görseli Şekil 4'te verilmiştir.

Şekil 2

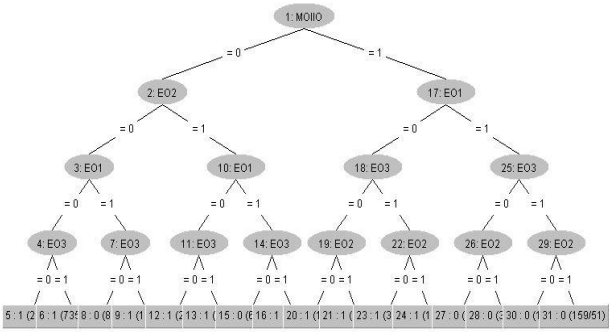
J48 Algoritmasına Göre Sınıflandırma Sonuçları



Şekil 2'ye bakıldığında, karar ağacının kökünü ilkokuldan önce ebeveyn tarafından çocukla gerçekleştirilen erken öğrenme aktivitelerinin (EÖ1) olduğu görülmektedir. Buna göre, ilkokuldan önce ebeveyn tarafından yapılan erken öğrenmeye yönelik aktivitelerin matematiksel akıl yürütme becerisinin sınıflandırılması üzerinde modele göre en etkili değişken olduğu ifade edilebilir. EÖ1 ölçeğinin içeriği; bir ebeveynin çocuğuna ilkokula başlamadan önce kitap okuması, hikâye anlatması, şarkı söylemesi, yaptığı şeylerden bahsetmesi, okuduklarından bahsetmesi, harf ya da kelime yazması, ritimli bir şekilde sayıları sayması, şekiller çizmesi, bir şeyleri tartması; çocuğu ile sayı oyuncakları ile oynaması, tahta ve kart oyunları oynaması, kelime oyunları oynaması ve alfabe içerikli oyuncaklarla oynaması şeklindedir. Şekil 2'ye göre, erken öğrenme aktivitelerinin düşük olduğu katılımcıların da matematiksel akıl yürütme becerilerinin yüksek olabileceği görülmektedir. Erken öğrenme aktivitelerinin yüksek olduğu katılımcılarda ise matematik öğrenme ilişkin ilginin olduğu görülmektedir. MÖİİÖ ölçeğinin içeriği; çocuğun matematik öğrenmekten hoşlanması, matematikte ilginç şeyler öğrenmesi, matematiği sevmesi, sayılarla ilgili ödevleri sevmesi, matematik problemlerini çözmekten hoşlanması, matematik derslerini heyecanla beklemesi, matematiğin favori derslerinden biri olması, matematiği sıkıcı bulması ve matematik çalışmak istememesi şeklindedir. Matematik öğrenmeye ilişkin ilgisi yüksek olan katılımcıların matematiksel akıl yürütme becerisinin düşük olabileceği de görülmektedir. Bununla beraber, matematik öğrenmeye ilişkin ilgisi düşük olan katılımcıların da ilkokul birinci sınıftaki sayma, yazılan sayıları fark etme ve sayıları yazma ile ilgili (EÖ3) becerilerine göre matematiksel akıl yürütme becerisinin şekillendiği söylenebilir. Buna göre, ilkokul birinci sınıfta katılımcıların sayma, yazılan sayıları fark etme ve sayıları yazma ile ilgili becerileri yüksek ise matematiksel akıl yürütme becerisinin yüksek, düşük ise matematiksel akıl yürütme becerisinin düşük olabileceği yorumu yapılabilir.

Şekil 3

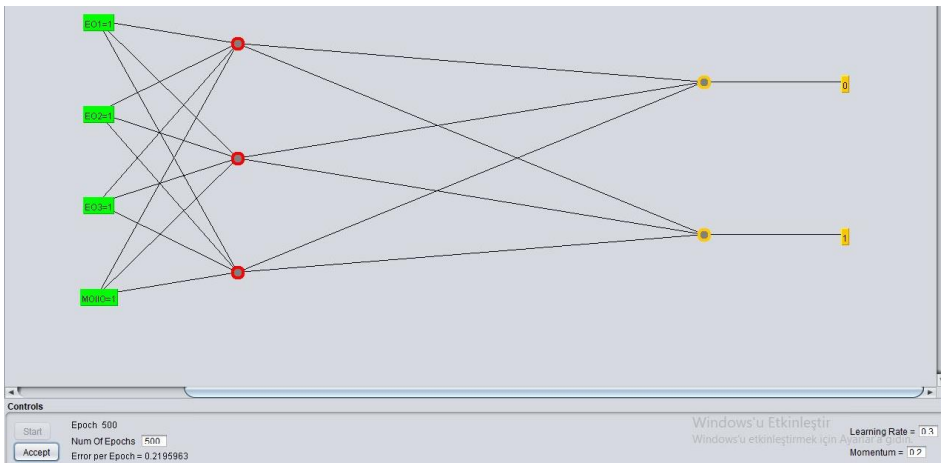
Rastgele Ağaç Algoritmasına Göre Sınıflandırma Sonuçları



Şekil 3'e bakıldığında, katılımcıların matematiksel akıl yürütme becerisinin sınıflandırılmasına ilişkin rastgele ağaç algoritmasının olası bir sonucu görülmektedir. Bu olası sonuçların J48 (C4.5) algoritmasından elde edilen sonuçlarla bazı noktalarda benzerlik gösterdiği söylenebilir. Bu bağlamda, erken öğrenme sürecindeki yaşantıların matematiksel akıl yürütüme becerisinin yüksek olması üzerinde etkili olabileceği olasılığı olduğu gibi bunun tersini de söylemek mümkündür.

Şekil 4

Çok Katmanlı Algılayıcı Algoritmasına Göre Sınıflandırma Sonuçları



Şekil 4'e bakıldığında, katılımcıların matematiksel akıl yürütme becerisinin sınıflandırılmasına ilişkin çok katmanlı algılayıcı algoritmasının uygulanması sonucu oluşan sinir ağı modeli görülmektedir. Giriş katmanın dört birimden (EÖ1, EÖ2, EÖ3, MÖİİÖ), gizli katmanın üç birimden ve çıkış katmanın iki birimden oluştuğu görülmektedir. Bununla birlikte, Şekil 4'te üç düğümden oluşan bir gizli katmanın olduğu ve buna göre matematiksel akıl yürütme becerisinin yüksek ve düşük şeklinde olduğu iki sınıflandırmaya ayrıldığı görülmektedir. Başka bir ifadeyle, çok katmanlı algılayıcı algoritmasının erken

öğrenme ve matematik öğrenmeye ilişkin ilgi değişkenleri ile matematiksel akıl yürütme becerisi arasındaki bağlantıların haritasını çizdiği söylenebilir. Bu harita üzerinde ilgili değişkenlerin bağlantıları incelenebilir.

Tartışma, Sonuç ve Öneriler

Bu araştırmanın temel amacı, TIMMS sınavına katılan ilkokul 4. sınıf öğrencilerin erken öğrenme süreçlerinin ve matematik öğrenmeye ilişkin ilgilerinin matematiksel akıl yürütme becerilerinin yüksek veya düşük şeklinde sınıflandırılması üzerindeki etkisini incelemektir. Bu bağlamda, okul öncesi eğitimin çocukların bilişsel ve sosyal gelişimlerini desteklediğine (Kandır ve Orçan, 2011), üst düzey akademik beceriler için okul öncesindeki erken akademik becerilerin çocuklara kazandırılmasının bir gereklilik olduğuna (Uyanık ve Kandır, 2010) ve okul öncesi eğitimi alan 48-66 aylık çocukların matematik yetenek puanlarının daha yüksek olduğuna (Avcı, 2015) ilişkin araştırmalar vardır. Örneğin Bozgün ve Uluçınar-Sağır (2018) okul öncesi eğitimi alan çocukların okuma ve yazmayı öğrenme sürecine ilişkin hazırbulunuşluk düzeylerinin okul öncesi eğitimi almayan çocuklara göre daha yüksek olduğu sonucuna ulaşmıştır. Bu sonuçlara göre, erken çocukluk dönemindeki süreçlerin ve matematik öğrenmeye ilişkin ilginin matematiksel akıl yürütme üzerinde etkili olabileceği de düşünülebilir. Ancak yapılan araştırmada elde edilen sonuçların beklenilenden bazı yönlerden farklılık gösterdiği söylenebilir. Öncelikle J48 algoritma sonuçları incelendiğinde, WEKA programı aracılığı ile J48 algoritmasının doğru sınıflandırma oranının %65.47 olduğu görülmektedir. Bu sınıflandırma oranı kapsamında, J48 algoritmasından elde edilen sonuçlara göre, karar ağacının kökünde yer alan okul öncesi eğitimin en etkili değişken olduğu görülmüştür. Buna göre, okula başlamadan önce çocuklarıyla değişik ve eğlenceli aktiviteler (hikaye anlatma, ritimli sayma, oyun oynama, şekiller çizme vb.) yapan ebeveynlerin çocuklarının matematik öğrenmeye ilişkin ilgilerinin olduğu ancak ilgili aktiviteleri yapmayan ebeveynlerin çocuklarının da matematiksel akıl yürütme becerilerinin yüksek olabileceği görülmüştür. Başka bir ifadeyle, ilkokuldan önce ebeveynleri ile eğlenceli ve farklı etkinlikler yapmayan çocukların matematiksel akıl yürütme becerilerinin yüksek, yapan çocukların da matematik öğrenmeye ilişkin ilgilerinin olabileceği görülmüştür. Matematik öğrenmeye ilişkin ilgi düzeyleri yüksek olan çocukların matematiksel akıl yürütme becerilerinin düşük olabileceği sonucu da ortaya çıkmıştır. Ancak matematik öğrenmeye ilişkin ilgi düzeyleri düşük olan çocukların ilkokul birinci sınıftaki temel becerilerine (sayma, yazılan sayıları fark etme ve sayıları yazma) göre matematiksel akıl yürütme becerilerinin şekillendiği de görülmektedir. Bu sonuçlara ilişkin olarak erken öğrenme sürecinde çocuklarla yapılan etkinliklerin ve çocukların matematik öğrenmeye ilişkin ilgilerinin ilkokul 4. sınıfta matematiksel akıl yürütme becerilerinin yüksek olması üzerinde yeterince etkili olmadığı yorumu yapılabilir. Bu durumda, başka değişkenlerin etkili olabileceği yorumu yapılabilir. Ancak doğru sınıflandırma oranının %65.17 olduğu rastgele ağaç algoritmasından elde edilen olası sonuca göre matematik öğrenmeye ilişkin ilginin yüksek olduğu çocuklardan erken öğrenme imkanlarına sahip olanların ilkokul birinci sınıfta daha başarılı olabileceği ve bunun sonrasında matematiksel akıl yürütme becerisinin yüksek olabileceği yorumu yapılabilir. İlginin motivasyonel bir değişken olarak herhangi bir durum, nesne vb. bir şeyle ilişki kurma noktasındaki bir

yatkınlık (Hidi ve Renninger, 2006) olduğu ve erken öğrenme sürecinin bilişsel ve sosyal gelişim üzerindeki etkisi (Kandır ve Orçan, 2011) düşünüldüğünde, yapılan yorumun tutarlı olduğu söylenebilir. Ayrıca rastgele ağaç algoritmasının olası sonuçlarına göre bu yorumun tersine bir yorum da yapılabilir. Bununla birlikte, doğru sınıflandırma oranının %65.44 olduğu çok katmalı algılayıcı algoritmasının sınıflandırma sonuçları dikkate alındığında erken öğrenme sürecinin ve matematik öğrenmeye ilişkin ilginin kendi içinde üç gizli yapıya ayrılarak matematiksel akıl yürütme becerisinin yüksek veya düşük olarak sınıflandırılmasında etkili olduğu söylenebilir. Bu sonucun J48 algoritmasından elde edilen sonuçla bir ölçüde tutarlı olduğu da söylenebilir. Ayrıca kullanılan algoritmalar içerisinde en iyi sınıflandırma oranını veren algoritmanın J48 olduğu görülmüştür.

Araştırmadan elde edilen sonuçlara göre gelecekte yapılacak olan çalışmalara yönelik bazı öneriler sunulabilir. Bu bağlamda, erken öğrenme sürecinde yapılan etkinliklerin ilkokul 4. sınıfta öğrenim gören öğrencilerin matematiksel akıl yürütme becerileri üzerinde ne derece etkili olduğuna ilişkin daha farklı araştırmalar yapılabilir. Ayrıca erken çocukluk döneminde yapılan etkinliklerin niteliğine ilişkin araştırmalar da yapılabilir. Bu bağlamda, matematiksel akıl yürütme becerisini, temel eğitimi ve eğitsel veri madenciliğini içeren daha geniş çaplı çalışmalar yapılmalıdır.

Kaynaklar

Akbaş, U. ve Koğar, H. (2020). *Nicel araştırmalarda kayıp veriler ve uç değerler: Çözüm önerileri ve SPSS uygulamaları.* Pegem Akademi.

Akçetin, E. ve Çelik, U. (2014). İstenmeyen elektronik posta (spam) tespitinde karar ağacı algoritmalarının performans kıyaslaması. *Journal of Internet Applications and Management, 5*(2), 43-56. https://doi.org/10.5505/iuyd.2014.43531

Akgün, K. ve Özek, M. B. (2020). Eğitsel veri madenciliği yöntemi ile ilgili yapılmış çalışmaların incelenmesi: içerik analizi. *Uluslararası Eğitim Bilim ve Teknoloji Dergisi, 6*(3), 197-213. https://doi.org/10.47714/uebt.753526

Aksu, G. ve Güzeller, C. O. (2016). PISA 2012 matematik okuryazarlığı puanlarının karar ağacı yöntemiyle sınıflandırılması: Türkiye örneklemi. *Eğitim ve Bilim, 41*(185), 101-122. http://dx.doi.org/10.15390/EB.2016.4766

Altaş, D. ve Gülpınar, V. (2012). Karar ağaçları ve yapay sinir ağlarının sınıflandırma performanslarının karşılaştırılması. *Trakya Üniversitesi Sosyal Bilimler Dergisi*, *14*(1), 1-22. https://dergipark.org.tr/tr/pub/trakyasobed/issue/30250/326695

Avcı, K. (2015). *Okul öncesi eğitimi alan 48-66 aylık çocukların matematik becerilerinin bazı değişkenler açısından incelenmesi* (Tez No. 381204) [Yüksek Lisans tezi, Çanakkale Onsekiz Mart Üniversitesi]. Yükseköğretim Kurumu Tez Merkezi.

Aydemir, E. (2018). *WEKA ile yapay zekâ.* Seçkin Yayıncılık.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246. https://doi.org/10.1037/0033-2909.107.2.238

Bhargava, N., Sharma, G., Bhargava, R., and Mathuria, M. (2013). Decision tree analysis on j48 algorithm for data mining. *Intrenational Journal of Advanced Research in Computer Science and Software Engineering, 3*(6), 1114-1119.

Bozgün, K., ve Uluçınar-Sağır, Ş. (2018). Okuma yazmayı öğrenme sürecinde okul öncesi eğitimin etkisi. *2nd International Symposium on Innovative Approaches in Scientific Studies, 3*, 1110-1115. http://www.set-science.com/manage/uploads/ISAS2018-Winter_0039/SETSCI_ISAS2018-Winter_0039_00214.pdf

Can, A. (2018). *SPSS ile bilimsel araştırma sürecinde nicel veri analizi* (6. baskı). Pegem Akademi.

Coşkun, C. ve Baykal, A. (2011). Veri madenciliğinde sınıflandırma algoritmalarının bir örnek üzerinde karşılaştırılması. *Akademik Bilişim*, 1-8. https://ab.org.tr/ab11/bildiri/67.pdf

Creswell, J. W. (2014). *Research design: Qualitative, quantitative and mixed methods approaches* (4th ed.). Sage.

Çokluk, Ö., Şekercioğlu, G. ve Büyüköztürk, Ş. (2018). *Sosyal bilimler için çok değişkenli istatistik SPSS ve LISREL uygulamaları* (5. baskı). Pegem Akademi.

DeVellis, R. F. (2014). Faktör analizi (D. Çakıcı-Eser, Çev.). T. Kotan (Ed.), *Ölçek geliştirme: Kuram ve uygulamalar* içinde (s. 115-158). Nobel Akademik Yayıncılık.

Doğan, O. (2017). Türkiye'de veri madenciliği konusunda yapılan lisansüstü tezler üzerine bir araştırma. *Gazi Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 19*(3), 929-951. https://dergipark.org.tr/tr/pub/gaziuiibfd/issue/36599/416570

Fan, W., Wang, H., Yu, P. S., and Ma, S. (2003). *Is random model better? On its accuracy and efficiency. The Third IEEE International Conference on Data Mining.* IEEE Computer Society.

Graczyk, M., Lasota, T., & Trawiński, B. (2009, October). *Comparative analysis of premises valuation models using KEEL, RapidMiner, and WEKA.* International Conference on Computational Collective Intelligence içinde (s. 800-812). Springer, Berlin, Heidelberg.

Günsoy, G. (2009). Beşeri sermaye ve insani gelişme için erken çocukluk eğitiminin önemi. *Bilgi Ekonomisi ve Yönetimi Dergisi, 4*(2), 23-43.

Hamsagayathri, P., and Sampath, P. (2017). Performance analysis of breast cancer classification using decision tree classifiers. *International Journal of Current Pharmaceutical Research, 9*(2), 19-25.

Han, J., Kamber, M., and Pei, J. (2012). *Data mining: Concepts and techniques.* Elsevier.

Haylock, D., and Cockburn, A. (2014). Matematiği anlama (S. Doğan, Çev.). Z. Yılmaz (Ed.), *Küçük çocuklar için matematiği anlama* içinde (4. baskı), (ss. 5-28). Nobel Akademik Yayıncılık (2013).

Hidi, S., and Renninger, K. A. (2006). The Four-Phase Model of Interest Development. *Educational Psychologist, 41*(2), 111-127.

Ian, H. W., and Eibe, F. (2005). *Data mining: Practical machine learning tools and techniques.* Elsevier Inc.

Kalmegh, S. (2015). Analysis of WEKA data mining algorithm REP tree, simple cart and random tree for classification of Indian news. *International Journal of Innovative Science, Engineering & Technology, 2*(2), 438-446. http://ijiset.com/vol2/v2s2/IJISET_V2_I2_63.pdf

Kaluza, B. (2016). *Machine Learning in Java.* Pact Publishing.

Kandır, A. ve Orçan, M. (2011). Beş-altı yaş çocuklarının erken öğrenme becerileri ile sosyal uyum ve becerilerinin karşılaştırmalı olarak incelenmesi. *İlköğretim Online, 10*(1), 40-50. https://dergipark.org.tr/tr/pub/ilkonline/issue/8593/106845

Kara, Ş. E. ve Şamlı, R. (2021). Yazılım projelerinin maliyet tahmini için WEKA'da makine öğrenmesi algoritmalarının karşılaştırmalı analizi. *Avrupa Bilim ve Teknoloji Dergisi*, (23), 415-426. https://doi.org/10.31590/ejosat.877296

Karasar, N. (2015). *Bilimsel araştırma yöntemi* (28. baskı). Nobel Akademik Yayıncılık.

Kutlu, Ö., Yıldırım, Ö., Bilican, S., and Kumandaş, H. (2011). İlköğretim 5. sınıf öğrencilerinin okuduğunu anlamada başarılı olup-olmama durumlarının kestirilmesinde etkili olan değişkenlerin incelenmesi. *Eğitimde ve Psikolojide Ölçme Değerlendirme Dergisi, 2*(1), 132-139. https://dergipark.org.tr/tr/pub/epod/issue/5806/77235

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-74. https://www.jstor.org/stable/2529310

Morgan, C. T. (2017). Psikolojik testler (S. Topçu, Çev.). S. Karakaş ve R. Eski (Ed.), *Psikolojiye giriş* (22. baskı) içinde (s. 259-284). Eğitim Yayınevi.

Olkun, S ve Toluk-Uçar, Z. (2006). *İlköğretimde matematik öğretiminde çağdaş yaklaşımlar.* Ekinoks.

Özdamar, K. (2017). *Ölçek ve test geliştirme yapısal eşitlik modellemesi: IBM SPSS, IBM SPSS AMOS ve MINITAB uygulamalı.* Nisan Kitabevi.

Pallant, J. (2016). *SPSS kullanma kılavuzu: SPSS ile adım adım veri analizi* [SPSS survival manual: A step by step guide to data analysis using IBM SPSS](S. Balcı ve B. Ahi, Çev.). Anı Yayıncılık (2015).

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning, 1*(1), 81-106. https://doi.org/10.1007/BF00116251

Quinlan, J. R. (1993). *C4.5: Programs for machine learning.* Morgan Kaufman Publishers, Inc.

Rokach, R., & Maimon, O. (2010). Classification tress. In O. Maimon and L. Rokach (Ed.) *Data mining and knowledge discovery handbook* (pp. 148-174). Springer.

Schumacher, R., and Lomax, R. (2004). *A beginner's guide to structual equation modelling.* Lawrence Erlbaum Associates.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*(2), 173–180. https://doi.org/10.1207/s15327906mbr2502_4

Tabachnick, B. G., and Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Pearson Education. Inc.

Tapkan, P., Özbakır, L., & Baykasoğlu, A. (2011, Eylül-Ekim). *WEKA ile veri madenciliği süreci ve örnek uygulama.* Endüstri mühendisliği yazılımları ve uygulamaları kongresi, İzmir, Türkiye. tarihinden http://embk.mmoizmir.org/wp-content/uploads/2016/05/emyk18.pdf

Tavşancıl, E. (2014). *Tutumların ölçülmesi ve SPSS ile ileri veri analizi.* Nobel Akademik Yayıncılık.

The World Bank. (2021). *Early Childhood Development.* https://www.worldbank.org/en/topic/earlychildhooddevelopment

Trends in International Mathematics and Science Study. (2019). *TIMSS 2019 mathematics framework.* https://timss2019.org/wp-content/uploads/frameworks/T19-Assessment-Frameworks-Chapter-1.pdf

Uğur, A. ve Kınacı, A. C. (2006, 21-23 Aralık). *Yapay zeka teknikleri ve yapay sinir ağları kullanılarak web sayfalarının sınıflandırılması.* XI. "Türkiye'de İnternet" Konferansı, Ankara, Türkiye. http://inet-tr.org.tr/inetconf11/kitap/ugur_kinaci_inet06.pdf

Uyanık, Ö. ve Kandır, A. (2010). Okul öncesi dönemde erken akademik beceriler. *Kuramsal Eğitimbilim Dergisi, 3*(2), 118-134. https://dergipark.org.tr/tr/download/article-file/304144

Üstün, U., Özdemir, E., Cansız, M. ve Cansız, N. (2019). Türkiye'deki öğrencilerin fen okuryazarlığını etkileyen faktörler nelerdir? PISA 2015 verisine dayalı bir hiyerarşik doğrusal modelleme çalışması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 35*(3), 720-732. https://doi.org/10.16986/HUJE.2019050786

Wood, T. (2019). What is the F-score? https://deepai.org/machine-learning-glossary-and-terms/f-score

Yavuzsoy-Köse, N. (2016). Cumhuriyetten günümüze kadar ilkokul matematik dersi öğretim programlarında matematiksel muhakeme. M. F. Özmantar, A. Öztürk ve E. Bay (Eds.) *Reform ve değişim bağlamında ilkokul matematik öğretim programları* içinde (ss. 317-246). Pegem Akademi Yayıncılık.

Zhang, P. G. (2010). Neural networks for data mining. In O. Maimon & L. Rokach (Eds.) *Data mining and knowledge discovery handbook* (ss. 419-444). Springer.

PISA 2018 uygulamasında hızlı tahmin davranışının farklı değişkenlere göre incelenmesi

Zeynep Nur Arpaguş

Giriş

Uluslararası büyük ölçekli değerlendirmeler, eğitim paydaşlarına ulusal düzeyde öğrencilerin güçlü ve zayıf yönlerini belirlemelerine ve öğrenci performansı ile performansı etkileyebilecek faktörler arasındaki ilişkiyi incelemelerine olanak tanıyan bilgiler sağlama potansiyeline sahiptir (Rios ve Guo, 2020). Uluslararası Öğrenci Değerlendirme Programı (PISA), üçer yıllık dönemler hâlinde, 15 yaş grubundaki öğrencilerin kazanmış oldukları bilgi ve becerileri değerlendiren bir araştırmadır (MEB, 2019). PISA, ülkelerin eğitim reformlarına yol gösterici olması bakımından politika yapıcılar için birincil öneme sahip iken öğrenciler ise sonuçlarından herhangi beklentilerinin olmadığı bu sınavlarda yeterince çaba göstermeyebilirler. Bu durum bahsi geçen sınavların geçerliğini olumsuz etkilerken, bu öğrencilerin yanıtları kullanılarak Madde Tepki Kuramına dayalı olarak yapılan yetenek ve madde parametre kestirimlerinde de yanlılığa sebep olabilmektedir (Rios ve Guo, 2020).

Öğrencilerin sınavda gösterdikleri çabayı ölçmek için farklı yöntem ve teknikler önerilmiş, öğrencilerin sınavda ne kadar çaba gösterdiklerini belirlemek amacıyla anket soruları yöneltilmiş ancak anketlerden elde edilen sonuçlarda doğruluk ve geçerlik problemleri gözlenmiştir (Wise ve De Mars, 2005). Alternatif olarak Wise ve Kong (2005), bilgisayar ortamında uygulanan sınavlarda öğrencilerin maddelere reaksiyon süresine göre gösterdikleri çabayı belirlemek amacıyla yanıt süresi çabasının (response time effort) kullanılmasını önermiştir. Öğrencinin maddeyi yanıtlama süresinin belirlenen eşik değerden az olmasının öğrencinin yeterli çaba göstermemesine karşılık geldiği varsayılmakta ve bu durum hızlı tahmin davranışı olarak tanımlanmaktadır (Debeer, Buchholz, Hartig ve Janssen, 2014). Hızlı tahmin davranışı gösteren öğrenci, bir maddeyi okuyup anlamak için yeterli süre geçmeden hızlıca yanıt vermektedir (Schnipke, 1995; Wise, 2017). Yanıtlama sürelerindeki farklılıklar; test çözme çabası, motivasyon, maddelerin güçlük seviyesi, ve maddenin testteki yerine göre değişebilmektedir (Guo ve diğ., 2016; Rios ve diğ., 2017; Van der Linden, 2011; Bilge, 2017; Wise, 2006). Cinsiyetin de yanıt süresi çabası üzerinde etkisi olabildiği belirlenmiş, DeMars ve diğ., (2013) genellikle kız öğrencilerin daha fazla çaba sarf ettiği sonucuna ulaşmışlardır. Bu sonuçlar, hızlı yanıt verme davranışının uluslararası karşılaştırmalı araştırmalardan elde edilen puanlara dayalı yapılan çıkarımların geçerliliğine potansiyel bir tehdit olduğunu göstermektedir.

Ercikan ve diğ., (2020), Değişen madde fonksiyonu (DMF) analizlerinin grupların toplam yanıt sürelerine göre eşleştirilerek yanıt süreçlerindeki farklılaşmanın belirlenmesi amacıyla kullanılmasını önermiştir. Genel olarak DMF analizleri, aynı yetenek seviyesine sahip farklı grupların bir maddeyi doğru yanıtlama oranlarının farklılaşıp farklılaşmadığını belirlemek amacıyla yapılmakta ve alt gruplar toplam puanlarına göre eşleştirilmektedir (Dorans, 2013; Ercikan, 1998; Ercikan ve Lyons-Thomas, 2013; Holland ve Thayer, 1988; Holland ve Wainer, 1993; Zieky, 1993, 2011; Zwick, 2012; akt. Ercikan ve diğ., 2020). DMF analizlerinde alt gruplar toplam yanıt sürelerine göre eşleştirildiğinde değişen yanıt süreleri, testi aynı sürede tamamlayan gruplar için farklı süreye ihtiyaç duyulan maddeleri işaret edebilir (Ercikan ve diğ., 2020). Bu önerilerden yola çıkılarak aşağıdaki araştırma sorularına yanıt aranacaktır:

1. Hızlı tahmin davranışı ülkelere ve cinsiyete göre farklılık göstermekte midir?
2. Hızlı tahmin davranışı ile madde güçlüğü ve öğrenci başarısı arasında ilişki var mıdır?
3. Öğrenciler toplam yanıt sürelerine göre eşleştirildiğinde, ülkelere göre farklı süreye ihtiyaç duyulan maddeler var mıdır?

Yöntem

Çalışma, öğrencilerin ve grupların çeşitli özelliklerinin inceleneceği betimsel bir araştırmadır. PISA 2018 uygulamasına katılan 79 ülkeden 70'inde matematik okuryazarlığı testi bilgisayar tabanlı olarak uygulanmış olup veride öğrencilerin maddeleri yanıtlama süreleri de yer almaktadır. Çalışma grubunu Kanada (598), İngiltere (506), Çin (469) ve Türkiye (255) olmak üzere 4 ülke oluşturmaktadır. Ülkeler belirlenirken Form-1 testini alan öğrenci sayılarının 200'ün üzerinde olmasına dikkat edilmiştir. PISA 2018 uygulaması matematik okuryazarlığı alanında 70 ülke bilgisayar tabanlı değerlendirilmiş olup çalışma 4 ülke ile sınırlandırılmıştır. Matematik okuryazarlığı alanında 24 farklı form kullanılmıştır ancak maddelerin testteki yerinin maddeyi hızlı yanıtlama davranışında değişikliğe sebep olmaması için çalışmaya sadece Form-1 testini alan öğrencilerin yanıtları dahil edilmiştir.

Öncelikle, altı ve daha fazla soruya yanıt vermeyen öğrenciler kayıp veri olarak kabul edilip analizden çıkarıldıktan sonra ülkelere ve cinsiyete göre betimsel istatistikler sunulacaktır. Her bir maddeye ait süre ortalamalarıyla ülke bazında kümülatif çizgi grafikleri çizdirilecektir. Ardından ülkelerin toplam yanıt süreleri hesaplanacak ve hızlı tahmin davranışına sahip olan öğrenciler tespit edilecektir. Her madde için bir yanıt süresi eşiği olarak %10 oranı kullanılarak (Goldhammer ve diğ., 2016) eşik değerden daha kısa sürede maddeyi yanıtlayanlar hızlı tahmin davranışı, eşik süresinde veya daha uzun sürede maddeyi yanıtlayanlar ise maddeyi çözme davranışı gösterenler olarak kabul edilecektir. Hızlı tahmin davranışında bulunan öğrencilerin ülkelere ve cinsiyete göre oranları istatistiksel testler ile karşılaştırılacaktır. Madde güçlüğü ve öğrencinin başarı durumu için ise korelasyon değerleri hesaplanacaktır. Son olarak, maddeleri aynı sürede cevaplayan öğrencilerin karşılaştırılmasında standartlaştırılmış P-DIF (Dorans ve Kulick, 1986) istatistiğinden yararlanılacaktır. Çalışmada, betimsel istatistiklerin hesaplanmasında SPSS (vers. 26), DMF belirlenmesinde ise difR paketi kullanılacaktır.

Sonuçlar

Beklenen sonuçlar araştırma problemlerinin sırasıyla verilmiştir. "Hızlı tahmin davranışı eşik değeri ülkelere ve cinsiyete göre farklılık göstermekte midir?" araştırma sorusuna yönelik olarak belirlenen eşik değerinden daha kısa sürede yanıt veren öğrenci oranlarının ülkelerin başarı ortalamalarıyla ilişkili olarak farklılık göstermesi ve alanyazına göre hızlı tahmin davranışında bulunan erkek öğrenci sayısının kız öğrencilerden fazla olması beklenmektedir. "Hızlı tahmin davranışı ile madde güçlüğü ve öğrenci başarısı arasında ilişki var mıdır?" araştırma sorusuna yönelik olarak ise madde güçlüğü arttıkça öğrencilerin hızlı tahmin davranışında bulunma sıklığının artması öğrencinin başarısı ile hızlı yanıt davranışı arasında ise negatif yönlü ilişki olması beklenmektedir. "Öğrenciler toplam yanıt sürelerine göre eşleştirildiğinde, ülkelere göre farklı süreye ihtiyaç duyulan maddeler var mıdır?" araştırma sorusuna yönelik olarak testlerin farklı dillere çevrilerek uygulanması sebebiyle toplam yanıt süreleri aynı olsa da bazı maddelerin yanıtlanması için ihtiyaç duyulan sürelerin farklılaşabileceği öngörülmektedir.

Kaynaklar

Debeer, D., Buchholz, J., Hartig, J., and Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics*, *39*(6), 502-523. https://doi.org/10.3102/1076998614558485

Dorans, N., and Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement,* 23(4), 355–368. http://www.jstor.org/stable/1434554

Ercikan, K., Guo, H., and He, Q. (2020). Use of response process data to ınform group comparisons and fairness research. *Educational Assessment*, *25*(3), 179-197. https://doi.org/10.1080/10627197.2020.1804353

Goldhammer, F., Martens, T., Christoph, G., and Lüdtke O. (2016). Test-taking engagement in PIAAC. *OECD Education Working Papers,* No. 133. OECD Publishing. https://doi.org/10.1787/5jlzfl6fhxs2-en.

Guo, H., Rios, J., Haberman, S., Liu, O. L., Wang, J., and Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, *29*(3), 173–183. https://doi.org/10.1080/08957347.2016.1171766

MEB (2019). *PISA 2018 ulusal ön raporu* (Eğitim Analiz ve Değerlendirme Raporları Serisi, No. 10). Miili Eğitim Bakanlığı. https://www.meb.gov.tr/meb_iys_dosyalar/2019_12/03105347_PISA_2018_Turkiye_On_Raporu.pdf

Rios, J., and Guo H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of differential noneffortful responding on an ınternational college-level assessment of critical thinking. *Applied Measurement in Education*, *33*(4), 263-279. https://doi.org/10.1080/08957347.2020.1789141

Rios, J., Guo, H., Mao, L., and Liu, O. L. (2017). Evaluating the impact of careless responses on aggregated scores: To filter unmotivated examinees or not? *International Journal of Testing*, 17, 74–104. https://doi.org/10.1080/15305058.2016.1231193

Schnipke, D. L. (1996). Assessing speededness in computer-based tests using item response times (Publication No. 9617600) [Doctoral dissertation, Johns Hopkins University]. ProQuest Dissertations & Theses Global.

Van der Linden, W. J. (2011). Test design and speededness. *Journal of Educational Measurement*, *48*(1), 44–60. http://www.jstor.org/stable/23018064

Wise, S., and DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10, 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S., and Gao, L. (2017). A general approach to measuring testtaking effort on computer-based tests. *Applied Measurement in Education*, *30*(4), 343-354. https://doi.org/10.1080/08957347.2017.1353992

Wise, S. L., and Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, 18, 163–183. https://doi.org/10.1207/s15324818ame1802_2

LGS sekizinci sınıf testlerinin derslere ve konulara göre önem düzeylerinin sınıflama ve sıralama yargılarıyla ölçekleme ile analizi

Özge Öncü, Ayşenur Tavlıca ve Hakan Koğar

Öz

Bu çalışmada, sekizinci sınıfta öğrenim görmekte olan öğrencilerin LGS kapsamında cevapladıkları Türkçe, T.C. İnkılap Tarihi ve Atatürkçülük, Din Kültürü ve Ahlak Bilgisi, Yabancı Dil (İngilizce), Matematik ve Fen Bilimleri testlerini hangi sıralamaya göre cevap verdiklerini belirlemek; Türkçe ve Matematik testlerinde konular dâhilinde duydukları kaygı düzeyini ortaya koymak amaçlanmıştır. Çalışma grubunu Antalya ve Hatay illerindeki devlet okullarında 8.sınıfta öğrenim gören 220 öğrenci oluşturmaktadır. Öğrencilerin LGS kapsamındaki dersler bağlamında sıralama yargılarına dayalı ölçekleme ile izledikleri cevaplama sırası belirlenmiştir. Türkçe ve Matematik testlerinin içerdiği konulara duydukları kaygı düzeyleri ise sınıflama yargılarına dayalı olarak ölçeklenmiştir. Elde edilen bulguların güvenirliğini ortaya koymak amacıyla iç tutarlılık katsayıları hesaplanmıştır. Araştırmanın sonucunda, öğrencilerin LGS testlerini cevaplamak için izledikleri sıralamanın sözel bölüm için Türkçe, T.C. İnkılap Tarihi ve Atatürkçülük, Din Kültürü ve Ahlak Bilgisi ve Yabancı Dil; sayısal bölüm için ise Fen Bilimleri ve Matematik şeklinde ilerlediği görülmüştür. Matematik konuları içerisinde en çok kaygı uyandıran konu *dönüşüm geometrisi* ve *geometrik cisimler* iken, en az kaygı uyandıran konu *çarpanlar* ve *katlar* olarak belirlenmiştir. Türkçe konuları içerisinde en çok kaygı uyandıran konu *yazım bilgisi* iken, en az kaygı uyandıran konunun *sözcükte anlam* olduğu görülmüştür.

Kaynaklar

Anıl, D. ve İnal, H. (2017). *Psikofizikte ölçekleme uygulamaları.* Pegem Akademi

Arık, R. S. ve Kutlu Ö. (2013). Öğretmenlerin ölçme ve değerlendirme alanı yeterliklerinin yargıcı kararlarına dayalı ölçeklenmesi. *Eğitim Bilimleri Araştırma Dergisi, 3*(2), 163-192. https://dergipark.org.tr/en/download/article-file/697961

Aykaç, N. ve Atar, E. (2014, 16-18 Ocak). Geçmişten günümüze ilköğretimden ortaöğretime geçiş sisteminin değerlendirilmesi. A. Akdoğanbulut-İnsan, ve A. Yavuz-Akengin (Eds.), *Cumhuriyet'in kuruluşundan günümüze eğitimde kademeler arası geçiş ve yeni modeller uluslararası kongresi* içinde (ss. 83–104). Atatürk Araştırma Merkezi. https://www.atam.gov.tr/wp-content/uploads/CUMHUR%C4%B0YET%E2%80%99%C4%B0N-KURULU%C5%9EUNDAN-G%C3%9CN%C3%9CM%C3%9CZE-

E%C4%9E%C4%B0T%C4%B0MDE-KADEMELER-ARASI-GE%C3%87%C4%B0%C5%9E1.pdf

Bal, Ş., Koç, E. ve Yıldırım, H. İ. (2008). İlköğretim ikinci kademe fen bilgisi müfredatı ile liselere giriş sınavları fen bilgisi sorularının öğrencilerin kişisel bilgileri de dikkate alınarak karşılaştırılması. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi, 9*(3), 35-48.

Baykul, Y. ve Turgut, M. F. (1992). *Ölçekleme teknikleri* (2. baskı). ÖSYM Yayınları.

Bilen, M. (2006). *Plandan uygulamaya öğretim* (7. baskı). Anı Yayıncılık.

Bökeoğlu, Ö. Ç. ve Özdem, G. (2009). Danimarka lisansüstü öğretim sistemi. E. Doğan-Kılıç (Ed.), *Lisansüstü öğretim sistemleri* içinde (ss.107-132). Pegem Akademi.

Can, E. (2015). Qualitative obstacles in Turkish education system and suggestions. *The Anthropologist, 20*(1-2), 289–296.

Can, E. (2017). Öğrenci görüşlerine göre merkezi sınavların etkilerinin belirlenmesi. *Akademik Sosyal Araştırmalar Dergisi, 5*(58), 108–122.

Çelik, Z. (2011, 10-11 Aralık). *Ortaöğretime geçiş sınav sistemleri ve politikaları. 21. yüzyılda Türkiye'nin eğitim ve bilim politikaları.* 21. Yüzyılda Türkiye'nin Eğitim ve Bilim Politikaları Sempozyumu'nda sunulmuş sözlü bildiri.

Çelik, Z., Boz, N., Arkan, Z. ve Toklucu, D. K. (2017). TEOG yerleştirme sistemi: güçlükler ve öneriler. SETA (Siyaset, Ekonomi ve Toplum Araştırmaları Vakfı), 94(1). https://www.researchgate.net/publication/320628250 adresinden alınmıştır.

Çepni, S., Özsevgeç, T. ve Gökdere, M. (2003). Bilişsel gelişim ve formal operasyon dönem özelliklerine göre ÖSS fizik ve lise fizik sorularının incelenmesi. *Milli Eğitim Dergisi,157*, 30-39.

Demir, S. B. ve Yılmaz T. A. (2019). En iyisi bu mu? Türkiye'de yeni ortaöğretime geçiş politikasının velilerin görüşlerine göre değerlendirilmesi. *Bolu Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 19*(1), 164-183. https://dx.doi.org/10.17240/aibuefd.2019.19.43815-445515

Edwards, A. L. (1957). *Techniques of attitude scale contruction.* Appleton-Century-Crofts.

Eurypedia. (2013). The structure of the European education systems 2013/14:schematicdiagrams. http://www.indire.it/lucabas/lkmw_file/eurydice/education_structures_2013_EN.pdf

Ergün, M. (2014, 16-18 Ocak). *Eğitimde kademelerin oluşması ve kademeler arası geçiş düzenlemelerine tarihi bakış.* Cumhuriyet'in kuruluşundan günümüze eğitimde kademeler arası geçiş ve yeni modeller uluslararası kongresi (ss. 1-40). Atatürk Kültür, Dil ve Tarih Yüksek Kurumu Atatürk Araştırma Merkezi. https://www.atam.gov.tr/wp-content/uploads/CUMHUR%C4%B0YET%E2%80%99%C4%B0N-KURULU%C5%9EUNDAN-G%C3%9CN%C3%9CM%C3%9CZE-E%C4%9E%C4%B0T%C4%B0MDE-KADEMELER-ARASI-GE%C3%87%C4%B0%C5%9E1.pdf

Gelbal, S. ve Kara, Y. (2013). İlköğretim öğrencilerinin başarılarını etkileyen özelliklerin tam sıralama halinde ikili karşılaştırmalar yöntemiyle ölçeklenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 4*(1), 33-51. https://dergipark.org.tr/tr/pub/epod/issue/5801/77217

Gelbal, S. ve Sayın, A. (2014). Başarıyı etkileyen faktörlerin önem derecelerinin ardışık aralıklar yöntemiyle ölçeklenmesi. *Amasya Üniversitesi Eğitim Fakültesi Dergisi 3*(1), 1-26.

Görmez, M. ve Coşkun, İ. (2015). 1. Yılında temel eğitimden ortaöğretime geçiş reformunun değerlendirilmesi. *Siyaset, Ekonomi ve Toplum Araştırmaları Vakfı (SETA),*114, 1-21. http://file.setav.org/Files/Pdf/20150115172839_1.-yilinda-temel-egitimdeortaogretime-gecisreformunun-degerlendirilmesi-pdf.pdf

Gür B., S., Çelik Z. Ve Coşkun İ. (2013). Türkiye'de ortaöğretimin geleceği: Hiyerarşi mi, eşitlik mi? *Siyaset, Ekonomi ve Toplum Araştırmaları Vakfı (SETA),* 69, 1-28 http://file.setav.org/Files/Pdf/20130802120003.http://file. setav.org/Files/Pdf/20130802120003_ortaogretim_analiz2.pdf adresinden alındı.

Jain, C., and Prasad, N. (2017). Indian education system: structure and key challenges. Quality of Secondary Education in India, 67–78. https://doi.org/10.1007/978-981-10-4929-3_7

Linn, R. L. A. and Gronlund, N. E. (1995). *Measurement and assessment in teaching* (7th ed.). PrenticeHall.

Milli Eğitim Bakanlığı (MEB) (2018). *Sınavla öğrenci alacak ortaöğretim kurumlarına ilişkin merkezi sınav başvuru ve uygulama kılavuzu.* Milli Eğitim Bakanlığı. https://www.meb.gov.tr/meb_iys_dosyalar/2020_05/06105923_BasYvuru_ve_Uygulama_KYlavuzu_2020_GuYncel.pdf

Tezbaşaran A. (2004). Likert tipi ölçeklere madde seçmede geleneksel madde analizi tekniklerinin karşılaştırılması. *Türk Psikoloji Dergisi, 19*(54), 77–90.

Torgerson, W. S. (1958). *Theory and methods of scaling.* John Wiley & Sons Inc.

Ulusoy, B. (2020) 8. Sınıf öğrencilerinin liselere geçiş sınavı (LGS)'na ilişkin algılarının metaforlar aracılığıyla incelenmesi. *Necmettin Erbakan Eğitim Fakültesi Dergisi 2*(2), 186-202. https://doi.org/10.51119/ereegf.2020.5

Wu Y. (2015) The examination system in China: The case of zhongkao mathematics. In Cho S. (Ed.) *Selected regular lectures from the 12th ınternational congress on mathematical education* (pp. 897-914). Springer. https://doi.org/10.1007/978-3-319-17187-6_50

Ölçme ve değerlendirme dersi başarısının yordanmasına ilişkin bir çalışma

Şeyma Uyar ve Neşe Öztürk Gübeş

Giriş

Öğretmenler, bir toplumun geleceği olarak nitelendirilen bireylerin yetiştirilmesi noktasında görevlendirilmiş eğitim sürecinin önemli üyeleridir. Böylesine önemli bir görevin sorumluluğu öğretmenlerin daha yeterli ve nitelikli olmasını gerektirmektedir (MEB, 1999). Bu doğrultuda ulusal ve uluslararası uzmanlar, akademisyenler, öğretmen görüşleri ile çalıştay ve pilot uygulamalar neticesinde öğretmen yeterlikleri belirlenmiş olup eğitim alanındaki gelişmeler ve eğitim sitemindeki yeniliklere uygun olarak "Öğretmenlik Mesleği Genel Yeterlikleri" tanımlanmıştır. Öğretmenlik Mesleği Genel Yeterliklerinin, bir öğretmende var olması gereken özellikleri somut bir biçimde ortaya koyması ve bu alanda geliştirilecek politikalara referans olması beklenmektedir. Öte yandan öğretmenlerin güçlü ve geliştirmesi gereken yönlerini objektif biçimde görmelerine katkı sağlaması öngörülmektedir. Bu özellikler mesleki bilgi, mesleki beceri, tutum ve değerler olmak üzere üç yeterlik alanında ele alınmıştır. Mesleki beceri yeterlik alanında dört alt yeterlik belirlenmiş olup bu alt yeterliklerinden biri de ölçme ve değerlendirmedir. Bu alanda öğretmenlerden ölçme ve değerlendirme yöntem, teknik ve araçlarını amacına uygun kullanmaları beklenmektedir (MEB, 2017).

Ölçme ve değerlendirme becerisi, öğretmen adaylarına lisans eğitiminde ölçme ve değerlendirme dersi kapsamında kazandırılmaya çalışılmaktadır. Öğretmen adayları, bir diğer deyişle geleceğin öğretmenleri ölçme ve değerlendirme dersinde kazandıkları bilgi ve beceriler doğrultusunda eğitim-öğretim faaliyetlerinin kalite kontrolünü yaparlar. Bu nedenle ileride uygulanacak olan değerlendirme faaliyetlerinin niteliği öğretmen adaylarının bu dersteki başarısından bağımsız düşünülememektedir. Öğrenci ders başarısı zihinsel, çevresel ve duyuşsal pek çok faktörden etkilenebilmektedir (Sarıer, 2016). Alanyazında özyeterliğin öğrenci başarısının önemli bir yordayıcısı olduğu belirtilmektedir (Bandura ve Wood, 1989; Maddux, 2002; Öztürk ve Kurtuluş, 2017; Pintrich ve De Groot 1990; Zimmerman, 2000). Bazı çalışmalarda özyeterliğin ve üst biliş öğrenme becerilerinin (yürütücü biliş, metacognition, bilişötesi) birbiriyle ilişkili olduğu ve öğrenci performansını yordadığı vurgulanmaktadır. Öğrenme süreçleri üzerinde denetimi sağlayan üstbiliş becerilerinin yüksek olması öğrenci performansında da artış sağlayabilmektedir (Kanfer ve Ackkerman, 1989; Bouffard-Bouchard ve diğ., 1991, Desoete ve diğ., 2001). Hatta üst biliş ile performans arasındaki ilişkiye öz yeterlik inancının aracılık ettiği de

belirtilmektedir (Bandura ve Wood, 1989; Coutinho, 2007, 2008). Öte yandan yapılan çalışmalarda öğrencinin cinsiyeti, anne ve baba eğitim durumu (Kart ve Gülleroğlu, 2012); derse yönelik tutumu, genel akademik başarısı ve Anadolu öğretmen lisesinden mezun olma durumlarının (Kurşun ve Çobanoğlu-Aktan, 2016) ölçme değerlendirme dersindeki başarıyı istatistiksel olarak anlamlı bir şekilde yordadığı ifade edilmektedir.

Yapılan çalışmalar doğrultusunda eğitimde ölçme ve değerlendirme dersi başarısını etkileyen faktörlerin belirlenip buna göre önlemler alınması önemli görülmektedir. Literatür incelendiğinde ölçme ve değerlendirme dersi başarısını etkileyen faktörlerin araştırıldığı sınırlı sayıda çalışma olduğu ifade edilebilir. Bu nedenle araştırmanın amacı, öğretmen adaylarının eğitimde ölçme ve değerlendirme dersi başarısı ile genel akademik ortalama, cinsiyet, ölçme ve değerlendirme genel yeterlik algısı ve yürütücü biliş becerileri arasındaki ilişkileri belirlemek olarak ele alınmıştır. Bu amaç doğrultusunda ele alınan problem şu şekilde ifade edilmektedir:

Öğretmen adaylarının genel akademik ortalama, cinsiyet, ölçme ve değerlendirme genel yeterlik algısı ve yürütücü biliş becerileri ölçme ve değerlendirme dersindeki başarısını anlamlı bir şekilde yordamakta mıdır?

Yöntem

Araştırmanın çalışma grubunu, 2019-2020 bahar döneminde ölçme ve değerlendirme dersini almış olan 158 öğrenci oluşturmaktadır. Öğrencilerin %16.8 (n=26)'ini fen bilgisi, %18.1 (n=28)'ini İngilizce, %14.8 (n=23)'ini okul öncesi, %30.3 (n=47)'ünü sosyal bilgiler ve %20.00 (n=31)'sini Türkçe eğitiminde öğrenim görmektedir. Araştırma kapsamında Nartgün (2008) tarafından geliştirilen "Öğretmen Adayları İçin Ölçme ve Değerlendirme Genel Yeterlik Algısı Ölçeği ile Altındağ ve Senemoğlu (2013) tarafından geliştirilen "Yürütücü Biliş Becerileri Ölçeği" çevrim içi olarak uygulanmıştır. Öğretmen Adayları İçin Ölçme ve Değerlendirme Genel Yeterlik Algısı Ölçeği; "temel kavramlar", "ölçme teknikleri" ve "istatistiksel çözümleme ve raporlama" olmak üzere üç faktörlü bir yapıya sahiptir. Beşli likert tipinde puanlanan ölçek, toplam 24 maddeden oluşmaktadır. Ölçeğin alt boyutları için Cronbach alfa güvenirlik katsayıları sırayla .79, .86 ve .86 bulunmuştur. Yürütücü Biliş Becerileri Ölçeği, tek boyutludur ve Likert tipi puanlanan 30 maddeden oluşmaktadır. Ölçeğe ait Cronbach alfa güvenirlik katsayısı .91 olarak bulunmuştur.

Araştırma kapsamında, ölçme ve değerlendirme başarısını belirlemek amacıyla öğrencilerden derse ilişkin seçilen 24 temel kavramı örneklendirdikleri bir hikâye yazmaları istenmiştir. Hikâyeler, temel kavram doğru örneklendirildiyse "1", yanlış ya da örnek verilmediyse "0" verilerek puanlanmış, elde edilen puanlar 100'lük sisteme çevrilmiştir.

Veri analizinin ilk aşamasında veri setinin regresyon analizi için uygunluğu incelenmiştir. Bunun için öncelikle ölçek toplam puanlarının z değerleri hesaplanarak tek değişkenli uç değerler incelenmiş ve z değeri +/-3 aralığının dışında kalan herhangi bir uç değer bulunmamıştır. Çok değişkenli uç değerleri belirlemek için mahalonobis uzaklıkları hesaplanmıştır, .001 anlamlılık düzeyinde ki kare tablo değeri

22.457'den büyük olan üç veri çıkartılmış ve analizlere 155 öğrenciye ait veri ile devam edilmiştir. Tek değişkenli normallik varsayımı için çarpıklık ve basıklık katsayıları incelenmiş ve +/- 1 aralığında oldukları görülmüştür. Tolerans değerlerinin minimum 0.52 olup 0.20'den düşük olmaması, VIF değerlerinin maksimum 1.91 olup 10'dan küçük olmasına dayalı olarak veri setinde çoklu bağlantı problemi olmadığı tespit edilmiştir. Veri analizinin ikinci aşamasında hiyerarşik çoklu doğrusal regresyon analizi yapılmıştır.

Sonuçlar

Modele ilk olarak geçmiş araştırmalarda ölçme ve değerlendirme başarısının istatistiksel olarak anlamlı yordayıcısı olduğu belirlenen genel akademik başarı ve cinsiyet (kukla kadın) değişkenleri alınmış daha sonra temel kavramlar, ölçme teknikleri ve istatistiksel çözümleme ve raporlama puanları ile yürütücü biliş becerileri puanı dahil edilmiştir.

Hiyerarşik çoklu doğrusal regresyon analizinin ilk bloğunda, modele alınan genel akademik ortalama ve cinsiyet değişkenlerinin öğrencilerin ölçme ve değerlendirme başarılarını istatistiksel olarak anlamlı bir şekilde yordadığı görülmüştür [$F_{(2,155)}$= 9.16, p< .05, R^2=0.11, R_{adj}= 0.09]. Regresyon modelinde yer alan tüm yordayıcılar, ölçme ve değerlendirme başarı puanındaki değişkenliğin %11'ini açıklamaktadır. Birinci blok için kurulan regresyon modelindeki genel akademik ortalama istatistiksel olarak anlamlı (β= 0.27, t= 3.48 p< .05) bir yordayıcı iken cinsiyet değişkeni (β= 0.12, t= 1.50 p> .05) anlamlı bir yordayıcı değildir.

İkinci blokta modele alınan temel kavramlar, ölçme teknikleri, istatistiksel çözümleme ve raporlama GYA ile üst biliş beceri değişkenlerinin öğrencilerin ölçme ve değerlendirme başarısını istatistiksel olarak anlamlı bir şekilde yordadığı [$F_{(6,151)}$= 5.34, p< .05, R^2=0.18, R_{adj}= 0.14] ve toplam varyansın %18'ini açıkladığı görülmüştür. İkinci blokta modele dahil edilen değişkenlerin açıklanan varyansa katkısı %7 olmuştur. Model 2'de genel akademik ortalama (β=0.26, t= 3.26 p< .05) ve temel kavramlar GYA (β= 0.29, t= 3.09 p< .05) istatistiksel olarak anlamlı yordayıcılardır.

Kaynaklar

Altındağ, M. ve Senemoğlu, N. (2013). Metacognitive skills scale. *Hacettepe University Journal of Education*, *28*(1), 15-26.

Bandura, A., & Wood, R. (1989). Effect of perceived controllability and performance standards on self-regulation of complex deci-sion making. *Journal of Personality and Social Psychology, 56*(5), 805-814.

Bouffard-Bouchard, T., Parent, S., & Larivèe, S. (1991). Influence of self-efficacy on self-regulation and performance among junior and senior high-school aged students. *International Journal of Behavioral Development, 14,* 153-164.

Coutinho, S. (2007). The relationship between goals, metacognition, and academic success. *Educate, 7*(1), 39-47.

Coutinho, S. (2008). Self-efficacy, metacognition, and performance. *North American Journal of Psycholog, 10*(1), 165.

Desoete, A. H. Roeyers., A., & Buysse. E. (2001). Metacognition and mathematical problem solving in grade 3. *Journal of Learning Disabilities*, *34*(5), 435-447.

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abili-ties: An integrative/aptitude-treatment interaction approach to skill acquisition [Monograph]. *Journal of Applied Psychology, 74,* 657-690.

Kart, A. ve Gülleroğlu, H. D. (2013). Demografik ve duyuşsal değişkenlerin ölçme ve değerlendirme başarısını yordama gücü. *YYÜ Eğitim Fakültesi Dergisi, 10*(1), 11-30.

Kurşun, K. ve Çobanoğlu Aktan, D. (2016). Eğitimde ölçme ve değerlendirme dersinde başarıyı etkileyen faktörlerin çoklu göstergeler çoklu nedenler modeliyle incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 7*(2), 372-387.

Maddux, J. E. (2002). Self-efficacy: The power of believing you can. In C. R. Snyder and S. Lopez (Eds.), *Handbook of positive psychology* (pp. 335-343). Oxford University Press.

MEB EARGED (1999). *Çağdaş öğretmen profili.* Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı.

MEB (2017). *Öğretmenlik mesleği genel yeterlikleri.* Öğretmen Yetiştirme ve Geliştirme Genel Müdürlüğü. https://oygm.meb.gov.tr/meb_iys_dosyalar/2017_12/11115355_YYRETMENLYK_MESLEYY_GENEL_YETERLYKLERY.pdf

Nartgün, Z. (2008). Öğretmen adayları için ölçme ve değerlendirme genel yeterlik algısı ölçeği: Geçerlik ve güvenirlik çalışması. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 8*(2), 85-94.

Öztürk, B., & Kurtuluş, A. (2017). Ortaokul öğrencilerinin üstbilişsel farkındalık düzeyi ile matematik öz yeterlik algısının matematik başarısına etkisi. *Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi*, *31*, 762-778.

Pintrich, P. R., and De Groot, E. (1990). Motivational and self regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*(1), 33–40.

Sarıer, Y. (2016). Türkiye'de öğrencilerin akademik başarısını etkileyen faktörler: Bir meta-analiz çalışması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 31*(3), 609- 627.

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, *25*, 82-91.

Lisansüstü öğrencilerin bilişsel madde yazmaya ilişkin özyeterlik düzeyleri ve madde yazmada dikkat ettikleri noktalar: TÜBİTAK 2237-A etkinliği

Şeyma Uyar ve Nuri Doğan

Giriş

Öğrenciler ilkokuldan yükseköğretime kadar çeşitli ölçme ve değerlendirme faaliyetlerine tabi tutulurlar. Öğrenme ve öğretme stratejilerinin yapılandırmacı anlayışa uygun olarak öğretmen merkezli bir yapıdan öğrenci merkezli bir yapıya doğru kaymasıyla tamamlayıcı ölçme ve değerlendirme yöntemleri de bu faaliyetlere dahil olmuştur (MEB, 2005). Yine de ölçme ve değerlendirme faaliyetleri arasında en büyük ağırlığın sınavlarda olduğu söylenebilir. Ölçme ve değerlendirme sisteminin kurulu olduğu düzen daha çok sorular ve bu sorular için alınan yanıtların dönüt olarak kullanılması şeklinde düşünülebilir (Duran ve Tufan, 2017). Bu nedenle ölçme ve değerlendirme amacıyla geleneksel yöntemler olan çoktan seçmeli ve açık uçlu soruların daha sık kullanıldığı belirtilmektedir (Duran ve Tufan, 2017; Menteş, 2012). Yapılan çalışmalarda öğretmenlerin öğrenciyi tanıma ve başarılarını belirlemek için geleneksel ölçme araçlarının kullanımında kendilerini daha yeterli gördükleri ifade edilmektedir (Çelikkaya ve diğ., 2011; Gelbal ve Kelecioğlu, 2007). Öğretmenler kendilerini yeterli gördükleri ölçme araçlarını daha sık kullanma eğilimindedir (Gelbal ve Kelecioğlu, 2007). Öte yandan ülkemizde uygulanan ulusal sınavlara bakıldığında YKS (TYT, AYT ve YDT) ile liseye geçiş amacıyla uygulanan LGS'nin de çoktan seçmeli testlerden oluştuğu görülmektedir. Ülkemizin dahil olduğu uluslararası uygulamalar olan PISA ve TIMSS'de ise çoktan seçmeli ve açık uçlu maddelerin yer aldığı görülmektedir. Sınıf içi uygulamalarda çoktan seçmeli ve açık uçlu maddelerden oluşan testlerin kullanılmasının nedenlerinden biri de öğrencileri bu sınavlara hazırlamak olarak düşünülebilir.

Öğrencilerin gelişimlerinin izlendiği ve dönüt verildiği her türlü sınavda alınan kararların isabetli olmasının kullanılan ölçme aracının psikometrik özelliklerine bağlı olduğu düşünülmektedir. Ancak büyük özenle hazırlanmış geniş ölçekli sınavlarda bile bir sorunun ya da doğru yanıtın iptali şeklinde durumlarla karşılaşılmaktadır. Bu durum ölçme sistemine ilişkin güveni sarsabilmekte ve sonuçların doğruluğu hakkında kuşku yaratabilmektedir (Yüksel, 2019). Ölçme araçlarının geliştirilme aşamasında izlenecek bazı adımlarla bu konuda önlem almak mümkündür (Baykul, 2015). Öte yandan madde yazarının da bu adımlardan bağımsız olmadığı düşünülmektedir. Her bir gelişim sürecine uygun farklı ölçme tekniklerini tanıma, etkinleştirme ve geliştirmeye yönelik bilgi sahibi olmak ölçme ve

değerlendirmede etkililiği arttıran öğeler olarak düşünülebilir. İyi bir ölçme ve değerlendirme hazırlanan testte yer alan maddelerin yeterli olmasını gerektirir. Gerekli ölçütlere göre yazılamayan maddeler cevaplayıcıların aklını karıştırabilir, herhangi bir geribildirim için sistematik hataya sebep olabilir. Araştırmalar kusurlu maddelerden oluşan testlerden elde edilen sonuçların güvenirlik ve geçerliğinin daha düşük düzeyde olduğunu göstermektedir (Downing, 2005; Pais ve diğ., 2016; Tarrant ve diğ., 2006; Tarrant ve Ware, 2008). İyi bir madde yazmak madde yazarının konu alanına hâkim olmasını, akılcı ve iyi gelişmiş eğitimsel değerlere sahip olmasını, testin uygulanacağı gurubu tanımasını, sözle iletişimde uzman olmasını ve madde yazma teknikleri konusunda uzman olmasını gerektirmektedir (Ebel, 1951).

Nitelikli insan gücünü kazandırma sorumluluğu olan öğretmenlerin çağın gerektirdiği koşullara ayak uydurabilmek amacıyla kısa veya uzun süreli eğitim görmeleri gereklidir (Üstüner ve diğ., 2009). Yapılan çalışmalarda ilkokul öğretmenlerinin ölçme ve değerlendirme konusunda hizmet içi eğitime ihtiyaç duydukları (Şahin ve Ersoy, 2010); fen bilgisi/fizik öğretmenlerinin kavram geliştirme ve haritalama, fiziksel ölçme ve hata hesapları, işbirliği ile öğrenme süreci ve etkinlikleri hazırlama, çalışma yaprakları tasarlama ve geliştirme, etkin öğretim için etkinlik tasarlama ve geliştirme vb. noktalarında eksikliklerini giderilmesi gerektiği (Üstüner ve diğ., 2009), ilköğretim okullarında görev yapan öğretmenlerin ölçme ve değerlendirme alanında, özel öğretim yöntem ve teknikleri, eğitim teknolojisi ile eğitim programı alanlarında hizmet içi eğitime ihtiyaç duydukları belirtilmiştir (Tanyel, 1999).

Bu bağlamda araştırmanın amacı lisansüstü öğrencilerin madde yazma konusundaki eğitim öncesi ve sonrasındaki özyeterliklerini incelemek ve madde yazma noktasında dikkate aldıkları noktaları belirlemektir. Açıklamalar çerçevesinde araştırmada ele alınacak problem aşağıdaki gibi ifade edilebilir: Tübitak 2237-A Bilimsel Eğitim Etkinlikleri Desteği Programı kapsamında eğitim bilimleri ve öğretmen yetiştirme alanında lisansüstü öğrencilere yönelik "Nitelikli Soru Nasıl Yazılır?" eğitimi öncesinde ve sonrasında katılımcıların madde yazma ilkeleri kapsamındaki özyeterlilikleri ne düzeydedir? Madde yazma noktasında en çok dikkat ettikleri noktalar nelerdir? olarak belirlenmiştir.

Yöntem

Bu araştırmada nicel ve nitel yöntemlerin kullanıldığı karma araştırma yöntemi benimsenmiştir. Nicel ve nitel yöntemlerin birlikte kullanılması, araştırma problemine ilişkin çözümlemelerin ve yorumlamaların daha kapsamlı yapılmasına olanak sağlamaktadır (Creswell ve Plano-Clark, 2007). Eğitim öncesi ve sonrasına ilişkin özyeterlik düzeyleri çalışmanın nicel boyutu olarak hedeflenirken, öğrencilerin madde yazmada dikkat ettikleri noktalara ilişkin görüşleri ise nitel olarak toplanıp çözümlenmiştir. Nicel araştırma kapsamında tarama modeli kullanılmış, nitel araştırma kapsamında durum çalışması yöntemi kullanılmıştır.

Araştırmanın çalışma grubunu 3-6 Mayıs 2021 tarihlerinde TÜBİTAK 2237-A Bilimsel Eğitim Etkinliklerini Destekleme Programı kapsamında desteklenen "Nitelikli Soru Nasıl Yazılır? Bilişsel gelişim sürecine göre soru yazma, Sınıf İçi Ölçme ve Değerlendirme Araçlarında Geçerlilik, Güvenirlik, Madde Yanlılığı ve Bilişsel Tanı Modellerine Dayalı Değerlendirme" isimli eğitim etkinliğine seçilen 30 katılımcı

oluşturmaktadır. Katılımcılar Eğitim Bilimleri ve Öğretmen Yetiştirme alanında lisansüstü eğitim yapan kişilerden oluşmaktadır. Katılımcılar Matematik ve Fen Bilimleri Eğitimi, Sınıf Eğitimi, Türkçe Eğitimi ve Sosyal Bilgiler Eğitimi alanlarında lisansüstü eğitime kayıtlı öğrencilerdir. Bunlardan 8 (%27) tanesi aynı zamanda öğretmen olarak görev yapmaktadır.

Bu çalışmada katılımcıların eğitim öncesinde ve sonrasında madde yazma konusundaki özyeterliklerini incelemek amacıyla literatürde en çok ele alınan madde yazma ilkeleri (Baykul, 2015; Frey ve diğ., 2005; Haladayna, 2002) çerçevesinde araştırmacılar tarafından hazırlanan yapılandırılmış "madde yazma özyeterlik anketi" isimli anket formu çevrimiçi uygulanmıştır. Bu form 5'li Likert tipinde hazırlanmış (5: Çok yeterliyim, 4: Yeterliyim, 3: Orta düzeyde yeterliyim, 2: Yetersizim, 1: Çok yetersizim) 18 maddeden oluşmaktadır. Katılımcılardan eğitim öncesinde çoktan seçmeli ve açık uçlu madde yazarken hangi noktalara dikkat ettikleri açık uçlu olarak toplanmıştır.

Çalışmada lisansüstü öğrencilerin özyeterlik anketine verdikleri yanıtlar için yüzde ve frekans değerleri hesaplanmıştır. Eğitim öncesinde ve sonrasında madde yazma özyeterlik anket maddelerine verilen yanıtlar arasındaki farklar ki-kare testi ile incelenmiştir. Öte yandan madde yazma konusunda dikkate aldıkları noktalara ilişkin alınan yanıtlara içerik analizi yapılmıştır.

Sonuçlar

Bu araştırmadan elde edilen bulgular doğrultusunda lisansüstü öğrencilerin eğitim öncesinde madde yazma özyeterliklerine verdikleri yanıtlarda orta düzeyde yeterliyim kategorisinde yığılma gösterdiği görülmüştür. Eğitim sonrasında ise bu yeterlik düzeylerinin arttığı gözlenmiştir. Öğrencilerin %50'den fazlası eğitim öncesinde; bilişsel özelliğe uygun madde yazmada, madde köküne uygun çeldirici yazmada, eşleştirme tipi maddelerde öğrenci düzeyine uygun öncül sayısı ve doğru yanıt sayısı oluşturmada, madde yazmadan önce plan yapmada, maddeyi güvenirliğe katkı sağlayacak şekilde yazmada, geçerliğe katkı sağlayacak madde yazmada, seçeneklerin uzunluklarını birbirine yakın hazırlamada, açık uçlu maddeyi sınırlayarak yazmada, bağlam kullandığım maddelerde bağlamı günlük/gerçek hayatla ilişkili kurmada, bağlam kullandığım maddelerde bağlamın sorunun çözümüne katkı sağlaması konusunda, üst düzey becerilere uygun madde yazmada orta düzeyde yeterli olduğunu belirtmiştir. Doğru-yanlış tipi maddeleri basit yapıda yazmada ve maddeyi anlaşılır yazmada %50 ve üzeri oranında öğrenci kendilerini yeterli olarak ifade etmektedir. Madde yazmaya ilişkin hazırlanan maddelerde kendisini çok yetersiz olarak nitelendiren öğrenci bulunmamaktadır, çok yeterli olarak tanımlayan öğrenci sayısı ise 3 ve daha azdır.

Öğrencilerden nitel olarak toplanan bilgiler doğrultusunda çoğunun eğitim öncesinde daha çok açık ve anlaşılır madde yazmaya, madde kökünde gereksiz bilgi vermemeye, sınıf seviyesine uygun olmasına, madde yazmadan önce amaç ve kapsam belirleme noktalarına dikkat ettikleri görülmüştür.

Kaynaklar

Baykul, Y. (2015). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması* (3. baskı). Pegem Akademi.

Creswell, J. W., and Plano-Clark, V. L. (2007). *Designing and conducting mixed methods research*. Sage Publications.

Çelikkaya, T., Karakuş, U. ve Demirbaş, Ç. (2010). Sosyal bilgiler öğretmenlerinin ölçme-değerlendirme araçlarını kullanma düzeyleri ve karşılaştıkları sorunlar. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi, 11*(1), 57-76.

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education, 10*(2), 133-143. https://doi.org/10.1007/s10459-004-4019-5

Duran, E. ve Sezgin-Tufan, B. (2017). The effect of open-ended questions and multiple choice questions on comprehension. *International Journal of Languages Education, 5*(1), 242-254. https://doi.org/10.18298/ijlet.1676

Ebel, R. L. (1951). Writing the test item. In E. F. Linguist (Ed.), *Educational measurement* (pp. 185-245). American Council on Education.

Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education, 21*(4), 357-364. https://doi.org/10.1016/j.tate.2005.01.008

Gelbal, S. & Kelecioğlu, H. (2007). Öğretmenlerin ölçme ve değerlendirme yöntemleri hakkındaki yeterlik algıları ve karşılaştıkları sorunlar. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 33*(33) , 135-145. https://dergipark.org.tr/tr/pub/hunefd/issue/7805/102347

Haladyna, T. M., Downing, S. M., & Rodriguez M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement Education 15*(3), 309–334. https://doi.org/10.1207/S15324818AME1503_5

MEB (2005). *İlköğretim okulu ders programları ve öğretim kılavuzları*. Yakutiye Yayıncılık

Menteş, R. (2012). *Öğretmenlerin ölçme değerlendirme uygulamalarının belirlenmesi ve hizmet içi eğitim ihtiyaçlarının saptanması* (Yayımlanmamış Yüksek Lisans Tezi). Yakındoğu Üniversitesi, Lefkoşa.

Pais, J., Silva, A., Guimarães, B., Povo, A., Coelho, E., Silva-Pereira, F., and Severo, M. (2016). Do item-writing flaws reduce examinations psychometric quality? *BMC Research Notes, 9*(1), 399. https://doi.org/10.1186/s13104-016-2202-4

Şahin, Ç. ve Ersoy, E. (2010). Sınıf öğretmenlerinin ilköğretim birinci kademe fen ve teknoloji dersindeki ölçme-değerlendirmeye ilişkin görüşlerinin değerlendirilmesi. *Milli Eğitim Dergisi, 185*, 175-192.

Tanyel, A. (1999). *İlköğretim okullarında görevli sınıf öğretmenlerinin hizmet içi eğitim ihtiyacı* (Tez No. 81485) [Yüksek lisans tezi, Yıldız Teknik Üniversitesi]. Yükseköğretim Kurumu Tez Merkezi.

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse education in practice, 6*(6), 354-363. https://doi.org/10.1016/j.nepr.2006.07.002

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical education*, *42*(2), 198-206. https://doi.org/10.1111/j.1365-2923.2007.02957.x

Yüksel, K. B. (2019). *Yaratıcılığın ve madde yazarlığı eğitiminin ölçme aracının psikometrik özelliklerine etkisi* (Tez No. 484104] [Yüksek lisans tezi, Hacettepe Üniversitesi]. Yükseköğretim Kurumu Tez Merkezi.

Sadeleştirilmiş matematik maddeleriyle öğrenci performansı ve madde anlaşılırlığı arasındaki ilişki

Seher Yalçın ve Özgür Avcı

Giriş

Test sonuçlarına dayalı olarak verilen kararların önemli olması nedeniyle ölçme araçlarının titizlikle geliştirilmesi gerekmektedir. Bu kapsamda, geçerliği tehdit eden kaynakların ortadan kaldırılması ya da azaltılması önemlidir (Haladyna ve Downing, 2004). Türkiye'de 2017-2018 eğitim-öğretim yılından itibaren uygulanan Liseye Geçiş Sistemi (LGS)'ndeki matematik alt testi incelendiğinde, önceki yıllarda uygulanan ortaöğretime geçiş sınavlarından farklı olarak, soruların daha geniş bağlam içerisinde sunulduğu görülmektedir. Bu durum matematik alt testinde kullanılan maddelerin bazı dilsel özelliklerinin daha karmaşık hale gelmesine neden olabilir. Bu durum da maddenin anlaşılmasını zorlaştıracağından geçerliğin düşmesi söz konusu olabilir.

Maddelerin dilsel özelliklerinin sadeleştirilmesiyle ilgili çalışmalarda sadeleştirilmesi gereken dilsel özelliklere yönelik öneriler bulunmaktadır (Abedi, 1997, 2011). Bu öneriler incelenerek Türkçe yazılan maddeleri anlamayı güçleştirebilecek dilsel özellikler belirlenmiştir. Bu belirlemede uzman görüşlerinden, Türkçe okunabilirlik çalışmalarının sonuçlarından (Ateşman, 1997; Çetinkaya, 2010) ve Abedi (2006)'nin önerilerinden faydalanılmıştır.

Başarı testlerindeki maddelerin bazı dilsel özelliklerinin anlamayı güçleştirebilmesi ve bu test sonuçlarına dayalı olarak verilen kararların hayati olması bu çalışmanın yapılmasının önemli olduğunu düşündürmüştür. Bu nedenlerle bu çalışmada, matematik maddelerini anlamayı zorlaştırabilecek dilsel özelliklerin sadeleştirilmesiyle öğrencilerin madde üzerindeki performansı, maddeyi yanıtlama süreleri ve maddelerin anlaşılırlığı arasındaki ilişkiyi incelemek, ayrıca öğrencilerin yanıtlamak için madde seçiminde dilsel özelliklerin etkisinin olup olmadığını ve hangi dilsel özelliklerin madde seçiminde etkili olduğunu belirlemek ve maddelerin sadeleştirilme düzeyi ile maddeleri tercih eden öğrenci sayısı arasındaki ilişkiyi tespit etmek amaçlanmıştır.

Yöntem

Bu araştırmada, sadeleştirilmiş matematik maddeleriyle öğrenci performansı ve madde anlaşılırlığı arasındaki ilişkiyi sesli düşünme protokolleri ve yarı yapılandırılmış görüşmeler yoluyla incelemek

amaçlanmıştır. Bu amaç doğrultusunda araştırma çoklu durum çalışması modelinde tasarlanmıştır. Araştırmanın örneklemi 38 öğrenciden oluşmaktadır. Örneklemde yer alan 18 öğrenci ile sesli düşünme protokollerine hazırlık yapılmıştır. Geriye kalan 20 öğrenci ile esas çalışma yürütülmüştür.

Öğrencilerin yanıtladığı orijinal test formunda LGS'de matematik alt testinde kullanılan iki madde ve Milli Eğitim Bakanlığı'nın sitesinde yayınladığı Beceri Temelli Testler'den matematik alanındaki yedi madde kullanılmıştır. Orijinal formdaki maddelerin dilsel özelliklerinin sadeleştirilmesiyle sadeleştirilmiş test formu oluşturulmuştur. Maddeler kelime sıklığı/aşinalığı, tamlama uzunluğu, karmaşık cümle yapısı, cümle uzunluğu ve madde uzunluğu bakımından sadeleştirilmiştir.

Çalışmanın veri toplama süreci çevrimiçi olarak gerçekleştirilmiştir. Bu kapsamda, önce her bir öğrenciyle sesli düşünme protokolleri yürütülmüş ve hemen ardından yarı yapılandırılmış görüşmeler yapılmıştır. Sesli düşünme protokolleri sürecinde 10 kişilik öğrenci grubuna dilsel özellikleri sadeleştirilmeyen (orijinal) test formu uygulanmış ve diğer 10 kişilik öğrenci grubuna ise dilsel özellikleri sadeleştirilen (sadeleştirilmiş) test formu uygulanmıştır. Daha sonra yarı yapılandırılmış görüşmelerde öğrencilerden önce orijinal sonra sadeleştirilmiş maddeyi içinden okuması istenmiş ve dilsel özelliklerin sadeleştirilmesinin maddelerin anlaşılırlığı üzerindeki etkisi hakkında fikir verebilecek sorular yöneltilmiştir. Son olarak, öğrencilerle yürütülen sesli düşünme protokollerine ve yarı yapılandırılmış görüşmelere ait ses kayıtları yazıya geçirilmiş betimsel analiz uygulanmıştır. Görüşmelere ait yazılı kayıtlardan kodlar oluşturulmuştur. Daha sonra kodları içeren temalar oluşturulmuştur. Böylece sadeleştirilmiş maddeleri tercih eden öğrencilerin tercih gerekçeleri belirlenmiş ve elde edilen bulgular yorumlanarak sunulmuştur.

Sonuçlar

Çalışma sonuçlarına göre, öğrenciler orijinal ve sadeleştirilmiş maddeler arasında seçim yaptığında, sadeleştirilmiş maddelerin orijinal maddelere göre daha fazla yanıtlanmak istenmesindeki en etkili gerekçe madde uzunluğu dilsel özelliğidir. Yani öğrenciler genel olarak daha kısa maddeleri yanıtlamak istemektedirler. Kelime sıklığı/aşinalığı ve cümle uzunluğu dilsel özellikleri ise öğrencilerin maddeler arasında seçim yaparken sadeleştirilmiş maddeleri orijinal maddelere seçmesinde daha az etkilidir. Yani matematik maddelerinin daha yaygın kullanılan kelimelerle ve daha kısa cümlelerle oluşturulması, sadeleştirilmiş maddelerin orijinal maddelere tercih edilmesinde madde uzunluğu kadar etkili değildir. Ayrıca sesli düşünme protokolleri ve görüşmelere ait bulgular incelendiğinde, maddelerin bazı dilsel özelliklerini sadeleştirmenin maddenin ölçtüğü kazanıma ulaşan ve dil becerilerinde fazla eksiği olan öğrenciler için maddenin anlaşılırlığına katkı sağlayabileceği sonucuna ulaşılmıştır.

Kaynaklar

Abedi, J., Lord, C., and Plummer, J. R. (1997). *Final report of language background as a variable in NAEP mathematics performance*. National Center for Research on Evaluation, Standards, and Student Testing. https://cresst.org/wp-content/uploads/TECH429.pdf

Abedi, J. (2006). Language issues in item development. In S. M. Downing and T. H. Haladyna (Eds.), *Handbook of test development* (pp. 377-398). Lawrance Erlbaum Associates.

Abedi, J. (2011). Language issues in the design of accessible items. In S. N. Elliott, R. J. Kettler, P. A. Beddow, and A. Kurz (Eds.), *Handbook of accessible achievement tests for all students* (pp. 217-230). Springer.

Ateşman, E. (1997). Türkçede okunabilirliğin ölçülmesi. *Ankara Üniversitesi TÖMER Dil Dergisi, 58,* 171-174. https://dergipark.org.tr/tr/pub/mersinefd/issue/17396/181918

Çetinkaya, G. (2010). *Türkçe metinlerin okunabilirlik düzeylerinin tanımlanması ve sınıflandırılması* (Tez No. 265580) [Doktora Tezi, Ankara Üniversitesi]. Yükseköğretim Kurumu Tez Merkezi.

Haladyna, T. M. ve Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17-27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

COVID salgını döneminde Türkiye'de eğitimde ölçme alanında yapılan makalelerin içerik analizi

Seher Yalçın, Ezel Tavşancıl ve Çağla Alpayar

Giriş

Covid-19 salgını ile birlikte önce eğitime bir süre ara verilmiş ardından eğitimin devamlılığı için Milli Eğitim Bakanlığı ve Yüksek Öğretim Kurumu tarafından uzaktan eğitim uygulamalarına hızlı geçiş yaşanmıştır. Öğrenme sürecinin uzaktan eğitimle ve çoğunlukla dijital ortamlarda gerçekleştirilmesi ölçme ve değerlendirme uygulamalarını sınırlandırmış, ölçme ve değerlendirmeler çoğunlukla çevrimiçi sınav veya eve ödev şeklinde olmuştur. Ancak çevrim içi ortamda yapılan ölçme ve değerlendirmelerde kopya sorunları nedeniyle elde edilen puanların geçerliği ve güvenirliği kuşkuludur (Özer ve Suna, 2020). Ödev odaklı ölçme ve değerlendirme faaliyetleri ile süreci yürüten eğitmenler ise ödev okuma yükü ve geri bildirim verme zorluğunu dile getirmiştir (Kavan ve Adıgüzel, 2021). Öğretmenlerin tüm öğrencilere aynı puanı verip dersten geçirmesi, öğrencilere rastgele not verilmesi de öğrenciler hakkında verilecek kararlar için hayati önem taşıyabileceğinden yaşanan sorunlardan bir diğeridir (Sarı, 2020). Okul uygulamalarında yaşanan sorunlarla beraber, ulusal düzeyde yapılan geniş ölçekli sınavlar da süreçte zorluk yaşanan durumlardır. COVID salgının Türkiye'de başladığı Mart 2020 tarihinden beri bu konuda pek çok çalışma yapılmıştır. Ancak hızlı kararlar vermeyi ve politikalar geliştirmeyi gerektiren bu süreçte yapılan makaleler kadar makalelerde ele alınan konular, kullanılan araştırma türü, örnekleme ilişkin düzeyler, makalelerde ele alınan sorunlar ve çözüm önerilerinin neler olduğunu derleyebilmenin önemli olduğu gerekçesiyle bu çalışmada, 2020-2021 yıllarında Google akademik veri tabanında pandemi döneminde eğitimde ölçme alanında yapılan çalışmaların "konu alanı, araştırmanın modeli, örneklem türü, örneklem büyüklüğü, örneklem düzeyi, veri toplama aracı, verilerin analizi, ele alınan sorunlar ve öneriler" açısından incelenmesi amaçlanmıştır.

Yöntem

Bu çalışmada, 2020-2021 yıllarında "google akademik" veri tabanında pandemi döneminde eğitimde ölçme alanında yapılan makalelerin incelenmesi amaçlandığından betimsel içerik analizi türündedir. Makalelerin seçiminde amaçlı örnekleme yöntemlerinden ölçüt örnekleme yöntemi kullanılmıştır. Makalelerin seçim ölçütleri google akademik veri tabanında "covid and eğitimde ölçme" ile "pandemi and eğitimde ölçme" kavramlarını içermeleri, 2020-2021 yıllarında yapılmış olmaları ve

Türkiye örnekleminde veya Türk yazarlar tarafından yapılmış olmalarıdır. Bu veri tabanı, TRDizin, ERIC, SSCI gibi ulusal ve uluslararası indekslerde taranan makaleleri de içerdiği için seçilmiştir. Bu veri tabanında, 12.07.2021 tarihi itibariyle 2020 yılından beri yayınlanan 138 çalışmaya ulaşılmıştır. Bu çalışmalardan makale türünde olanlar ve doğrudan ilgili olan çalışmalar analize alınmıştır. Bu makaleler "konu alanı, araştırmanın modeli, örneklem türü, örneklem büyüklüğü, örneklem düzeyi, veri toplama aracı, verilerin analizi, ele alınan sorunlar ve öneriler" başlıklarında incelenmiştir. Araştırmacılar tarafından çalışmaya özgü bir form oluşturulmuştur. Tüm makaleler belirtilen başlıklara göre forma kodlanmış ya da yazılmıştır. Verilerin analizinde, içerik analizi türlerinden kategorisel ve frekans analizi kullanılmıştır. İçerik analizi, "sözel, yazılı ve diğer materyallerin içerdiği mesajı, anlam ve/veya dilbilgisi açısından nesnel ve sistematik olarak sınıflandırma, sayılara dönüştürme ve çıkarımda bulunmadır" (Tavşancıl ve Aslan, 2001, s. 22). Araştırmacılar tarafından paylaşılan makaleler kodlama formuna göre incelenmiştir. Bütün kodlama işlemi bittikten sonra kodlamalar arası tutarlılık incelenecektir. Bütün makalelerin inceleme süreci henüz tamamlanmamıştır.

Sonuçlar

Yapılan incelemelerde, salgın koşullarının güncelliği nedeniyle keşfedici bir yaklaşım izlenmiştir. Buna bağlı olarak yapılan çalışmalarda, nitel araştırma türünde ve görüşme yöntemi ile veri toplanmasının yaygın olduğu görülmektedir. Verilerin toplanmasında sıklıkla yapılandırılmış/yarı yapılandırılmış görüşme formlarından ve anketlerden yararlanılmıştır. Bu durum, salgının eğitim alanındaki etkilerini açıklayan değişkenlere yönelik standart araçların henüz yaygınlaşmaması ile açıklanabilir. Katılımcılar sıklıkla uygun ya da amaçsal örnekleme yoluyla belirlenmiştir. Alan yazındaki çalışmalardan (Şenyurt ve Özkan, 2017) farklı olarak katılımcılar çoğunlukla eğitimcilerden oluşmaktadır. Öğrencilerin katılımcı olduğu çalışmalar ise çoğunlukla üniversite düzeyindedir. Verilerin çözümlenmesinde alan yazındaki çalışmaları destekler şekilde (Kesim ve Türk, 2021) sıklıkla betimsel istatistiklerden ve içerik analizinden yararlanılmıştır. Araştırmalarda, ölçme ve değerlendirme uygulamalarına yönelik gözlenen sorunlar aşağıdaki gibi gruplandırılabilir:

- *Öğretmen kaynaklı*: Teknolojik alan bilgisi konusunda eksiklikleri (Baran ve Sadık, 2021), hızlı geri bildirim sağlayamamaları (Kavan ve Adıgüzel, 2021)

- *Öğrenci kaynaklı*: Fırsat eşitsizliği (Balaman ve Tiryaki, 2021), zamansızlık ve görev yükü artışı (Şahin, 2021), dezavantajlı bireylerin göz ardı edilmesi, görev takibinde isteksizlik (Bayram, 2021).

- *Ölçme aracı ve yöntemi kaynaklı*: Elektronik materyallerin ölçme ve değerlendirme boyutunun yetersiz kalması (Yüksekdağ, 2021), bazı derslerin doğasının çevrimiçi ölçme uygulamalarına elverişli olmaması (Dolmacı ve Dolmacı, 2020; Orşanlı ve Bekmezci, 2020), güvenlik sorunu (Tanrıkulu, 2021).

Araştırmalarda konuyla ilgili yapılan önerilerden bazıları; eğiticilerin, hizmet-içi eğitimlerle desteklenmesi gerekliliği (Karadağ, 2021), çevrimiçi kaynakların ölçme ve değerlendirme yönünde zenginleştirilmesinin ve uygulamaların merkezîleştirilmesinin gerektiği şeklindedir (Balaman ve Tiryaki, 2021).

Kaynaklar

Balaman, F. and Tiryaki, S. H. (2021). Corona virüs (Covid-19) nedeniyle mecburi yürütülen uzaktan eğitim hakkında öğretmen görüşleri. *Itobiad: Journal of the Human & Social Science Researches*, *10*(1), 52-84. https://doi.org/10.15869/itobiad.769798

Bayram, H. (2021). Challenges secondary school teachers face during the distance education process. *International Journal of Eurasian Education and Culture*, *6*(12), 613-658. http://dx.doi.org/10.35826/ijoecc.306

Dolmacı, M. ve Dolmacı, A. (2020). Eş zamanlı uzaktan eğitimle yabancı dil öğretiminde öğretim elemanlarının görüşleri: Bir covid-19 örneği. *Türk Eğitim Bilimleri Dergisi*, *18*(2), 706-732. https://doi.org/10.37217/tebd.783986

Karadağ, N. (2021). Açıköğretim sisteminde çevrimiçi sınav uygulamasının sınav hazırlama sürecine etkisi. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi, 7*(1), 45-60. https://dergipark.org.tr/tr/pub/auad/issue/60075/857187

Kavan, N. ve Adıgüzel, A. (2021). Türkçe öğretmenlerinin salgın süreci eğitim faaliyetlerine ilişkin görüşlerinin incelenmesi. *Electronic Journal of Education Sciences*, *10*(19), 138-155. https://dergipark.org.tr/tr/pub/ejedus/issue/62791/939231

Kesim, E. ve Türk, Y. K. (2021). Türkiye'de mesleki ve teknik eğitim alanında hazırlanan lisansüstü tezlerin incelenmesi. *Anadolu Journal of Educational Sciences International*, *11*(1), 287-322. https://doi.org/10.18039/ajesi.815033

Orçanlı, K. ve Bekmezci, M. (2020). Üniversite öğrencilerinin Covid-19 pandemisinde uzaktan eğitim algısının belirlenmesi ve bazı demografik değişkenlerle ilişkisi. *Uluslararası İktisadi ve İdari Bilimler Dergisi, 6*(2), 88-108.

Özer, M. ve Suna, H. E. (2020). COVİD-19 salgını ve eğitim. M. Şeker, A. Özer ve C. Korkut (Eds.), *Küresel Salgının Anatomisi İnsan ve Toplumun Geleceği* içinde (ss. 171-192). Türkiye Bilimler Akademisi.

Sarı, H. İ. (2020). Evde kal döneminde uzaktan eğitim: Ölçme ve değerlendirmeyi neden karantinaya almamalıyız? *Uluslararası Eğitim Araştırmacıları Dergisi*, *3*(1), 121-128. https://dergipark.org.tr/tr/pub/ueader/issue/55302/730598

Şahin, M. (2021). Opinions of university students on effects of distance learning in Turkey during the COVID-19 pandemic. *African Educational Research Journal, 9*(2), 526-543. https://doi.org/10.30918/AERJ.92.21.082

Şenyurt, S., ve Özkan, Y. Ö. (2017). Eğitimde ölçme ve değerlendirme alanında yapılan yüksek lisans tezlerinin tematik ve metodolojik açıdan incelenmesi. *İlkogretim Online*, *16*(2), 628-653. https://doi.org/10.17051/ilkonline.2017.304724

Tanrıkulu, F. (2021). Barriers encountered by Turkish teachers in the use of digital environment and content in the distance education process. *Journal of Language Education and Research*, *7*(1), 78-120. https://doi.org/10.17051/ilkonline.2017.304724

Tavşancıl, E. ve Aslan, E. (2001). *Sözel, yazılı ve diğer materyaller için içerik analizi ve uygulama örnekleri.* İstanbul.

Yüksekdağ, B. B. (2021). Covid-19 pandemisi döneminde öğrenme ve uzaktan hemşirelik eğitiminde paradigma değişimi. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi*, *7*(1), 61-73. https://dergipark.org.tr/tr/download/article-file/1519329

Yapay sinir ağları ile madde parametre kestiriminin etkililiğinin incelenmesi

Eda Akdoğdu ve Kübra Atalay Kabasakal

Giriş

Ölçme aracına cevap veren bireylerin yeteneği, madde parametreleri ve cevapları arasındaki ilişkiyi modelleyen Madde Tepki Kuramı (MTK) güçlük (b) ve ayırt edicilik (a) parametrelerinin maddeden ve bireyden bağımsız kestirilebilmesi gibi sağladığı avantajlarla eğitimde ve psikolojideki ölçme uygulamalarında oldukça sık kullanılmaktadır (Embretson ve Reise, 2000; Harvey ve Hammer, 1999). Özellikle test geliştirme ve uygulamasında oldukça kullanışlı ve avantajlı olan MTK, parametrelerin az hata ile kestirilebilmesi için geniş madde havuzuna ve örneklem büyüklüğüne ihtiyaç duymaktadır (Goldman ve Raju, 1986; Hulin ve diğ., 1982). Sosyal bilimlerde yeni yeni kullanılmaya başlayan ve geleneksel analitik yöntemlerle ölçülmesi zor olan verilerin veya karmaşık öznitelik ilişkilerinin bulunduğu çeşitli uygulama alanlarında giderek daha fazla kullanılan (Shanmuganathan, 2016) yapay sinir ağlarının madde parametrelerinin kestiriminde küçük örneklemler için destek sağlayabileceği düşünülmektedir. Yapay Sinir Ağları (YSA) insan beyni gibi öğrenerek yeni bilgiler üretebilmek amacıyla gerçekleştirilen bilgisayar sistemleridir. YSA algoritması verilen girdi ve çıktılardaki değişikleri öğrenerek yeni gelen girdi seti için uygun çıktıyı verecek bir model oluşturur (Tan vd, 2016). Öğrenme süreci sonunda elde edilen çıktılar ve hedef çıktılar arasındaki farkın en aza indirilmesi hedeflenir (Tan ve diğ., 2016). Ayrıca model ile kestirimde oluşacak hatalar eğitim ve test sürecinde belirlenebilir. Büyük örneklemlere ulaşılamayan durumlarda YSA'nın MTK madde parametreleri kestirimi için bir alternatif olup olamayacağı incelenmek istenmiştir. Bu sebeple çalışmada küçük ve farklı örneklem büyüklüklerinde (n= (38, 75, 188, 375)) KTK'ya dayalı güçlük ve ayırt edicilik parametreleri ile eğitilmiş yapay sinir ağları ve MTK'ya dayalı analizlerle a ve b parametrelerinin kestirimi ve sonuçların karşılaştırılması amaçlanmaktadır.

Yöntem

MTK'ye dayalı ve yapay sinir ağları ile kestirilen madde parametrelerindeki hatanın farklı örneklem büyüklüklerinde nasıl değişeceğini gösteren bu araştırma Monte Carlo simülasyon çalışması aracılığıyla ele alınmıştır. Çalışma veri oluşturma, istatiksel analiz ve sonuçların özetlenmesi olarak üç adımdan oluşmaktadır ve bu adımlar R (4.1.0) yazılımında gerçekleştirilmiştir. Veri üretiminde 2PL

modele göre, a ve b parametreleri sırasıyla log-normal [$a \sim lnN(0.0, 0.2)$] ve normal dağılımdan [$b \sim N(0, 1)$] çekilmiştir. Ayrıca yetenek parametrelerinin belirlenmesi için de yine normal dağılım [$\theta \sim N(0, 1)$] tercih edilmiştir. 2000 birey ve 100 madde için üretilen 1-0 puanlanan veri setlerinin gerçek a ve b madde parametreleri "mirt" (Chalmers, 2012) paketi ile, güçlük ve ayırt edicilik "CTT" (Willse, 2018) paketi ile kestirilmiştir. Çalışmanın istatiksel analiz ve sonuçların özetlenmesi adımları ise şu şekilde devam etmiştir:

1. Üretilen 2000 bireyden oluşan veri setinden rastgele 50, 100, 250, 500 örneklem büyüklüğünde veriler çekilmiştir.
2. Çekilen her bir örneklem büyüklüğü için yapay sinir ağlarında (YSA) eğitim veri ve test veri seti oluşturularak (%75 eğitim verisi, %25 test verisi) eğitim veri setinde tek katman, tek nöronlu, geri beslemeli ve doğrusal aktivasyon fonksiyonuyla güçlük ve ayırt edicilik analizleri ile a ve b parametreleri iki ayrı modelde "neuralnet" (Fritsch ve diğ., 2019) paketi ile modellenmiştir.
3. Çekilen her bir örneklem büyüklüğü için 2PL model altında a ve b parametreleri "mirt" (Chalmers, 2012) paketi ile kestirilmiştir.
4. YSA ve MTK'ya göre yapılan kestirimlerde hataların ortalama karekökü (RMSE) ve ortalama mutlak hata (MAE) değerleri elde edilmiştir.
5. Bu süreç 50 replikasyonla gerçekleştirilerek a ve b parametrelerinin kestiriminde n= (38, 75, 188, 375) örneklem büyüklüklerinde RMSE ve MAE değerlerinin ortalamaları elde edilmiştir.

Sonuçlar

Tablo 1

Küçük Örneklem Büyüklüklerinde YSA ve MTK için Madde Parametresi Kestirim Hataları

Örneklem büyüklüğü	YSA (neuralnet)				MTK (mirt)			
	a		b		a		b	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
38	0.124	0.100	0.064	0.046	0.567	0.426	0.311	0.243
75	0.132	0.110	0.066	0.050	0.426	0.341	0.183	0.150
188	0.099	0.080	0.061	0.050	0.264	0.219	0.125	0.100
375	0.178	0.141	0.070	0.055	0.161	0.132	0.081	0.066

YSA ve MTK'da küçük örneklemlerde a ve b parametrelerinden elde edilen hata değerleri Tablo 1'de yer almaktadır. Bu hatalar incelendiğinde YSA ile tahmin etme sürecinde örneklem büyüklükleri arttıkça hatalarda dramatik olarak bir artış azalış olmadığı, MTK'da ise beklendiği gibi örneklem büyüklüğü arttıkça hatanın azaldığı görülmüştür. Her iki yöntemde de b parametrelerinin a parametresine göre daha iyi kestirildiği tespit edilmiştir. Ayrıca hem a hem b parametresi için 38, 75 ve 188 örneklem büyüklüklerinde YSA'nın daha iyi kestirim yaptığı ve 375 örneklem büyüklüğünde ise

yalnızca a parametresi için MTK'dan kötü sonuçlar verdiği sonucuna ulaşılmıştır. Yapılan 50 replikasyon sonucunda YSA 188 örneklem büyüklüğü için en iyi tahminleri vermiştir. Bu örneklem büyüklüğü civarında tekrar simülasyon çalışmaları yapılarak uygun örneklem büyüklüğü tespit edilebilir. Ayrıca parametreleri tahmin edilecek madde sayısı değiştiğinde nasıl sonuçlar alınacağı, test veri setinde hataların nasıl farklılaşacağı, farklı bir makine öğrenme algoritmasıyla daha az hata elde edilip edilemeyeceği incelenebilir.

Kaynaklar

Chalmers, R. P. (2012). *mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Embretson, S. E., and Reise, S. P. (2013). *Item response theory for psychologists.* Lawrence Erlbaum Associates.

Fritsch, S., Guenther, F., and Wright M. N. (2019). *neuralnet: Training of neural networks* (version 1.44.2) [Computer Software]. https://cran.r-project.org/package=neuralnet

Harvey R. J, Hammer, A. L. (1999) Item response theory. *The Counseling Psychologist, 27*(3), 353-383. https://doi.org/10.1177/0011000099273004.

Shanmuganathan, S. (2016). Artificial neural network modelling: An introduction. In S. Shanmuganathan S. and S. Samarasinghe (Eds.), *Artificial Neural Network Modelling* (pp. 1-14). Springer. https://doi.org/10.1007/978-3-319-28495-8_1

Tan, P. N., Steinbach, M., and Kumar, V. (2016). *Introduction to data mining.* Pearson Education India.

Willse, J. T. (2018). *CTT: Classical test theory functions* (version 2.3.3) [Computer Software]. https://cran.r-project.org/package=CTT

Investigation of performances of the ω and M_4 indexes in tests consisting of subtests based on a bifactor model under various conditions by using two-stage anlaysis via *lz**

Ebru Balta ve Arzu Uçar

Introduction

Educational and psychological tests have different subsections based on content categories. For example, a test of general ability can have subsections on mathematics, reading, and writing. Subsections based on content categories which have added value (Haberman et al. 2009) over total scores are perceived as subtests, so that test scores obtained from these subtests can be reported. There is an increasing interest in reporting subscores at examinee level. A subscore may be considered useful only when it provides a more accurate measure of the construct being measured than is provided by the total score (Haberman, 2005). Aberrant testing behaviors damages the validity of the psychometric properties of test and subtest scores. Thus, it is important to investigate aberrant testing behaviors in tests consisting of subtests that correlate with each other. There are several common types of aberrant testing behaviors to detect, including answer-copying, pre-knowledge cheating, creative thinking, lucky guessing, and random responding (Cizek and Wollack, 2017; Haberman and Lee, 2017; Karabatsos, 2003; Kingston and Clark, 2014; Lee and Haberman; 2016; Sijtsma and Meijer, 1992; Sinharay, 2017). The purpose of this study is to investigate the performances of answer copying and similarity indices in tests consisting of subtests that are related to each other based on bifactor model under various conditions by using two stage analysis via person fit statistic.

Method

In the study, the sample size was fixed at 3000, and the test length was 80 items using the Bifactor Item Response Theory, which is one of the models of Multidimensional Item Response Theory. In addition, the rate of individuals whose data was manipulated was fixed at 5%. In line with the purpose of the study, 900 dichotomous data sets were generated with 100 replication under the conditions of correlation between subtests (0.60, 0.70, 0.80) and copiers' ability (low ability, mid ability, high ability) for obtained means of Type I error rate of methods. In order to obtain the mean power rate of the methods, 4000 dichotomous data sets were generated with 100 replication under the conditions of correlation between subtests, copiers' ability and copying ratio (5%, 10%, 20%, 40%, 60%). There were

4900 data files for the power and type I error analysis. Each of the data files included 3000 data on person ability which formed a normal distribution. ω (omega) index and M_4 index (Maynes, 2014) were used in two-stage analyses to detect individuals who were suspected of being copiers. In the two-stage analyses, the ω and M_4 indexes were used after person fit statistics which are *lz** (Magis, Raîche, & Béland, 2012) . R.4.0.1 software was used for data generation and analyses. The ω index provides a measure of the standardized difference between two individuals' (C (suspected copier) and S (source)) answer matches (both true and false). The M4 index is using a generalized trinomial distribution to model the joint distribution of the number of matching correct responses and matching incorrect responses. We calculated the probability values of the copying and similarity indexes using the *"CopyDetect"* package (Zopluoglu, 2018) to obtain the type I error and power rate. We compared the probability values of the answer copying and similarity indexes with $\alpha = .05$. We assigned 1 to those with probability values less than or equal to .05, and 0 to those with a greater probability. [A value of 1 represents "test fraud due to answer copying" and 0 represents "no test fraud due to answer copying" for the suspected copier and source pair]. In the two-stage analysis, we first calculated the person fit statistic (for *lz** statistic) with the *"PerFit"* package (Tenderio, 2018).

Results

According to the results of the two-stage analysis performed in tests consisting of subtests that are related to each other based on bifactor model, the performances of answer copying (ω) and similarity indexes (M_4) were found to be higher under the conditions of very low copy ratio (5%) and high copy ratio (60%) compared to the medium copy ratio (40%). However, in the two-stage analysis performed in the total test, the ω and M_4 indexes showed lower performance than the subtest analysis. Similrlay, in the subtest analysis, the performances of answer copying (ω) and similarity indexes (M_4) were found to be higher under the conditions of correlation between subtests (0.60, 0.70).

References

Cizek, G., and Wollack, J. (2017). Identification of item preknowledge by the methods of information theory and combinatorial optimization. In G. Cizek and J. Wollack (Eds.), *Handbook of Quantitative Methods for Detecting Cheating on Tests,* (pp.217-233). Routledge.

Haberman, S. J. (2005). *When can subscores have value?* (ETS RR-05-08). Educational Testing Service.

Haberman, S., & Lee, Y. (2017). *A statistical procedure for testing unusually frequent exactly matching responses and nearly matching responses.* (Research Report No: RR-17-23). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12150

Haberman, S. J., Sinharay, S., and Puhan, G. (2006). *Subscores for institutions.* (ETS RR-06-13). Educational Testing Service. https://files.eric.ed.gov/fulltext/EJ1111382.pdf

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*(4), 277–298. https://doi.org/10.1207/s15324818ame1604_2

Kingston, N., & Clark, A. (2014). *Test fraud: Statistical detection and methodology.* Routledge.

Lee, Y., & Haberman, S. (2016). Investigating test-taking behaviors using timing and process data. *International Journal of Testing, 16*(3), 240–267. https://doi.org/10.1080/15305058.2015.1085385

Magis, D., Raîche, G., & Be´land, S. (2012). A didactic presentation of Snijders's l(z)* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics*, *37(1)*, 57–81. https://doi.org/10.3102/1076998610396894

Maynes, D. D. (2005). M_4: *A new answer-copying index*. Unpublished manuscript, Caveon Test Security, Midvale, UT. Retrieved from https://www.caveon.com/

Maynes, D. D (2014a). Detection of non-independent test taking by similarity analysis. In N. M. Kingston, and A. K. Clark (Eds.), *Test Fraud: Statistical Detection and Methodology* (pp. 52-80). Routledge.

Wollack, J. A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement*, *21*(4), 307–320. https://doi.org/10.1177/01466216970214002

Wollack, J. A. (2006). Simultaneous use of multiple answer copying indexes to improve detection rates. *Applied Measurement in Education*, *19*(4), 265–288. https://doi.org/10.1207/s15324818ame1904_3

Wollack, J. A., and Cohen, A. S. (1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement, 22*(2), 144–152. https://doi.org/10.1177/01466216980222004

Wollack, J. A., & Maynes, D. (2011, April). *Detection of test collusion using item response data* [Paper presentation]. National Council on Measurement in Education Annual Meeting. New Orleans, LA, USA.

Sijtsma, K., and Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's non-parametric IRT model. *Applied Psychological Measurement, 16*(2), 149–157. https://doi.org/10.1177/014662169201600204

Sinharay, S. (2017a). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics*, *42*, 46–68. https://doi.org/10.3102/1076998616673872

Sinharay, S. (2017b). Which statistic should be used to detect item preknowledge when the set of compromised items is known? *Applied Psychological Measurement, 41*(6), 403–421. https://doi.org/10.1177/0146621617698453

Tendeiro, J. N. (2018). *PerFit: Person Fit* (version 1.4.3) [Computer software]. https://cran.rproject.org/web/packages/PerFit/PerFit.pdf

Zopluoglu, C. (2018). *CopyDetect: Computing response similarity indices for multiple-choice tests* (version 1.3) [Computer software]. https://cran.rproject.org/web/packages/CopyDetect/CopyDetect.pdf

Veri madenciliği yöntemleri ile TIMMS 2019 Türkiye örneği Matematik başarısını sınıflamada belirlenen algoritmaların başarı oranlarının karşılaştırılması

Yasemin Yardım ve Tuncay Öğretmen

Giriş

Günümüzde teknolojinin büyük bir hızla gelişmesi sonucu ihtiyaçlar farklılaşmış ve bu farklılaşma bireysel ve toplumsal olarak beraberinde bazı değişimler getirmiştir. Toplumlar bu değişime ayak uydurabilmeleri için bilgiyi olduğu gibi almak yerine bilgiyi düzenleme, sunabilme ve günlük hayata uygulayabilme becerisini geliştirmek durumunda kalmışlardır. 21. yüzyıl becerileri denilen eleştirel düşünme, yaratıcılık, iletişim, işbirliği, bilgi ve bilim okuryazarlığı gibi pek çok becerinin yanında matematik okuryazarlığı da çok önemli bir yer tutmaktadır.

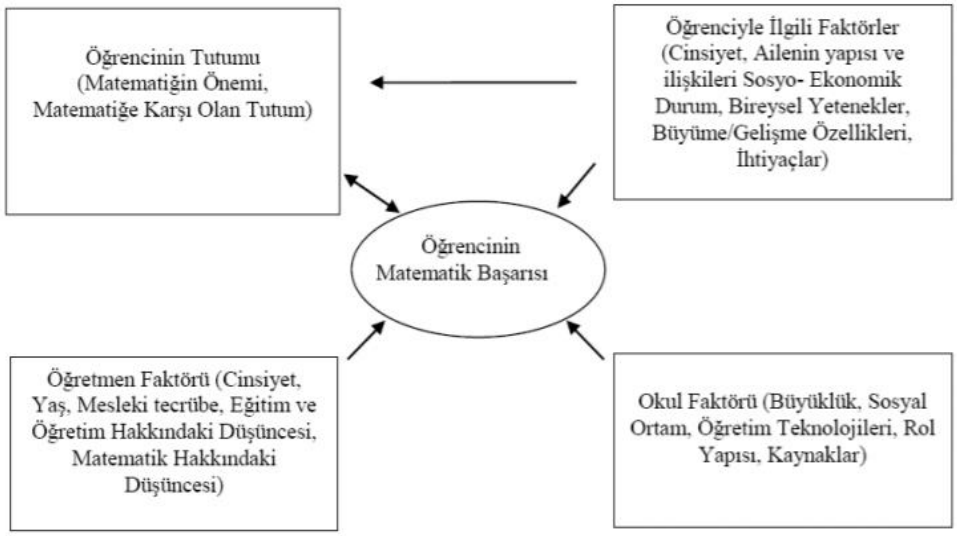
Son yıllarda yapılan Liselere Giriş Sınavı (LGS) ve Yüksek Öğretime Geçiş Sınavı (YKS) gibi ulusal sınavlarda öğrencilerin başarısının en düşük olduğu dersin matematik olduğu görülmektedir. Örneğin 2020-2021 eğitim öğretim yılında yapılan LGS sınavında matematikten 20 soru sorulmuş ve matematik doğru cevap ortalaması 4.20 olarak bulunmuştur. Bu sonuçlar 2020 yılı ile karşılaştırıldığında matematik doğru cevap ortalamasının birbirine oldukça yakın olduğu görülmüştür (MEB, 2021). Benzer şekilde 2019-TIMMS (Trends in International Mathematics and Science Study) sonuçları incelendiğinde Türkiye matematik ortalama puanının 496 olduğu ve TIMMS genel ortalaması olan 500 puanın altında kaldığı görülmektedir (TIMMS, 2019).

Ülkemizde yapılan ve uluslararası yapılan sınavlarda öğrencilerin matematik başarılarının düşük olması araştırmacıları bu düşük başarı oranının nedenlerini araştırmaya itmiştir. Yapılan çalışmalar incelendiğinde öğrenci başarısını etkileyen faktörlerin oldukça fazla olduğu görülmektedir. Öğrenmeyi hem bireyin özellikleri hem de etkileşimde bulunduğu çevrenin özellikleri etkilemektedir. Bireysel farklılıklardan kaynanan faktörler; zekâ, bilişsel stil farklılıkları, bireyin genel yeteneği, ön bilgileri, öğrenme yöntemi, duyuşsal özellikleri, cinsiyet, motivasyon, dikkat, kaygı, yaş olarak sayılabilir (Savaş, 2010).

Savaş ve diğ. yaptıkları bir araştırmada matematik başarısını etkileyen faktörleri Şekil 1'deki gibi özetlemişlerdir. Bu faktörler öğrencinin matematiğe karşı tutumu başta olmak üzere, öğrenciyle ilgili faktörler, öğretmen faktörü ve okul faktörü olarak dört grupta toplanmıştır (Savaş, 2010).

Şekil 1

Matematik Başarısını Etkileyen Faktörler



TIMSS (Trends in International Mathematics and Science Study-Uluslararası Matematik ve Fen Bilimleri Eğilimleri), dünya çapında matematik ve fen bilimlerinde öğrenci başarısını uluslararası karşılaştırmalı olarak ortaya koyan bir çalışmadır. 1995'ten bu yana dört yıllık bir değerlendirme döngüsünde yürütülen TIMSS, 1995, 1999, 2003, 2007, 2011, 2015 ve 2019'da dördüncü ve sekizinci sınıflardaki öğrenci başarısını yedi kez değerlendirerek ve 24 yıllık eğilim ölçümlerini biriktirmiştir. TIMSS, öğrencilerin matematik ve fen başarıları yanında ulusal, ev, okul ve sınıf bağlamları hakkında zengin bir bilgi dizisi toplar. Bu bağlam verileri, matematik ve fen başarısı ile ilgili eğitim faktörleri hakkında uluslararası karşılaştırmalı bakış açıları sağlar.

2019 yılında TIMSS, "eTIMSS" adlı kâğıt kalem değerlendirmesinin dijital bir versiyonunu sunarak bilgisayar tabanlı değerlendirmeye geçişe başlanmıştır. eTIMSS 2019 başarı verileri, hem eTIMSS hem de paperTIMSS değerlendirmeleri arasında göreceli performans karşılaştırmalarına izin verirken, aynı zamanda fazla mesai trend ölçümlerinin karşılaştırılabilirliğini sağlamak için ölçeklenmiştir (Foy ve diğ., 2020; von Davier, 2020). eTIMSS ve paperTIMSS arasında bir köprü sağlamak için, eTIMSS ülkeleri ayrıca paperTIMSS trend maddelerini tipik olarak aynı okullardaki ayrı bir öğrenci örneklemine uygulanmıştır. TIMSS 2019 Bridge verileri, eTIMSS ülkelerinin 2019'daki bilgisayar tabanlı verileri ile 2015'teki kâğıt tabanlı verilerinin yanı sıra TIMSS 2019 ülkelerinden gelen veriler arasında bir ara bağlantı (veya köprü) oluşturmaktadır.

Dördüncü ve sekizinci sınıflarda matematik ve fen eğitimini iyileştirmeyi amaçlayan ikincil analizleri desteklemek ve teşvik etmek için TIMSS 2019 Uluslararası Veritabanı, TIMSS 2019 projesi tarafından toplanan ve analiz edilen verileri araştırmacıların, analistlerin ve diğer kullanıcıların kullanımına sunmaktadır. Veritabanı, 64 ülke ve 8 kıyaslama katılımcısı için öğrenci başarı verilerinin yanı sıra öğrenci, ev, öğretmen, okul ve ulusal bağlam verilerini içerir. Her iki sınıf için ve Bridge verileri

de dahil olmak üzere, veritabanı, 682.680 öğrenci, 387.227 veli, 65.306 öğretmen, 24.316 okul müdürü ve her katılımcı ülkenin Ulusal Araştırma Koordinatörleri için kayıtları içerir. Bu çalışmada kullanılan veri setinde Türkiye örnekleminde 209 okuldan toplam 4077 sekizinci sınıf öğrenci verisi bulunmaktadır (TIMMS, 2019).

Yapılan araştırmalara bakıldığında öğrencilerin matematik başarıları üzerinde matemetiğe karşı tutumlarının önemli rol oynadığı görülmüş ve bu çalışmada TIMMS Türkiye 8. Sınıf öğrencilerinin Matematik başarılarını matematiğe karşı tutum ölçeğinden aldıkları puanlara göre veri madenciliği teknikleri kullanarak sınıflamada en başarılı algoritmaları araştırmak amaçlanmıştır (https://timss2019.org/international-database/., 2019).

Veri madenciliği ve makine öğrenmesi veriyi işlemek ve bilgiye ulaşmak için kullanılan en önemli yöntemlerden biridir. Bu yöntemler veri sayısının giderek artmasıyla sıklıkla kullanılmaya başlanmıştır. Toplanan verinin miktarı büyüdükçe, boyutu ve karmaşıklığının da artması ile birlikte, veri madenciliği ve makine öğrenmesi teknolojik ilerleme için gerekli bir bileşen olarak ortaya çıkmaktadır. Fazla sayıda verinin işlenmesi, analiz edilmesi önemli sonuçlar ortaya çıkarmaktadır. Makine öğrenmesi, birçok farklı istatistiksel prosedürü içeren disiplinler arası bir çalışma alanıdır. Makine öğrenmesi yöntemlerinin standart istatistik yöntemlerdeki gibi varsayımlarının olmaması yöntemlerinin tercih nedenlerinden birisidir. Örneğin, doğrusal regresyon gibi geleneksel istatistiksel yöntemlerden farklı olarak, makine öğrenmesi yöntemleri doğrusallık, homojenlik veya normallik varsayımları gerektirmez. Ayrıca veri madenciliği ve makine öğrenmesi, eldeki verileri bilgiye dönüştürmek; kümelemek, sınıflandırma yapmak, tahmin etmek için büyük veri setlerine uygulamakla ilgilidir. Makine öğrenmesi eğitim, tıp, astronomi, finans gibi farklı alanlarda sıklıkla uygulanmaktadır. Bu alanlardan eğitim alanında kullanılabilen makine öğrenmesi algoritmaları Eğitim Veri Madenciliği (EVM) olarak adlandırılır. EVM, eğitimcilere ve eğitim planlamacılarına büyük ve karmaşık eğitim veri kümelerini daha iyi anlamalarını ve bu veri kümelerinden çıkardıkları sonuçlara göre durum tespiti yapmalarına ve gelecek için planlamalar yapmalarına olanak sağlar. Temel istatistiksel analizlerle, veri kümelerinin genel davranışları yorumlanabilir. Bununla birlikte, bilinmeyen veya erişilemeyen yararlı bilgilerin keşfi EVM yöntemleri kullanılarak yapılabilir. Çıkarılan faydalı bilgiler, eğitimciler ve eğitim alanındaki karar vericiler tarafından kullanılabilir. Ayrıca, bu bilgiler eğitim sisteminin en önemli bileşeni olan öğrencilerin başarılarının mevcut durumunu izler ve bu nedenle hatalı eğitim stratejilerine çözümler önerilebilir. TIMSS, yalnızca eğitim politikalarının etkileri ve uygulamaları ile ilgili güvenilir bilgi sağlamakla kalmaz, aynı zamanda öğrenci başarısı açısından, katılımcı ülkelerin sonuçları arasında karşılaştırma yapmalarını sağlar (Filiz, 2019).

Bu araştırmanın amacı, EVM çerçevesinde TIMSS çalışmalarının güncel literatürüne katkıda bulunmaktır. Çalışmada TIMSS-2019 Türkiye sekizinci sınıf öğrencilerinin matematik başarılarının öğrenci matematik tutum anketinden elde edilen verilere göre sınıflandırılması için hangi EVM yönteminin daha uygun olduğunu belirlenmeye çalışılmıştır. En uygun EVM yöntemini bulmak için,

EVM çalışmalarında literatürde en sık kullanılan algoritmalar seçilmiştir. Bunlar; K-En Yakın Komşu (K-NN), Naive Bayes (NB), Yapay Sinir Ağları (YSA) ve Karar Ağacı (KA-J48) algoritmalarıdır.

TIMSS-2019 Türkiye sekizinci sınıf öğrencilerinin matematik başarılarının öğrenci matematik tutum anketinden elde edilen verilere göre sınıflandırılması için hangi Eğitim Veri Madenciliği yöntemi daha uygundur?

Veri Madenciliği

Veri madenciliğinin bir alt kolu olan makine öğrenmesi, veriden öğrenmeye dayalı yöntemler olarak düşünülebilir. Farklı bir bakış açısıyla bilgisayarın kendi kendine problemi çözmeyi öğrenmesi olarak tanımlanabilir. Makine öğrenmesi, elde edilen çıktıların sınıflandırılmasını, kümelenmesini ya da tahmin edilmesini sağlayacak algoritmaları barındırır. Makine öğrenmesinde farklı algoritmalar kullanılır. Eğer veri setinde çıktı biliniyorsa danışmanlı, bilinmiyorsa danışmansız algoritmalardan yararlanılır (Akın, 2014).

Veri madenciliği yöntemleri temel olarak tahmin, sınıflandırma, kümeleme ve birliktelik kuralları şeklinde 4 bölümden oluşur. Tahmin yöntemleri olarak Regresyon analizi, Bayes ağları, LR, KA ve YSA; sınıflandırma için k-NN, NB, YSA, KA, DVM ve LR; kümeleme için k-means, modele dayalı kümeleme, tam bağlantı kümeleme; birliktelik kuralları için Apriori, Carma, Sequence, GRI, Eclat, FP-Growth gibi yöntemler kullanılır (Li, 2003)

Sınıflandırma kavramı, basitçe bir veri kümesi (data set) üzerinde tanımlı olan çeşitli sınıflar arasında veriyi dağıtmaktır. Sınıflandırma algoritmaları, verilen eğitim kümesinden bu dağılım şeklini öğrenirler ve daha sonra sınıfının belirli olmadığı test verileri geldiğinde doğru şekilde sınıflandırmaya çalışırlar. Veri kümesi üzerinde verilen bu sınıfları belirten değerlere etiket (label) ismi verilir ve gerek eğitim gerekse test sırasında verinin sınıfının belirlenmesi için kullanılırlar (Şeker, 2021).

K-En Yakın Komşu Algoritması

Sınıflandırmada (classification) kullanılan bu algoritmaya göre sınıflandırma sırasında çıkarılan özelliklerden (feature extraction), sınıflandırılmak istenen yeni bireyin daha önceki bireylerden k tanesine yakınlığına bakılmasıdır (Şeker, 2021).

Başka bir deyişle, bir veri setinde sınıflandırma yapmak için değişkenler arasında en yakın komşuları bulur. Buradaki en önemli nokta veri noktaları arasındaki mesafedir. Genellikle Manhattan, Minkowski, Mahalanobis ve Öklid gibi uzaklık ölçüleri veriler arasındaki mesafenin hesaplanmasında kullanılır. (Filiz, 2019).

Naive Bayes Algoritması

NB algoritması veri madenciliğinde araştırmacılar tarafından kullanılan en etkili öğrenme algoritmalarından biri olarak kabul edilmektedir (Zhang, 2004). Bir tür Bayes ağı olan bu algoritmanın en iyi şekilde çalışması için iki koşulun gerçekleşmesi gerekir. Birinci koşul, sınıfların belirli koşullar

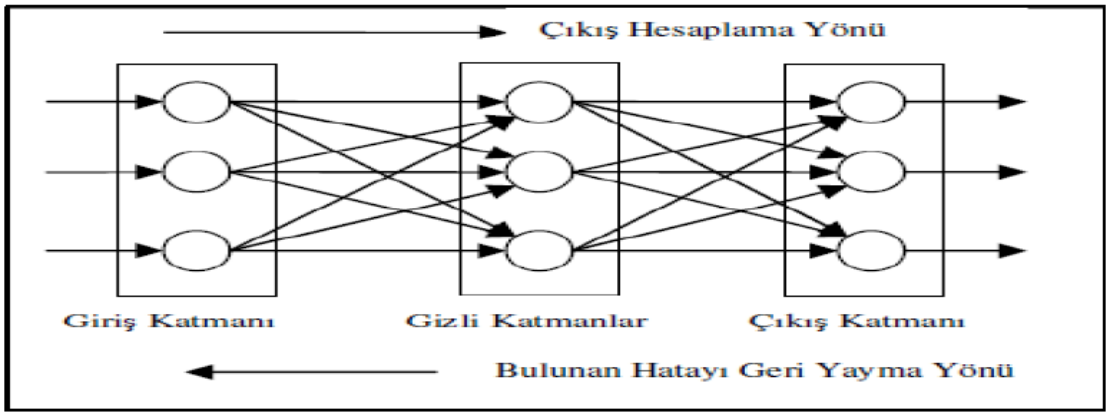
altında birbirlerinden bağımsız olmasıdır. İkincisi koşul ise sonuçları etkileyebilecek olan değişkenlerin belirli olmasıdır (John ve Langley, 1995). Bağımsızlık varsayımı yerine getirildiğinde, NB sınıflandırıcılarının öğrenme süreci daha basit hale gelir ve gözlemlenebilir faktörlerin vektörü kullanılarak en uygun şekilde atama belirlenir (Öz, 2016). Ayrıca eldeki veri setinin birleşik olasılıklarının sınıflandırma başarılarının bulunmasında da NB sınıflandırma algoritmasından yararlanılmaktadır (Amasyalı, 2006).

Yapay Sinir Ağları

İnsan beyni süreçlerini taklit edecek şekilde çalışan YSA, geçmişteki deneyimlerin sonuçlarına dayanarak hareket eder. Bu nedenle, karmaşık doğrusal olmayan durumlarla çalışırken bile, YSA modellerinde standart istatistik yöntemlerinde olduğu gibi varsayımlara gerek duymaz. Çok katmanlı algılayıcı (ÇKA), eğitim sürecinde geri yayılım algoritması kullanan bir yapay sinir ağı modelidir. ÇKA bir giriş katmanı, gizli katmanlar ve bir çıkış katmanı olarak üç unsurdan oluşur. Eldeki bilgi bir nörondan diğerine ağırlık değeri ile taşınır. Bir ÇKA algoritmasının ilk adımı rasgele ağırlıklar atamaktır. İkinci adımda, girişler (bağımsız değişkenler) sigmoid veya lojistik işlevini kullanarak ileri doğru yayılır, böylece her gizli katman için çıktı değerleri (bağımlı değişkenler) üretilir. Bundan sonra, üçüncü adımda ağırlıkları ve önyargıları güncelleyerek hata geriye doğru yayılır. Hatalar her çıktı ve gizli katman için hesaplanır. Son adımda, ağırlıklar ve önyargılar güncellenerek Adım 2'ye geri döndürülür. Genel hata en aza indirilene kadar adımlar tekrarlanır (Han, 2012). Yapay sinir ağları modelleri; ağın yapısına, ağdaki düğümlerin özelliklerine, kullanılmakta olan eşik değerine, ağın ileri ya da geri beslemeli olmasına, ağırlık matrisi değerlerinin sabit ya da değişken olmasına, ağırlık matrislerinin simetriklik veya asimetriklik durumuna ve eğitim ya da öğrenme kurallarına bağlıdır (Şen, 2004).

Şekil 2

Çok Katmanlı Algılayıcı Örneği (Öz, 2017)



Karar Ağacı Algoritmaları

KA algoritmaları, hem sınıflandırma hem de tahminde kullanılabilecek en yaygın veri madenciliği yöntemlerinden biridir. KA algoritmaları, kolay yorumlanabildiği için sınıflandırma tekniklerinde sıklıkla tercih edilmektedir. Genelleme hatasını minimum seviyeye düşürerek veri setinden bir karar ağacı oluşturulması amaçlanmaktadır. KA algoritmaları C4.5, RF, Raptree gibi farklı algoritmalar kullanılarak çözülmektedir (Maimon, 2005). Karar ağacı akış çizelgesi bir ağaç yapısına benzemektedir. Üretilen sınıflandırma kuralları bu ağaçtan kolayca belirlenebilmektedir (Han ve Kamber, 2001). Bu kurallar sayesinde programcıların da daha kolay program yazmaları sağlanır (Kartal, 2015).

Sınıflandırma Algoritmaları İçin Performans Ölçüm Kriterleri

Sınıflandırma algoritmalarının başarı oranının belirlenmesi sürecinde farklı performans kriterlerinin sonuçları hesaplanır ve karşılaştırılır. Hangi algoritmanın en etkili olduğunu belirlemek için birçok kriter kullanılmaktadır. Bu kriterler; doğru pozitif (DP) oranı, yanlış pozitif (YP) oranı, doğru negatif (DN) oranı, yanlış negatif (YN) oranı, hassasiyet (precision), f-ölçütü (f-measure), alıcı işlem karakteristiği eğrisi (Receiver Operating Characteristic [ROC]), kappa (κ) istatistiği, ortalama mutlak hata (Mean Absolute Error [MAE]), kök ortalama kareler hatası (Root Mean Square Error [RMSE]), Matthews korelasyon katsayısı (Matthews Correlation Coefficient [MCC]) olarak söylenebilir. Bu değerlerden bazılarının hesaplanabilmesi için karşılaştırma tablosunu bilmek gerekir. Karşılaştırma tablosunun bileşenleri Tablo 1'de verilmiştir (Öz, 2017).

Tablo 1

Karşılaştırma Tablosu

		Tahminlenen Değer	
		a	b
Gerçekleşen değer	a	DP	YN
	b	YP	DN

Tablo 1'de doğru pozitif (DP): doğru pozitif tahmin sayısını, yanlış pozitif (YP): yanlış pozitif tahmin sayısını, doğru negatif (DN): doğru negatif tahmin sayısını ve yanlış negatif (YN): yanlış negatif tahmin sayısını göstermektedir (Öz, 2017).

Doğru Pozitif Oranı (Duyarlılık)

Doğru sınıflandırılmış pozitif örneklerin, modeldeki toplam pozitif örnek sayısına oranıyla elde edilir ve Denklem 1'de gösterildiği biçimde ifade edilir (Filiz ve Öz, 2019).

$$DP\ Oranı = \frac{DP}{DP+YN} \quad (1)$$

Yanlış Pozitif Oranı

Gerçekte negatif olan fakat pozitif sınıflanmış olan örneklerin toplam negatif örnek sayısına oranı ile bulunur ve Denklem 2'de gösterildiği biçimde ifade edilir (Filiz, 2019).

$$\text{YP Oranı}=\frac{YP}{YP+DN} \tag{2}$$

Precision (Hassasiyet)

Hassasiyet değeri, doğru sınıflanmış pozitif örneklerin toplam pozitif örneklerin sayısına oranıdır. Denklem 3'te bu oran gösterilmiştir (Filiz, 2019).

$$\text{Hassasiyet}=\frac{DP}{DP+YP} \tag{3}$$

F-Measure (F-ölçütü)

Hassasiyet ve DP oranının harmonik ortalaması ile belirlenir. Denklem 4'te nasıl hesaplandığı görülmektedir (Filiz, 2019).

$$\text{F-ölçütü}=\frac{2.Hassasiyet.DP\ Oranı}{Hassasiyet+DP\ Oranı} \tag{4}$$

Alıcı İşlem Karakteristiği Eğrisi (ROC: Receiver Operating Characteristic)

ROC eğrisi genellikle sınıflandırma algoritmalarının performansını ölçmek için kullanılır, burada eğri altındaki alan sınıflandırıcının nasıl çalıştığını gösterir (Bradley, 1997) Bu eğri, Y ekseninde DP değerine ve X ekseninde (1-DN) değerine sahiptir. ROC eğrisinin altında kalan alanın değeri ne kadar yüksek olursa, algoritmanın yaptığı sınıflandırmanın da o kadar başarılı olduğu söylenebilir. Değerler sayısal olarak hesaplanabilir. ROC eğrisinin altındaki alan, algoritmaların performansı ile ilgili görsel sonuçların yanı sıra sayısal sonuçlar da sunmaktadır. Böylece, farklı algoritmaların performanslarının karşılaştırılması görsel olarak kolayca yorumlanabilir (Filiz ve Öz, 2019).

Kappa (κ) İstatistiği

κ istatistiği, bir sınıflandırma algoritmasının performans başarısı ile ilgilenir. Kategorik değişkenler için yapılan analizlerin anlaşılmasında kullanılan uygun bir istatistiki veridir. Aynı zamanda ki-kare tablosuna dayanan bir değerdir (Klar, 1996). κ istatistiği 1'e yaklaştıkça, sınıflandırıcılar arasındaki uyuşma daha yüksektir. p_o ve p_e'nin iki kategorik değişken arasındaki gözlemlenen ve beklenen değerleri göstermek üzere κ istatistiği Denklem 5 ile hesaplanır (Öz, 2017).

$$\kappa = \frac{p_o+p_e}{1-p_e} \tag{5}$$

p_o ve p_e değerleri de Denklem 6 ve Denklem 7 ile hesaplanır.

$$p_o=\frac{DP+DN}{DP+YN+YP+DN} \tag{6}$$

$$p_e=\frac{(DP+YP).(DP+YN)+(YN+DN).(YP+DN)}{(DP+YN+YP+DN)^2} \quad (7)$$

Ortalama Mutlak Hata (MAE: Mean Absolute Error)

Ortalama mutlak hata istatistiği, bir modelin tahmin edilen ve gözlenen değerleri arasındaki farkları göstermeye yardımcı olur (Matsuura, 2005). Bu değer, tahmin edilen ve gözlemlenen değerler arasındaki mutlak farkların ortalamasını hesaplar ve Denklem 8'de gösterildiği biçimde ifade edilir (Öz, 2017).

$$\text{MAE}=n^{-1}\sum_{i=1}^{n}|P_i-O_i| \quad (8)$$

P_i-O_i modelin tahmin hatasını gösterir.

Kök Ortalama Kare Hata (RMSE: Root Mean Square Error)

MAE'ye benzer şekilde bir modelin tahmin edilen ve gözlenen değerleri arasındaki farkları göstermeye yardımcı olur. Kök ortalama kare hatası, tahmin edilen ve gözlemlenen değerler arasındaki kare farklarının ortalama karekköüne eşittir ve Denklem 9'da gösterildiği biçimde ifade edilir.

$$\text{RMSE}=\sqrt{n^{-1}\sum_{i=1}^{n}|P_i-O_i|^2} \quad (9)$$

$Pi-Oi$ modelin tahmin hatasını gösterir.

Matthews Korelasyon Katsayısı (MCC: Matthews Correlation Coefficient)

Matthews korelasyon katsayısı (MCC) -1 ile 1 arasında değerler alır ve karşılaştırma matrisindeki bileşenler kullanılarak elde edilir. Pozitif değer elde eden MCC'lerin doğru tahminler ürettiği sonucuna varılabilir. MCC = 1 değeri mükemmel tahminler yağıldığı anlamına gelir. MCC, Denklem 10'da verildiği şekilde hesaplanır (S. Kılıç-Depren, 2017).

$$\text{MCC}=\frac{(DP.DN)-(YP.YN)}{\sqrt{(DP+YP).(DP+YN).(DN+YP).(DN+YN)}} \quad (10)$$

k-katlı Çapraz Doğrulama

Makine öğrenmesi yöntemlerini uygularken kritik noktalardan biri eğitim ve test verilerinin (kümelerinin) belirlenmesidir. Veri seti temel olarak eğitim ve test veri seti olarak iki gruba ayrılır. Bölünme, % 50 - % 50 ya da % 60 - % 40 gibi değerlerle belirlenebilir. k-katlı çapraz doğrulama yöntemi

veri setini k eşit parçaya böler. Bunların k-1 tanesi eğitim veri seti, kalan kısmı ise test verisidir. Her bir parça, test kümesi olarak alınarak işlem k kez tekrar edilir. Tüm sonuçların ortalaması hesaplandığında, sınıflandırma değerleri belirlenmiş olur (Öz, 2017).

Literatürde k değerinin 2, 5, 10 olarak kullanıldığı görülmektedir (Erpolat ve Öz, 2010). Breiman ve diğ. ()k = 10 seçmenin en iyi seçim olduğunu söylemişlerdir (Breiman, 1984). Bu çalışmada da, k değeri için sıklıkla kullanılan 10 değeri alınmıştır. Matematik başarısı veri seti için k=10 alınarak 4077 öğrencinin verileri ile 10-katlı çapraz doğrulama yapılmıştır. 10 parçanın 9 tanesi eğitim seti, 1 tanesi test grubu olarak seçilmiştir. 408 öğrencinin verisi eğitim setinde, 3999 öğrencinin verisi de test grubunda yer almaktadır. Bu işlem her defasında test grubunu değiştirmek suretiyle 10 kez tekrar edilir.

Yöntem

Bu çalışmada 2019-TIMMS veri setindeki Türkiye örnekleminden elde edilen veri seti kullanılmıştır (https://timss2019.org/international-database/., 2019). 2019-TIMSS veri setinde Türkiye'deki sekizinci sınıf öğrencilerinin matematik başarıları sınıflandırılmış ve bu sınıflandırma başarılarının bulunması için makine öğrenmesi algoritmalarından yararlanılmıştır. Bu işlemler için aşağıdaki adımlar izlenmiştir:

- TIMSS-2019 Türkiye 8. sınıf öğrencilerinin sonuçları odak grubu olarak ele alınmıştır. Veri setindeki mevcut olmayan ve eksik gözlemler yerine o setteki verilerin ortalaması atanmıştır. Bunun için SPSS programında eksik veriler için anlamlılık testi yapılmış ve $p > .05$ olduğu için eksik veri yerine değer atama işleminin gerçekleştirilebileceği anlaşılmıştır. Daha sonra verilerin dağılımı normal olduğu için eksik veriler yerine o serinin ortalamasının atanmasına karar verilmiştir. Daha sonra veriler ARFF dosya türüne dönüştürülmüştür.
- Eğitim ve test veri kümeleri, 10-katlı çapraz doğrulama kullanılarak elde edilmiştir.
- En iyi performans gösteren algoritma bulunmuştur.
- Algoritmaların sınıflandırma başarıları incelenmiştir.
- 8. Sınıf matematik başarısının algoritmalara göre sınıflandırılması ve algoritmaların karşılaştırma matrisleri Tablo 6 ve Tablo 7'de gösterilmiştir.

WEKA Programı

Bu çalışmada kullanılacak verinin bağımlı (Matematik Başarısı) ve bağımsız değişkenlerin (Öğrenci Matematik Tutum Anketi) değerleri arff dosyası olarak hazırlanmış ve analizleri yapmak amacıyla üzere WEKA programına aktarılmıştır. WEKA, veri madenciliğinde kullanılan makine öğrenmesi algoritmalarını ve öznitelik seçim algoritmalarını bünyesinde barındıran Java programlama dili tabanlı bir yazılımdır. Waikato üniversitesi tarafından geliştirilmiş ve Waikato Environment for Knowledge Analysis kelimelerinin baş harflerinden programın adı oluşturulmuştur (Frank, 2016).

2019-TIMSS çalışmasının sekizinci sınıftaki Türk öğrencilerinin matematik başarıları ele alınmıştır. Türkiye 8. Sınıf örnekleminde toplam 4077 öğrenci bulunmaktadır. Toplamda 4 değişken kullanılmış ve bağımsız değişkenler Tablo 2, Tablo 3 ve Tablo 4' te gösterilmiştir. Bu değişkenlerden "Matematik birinci değerlendirme sonucu" (BSMMAT01) değişkeni öğrencilerin matematik başarısını göstermektedir ve çalışmada bağımlı değişken olarak seçilmiştir. BSMMAT01 ortalama matematik puanları önce SPSS programında z-puanlarına dönüştürülmüştür. Daha sonra -1 den küçük olan z-puanları için "low", -1 ile 1 arasındaki z-puanları için "intermediate" ve 1 den büyük z-puanları için "high" olarak değer atanmış ve sınıflama etiketleri belirlenmiştir. TIMMS tarafından Türkiye'deki 8. Sınıf öğrencilerinin ortalama matematik başarı puanı 497 olarak belirlenmiştir. Öğrenci anketindeki Matematiğe karşı tutum ölçeğindeki Matematikte öğrencinin özgüveni, (BSBGSCM), Matematik öğrenmeyi sevmesi (BSBGSLM), Öğrencinin Matematiğe verdiği değer (BSBGSVM) değişkenleri ise matematik başarısını etkilediği düşünülen bağımsız değişkenler olarak seçilmiştir.

Matematik veri setinde öğrencinin matematik öğrenmeyi sevmesi (BSBGSLM) ölçeğindeki sorular ve cevap seçenekleri aşağıdaki Tablo 2'de verilmiştir.

Tablo 2

Matematik Veri Seti İçin Öğrencinin Matematiği Öğrenmeyi Sevmesi (BSBGSLM) Ölçeğinin İçeriği

Soru Kodu	Soru	Cevap
BSBM16A	Matematik öğrenirken eğlenirim.	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM16B	Keşke matematik çalışmak zorunda kalmasam.	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM16C	Matematik sıkıcıdır.	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM16D	Matematikte birçok ilginç şey öğrenirim.	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM16E	Matematiği severim.	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM16F	Rakam içeren herhangi bir okul çalışmasını severim.	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum

(devam ediyor)

Tablo 2 (devam)

Soru Kodu	Soru	Cevap
BSBM16G	Matematik problemlerini çözmeyi severim .	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM16H	Okulda matematik öğrenmeyi heyecanla bekliyorum.	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM16I	Matematik en sevdiğim derslerden biridir.	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum

Matematik veri setinde öğrencinin matematikteki özgüveni (BSBGSCM) ölçeğindeki sorular ve cevap seçenekleri aşağıdaki Tablo 3'te verilmiştir.

Tablo 3

Matematik Veri Seti İçin Öğrencinin Matematikteki Özgüveni (BSBGSCM) Ölçeğinin İçeriği

Soru Kodu	Soru	Cevap
BSBM19A	Matematikte genellikle başarılıyımdır.	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM19B	Birçok sınıf arkadaşıma göre matematikte daha çok zorlanırım.	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM19C	Matematik güçlü yanlarımdan biri değildir.	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM19D	Matematikteki şeyleri kolaylıkla öğrenirim.	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM19E	Matematik beni tedirgin eder.	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM19F	Matematik zor problemlerini çözmekte başarılıyım.	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum

(devam ediyor)

Tablo 3 (devam)

Soru Kodu	Soru	Cevap
BSBM19G	Öğretmenim matematikte iyi olduğumu söyler	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM19H	Matematik benim için diğer konulardan daha zordur	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM19I	Matematik beni şaşırtır	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum

Matematik veri setinde öğrencinin matematiğe verdiği değer (BSBGSVM) ölçeğindeki sorular ve cevap seçenekleri aşağıdaki Tablo 4'te verilmiştir.

Tablo 4

Matematik Veri Seti İçin Öğrencinin Matematiğe Verdiği Değer (BSBGSVM) Ölçeğinin İçeriği

Soru Kodu	Soru	Cevap
BSBM20A	Matematik öğrenmek günlük hayatta bana yardımcı olacaktır	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM20B	Diğer okul konularını öğrenmek için matematiğe ihtiyaç vardır	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM20C	Gitmek istediğim üniversite için matematikte başarılı olmak gerekir	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM20D	İstediğim iş için matematikte başarılı olmak gerekir	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM20E	Matematik kullanmayı içeren bir iş istiyorum	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM20F	Dünyada ilerlemek için matematiği öğrenmek önemlidir	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM20G	Yetişkin olduğumda matematik bana daha fazla iş imkânı sağlayacak	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum

(devam ediyor)

Tablo 4 (devam)

Soru Kodu	Soru	Cevap
BSBM20H	Ailem matematikte iyi düzeyde olmamın önemli olduğunu düşünür.	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum
BSBM20I	Matematikte başarılı olmak önemlidir.	1) Kesinlikle katılıyorum 2) Katılıyorum 3) Katılmıyorum 4) Kesinlikle katılmıyorum

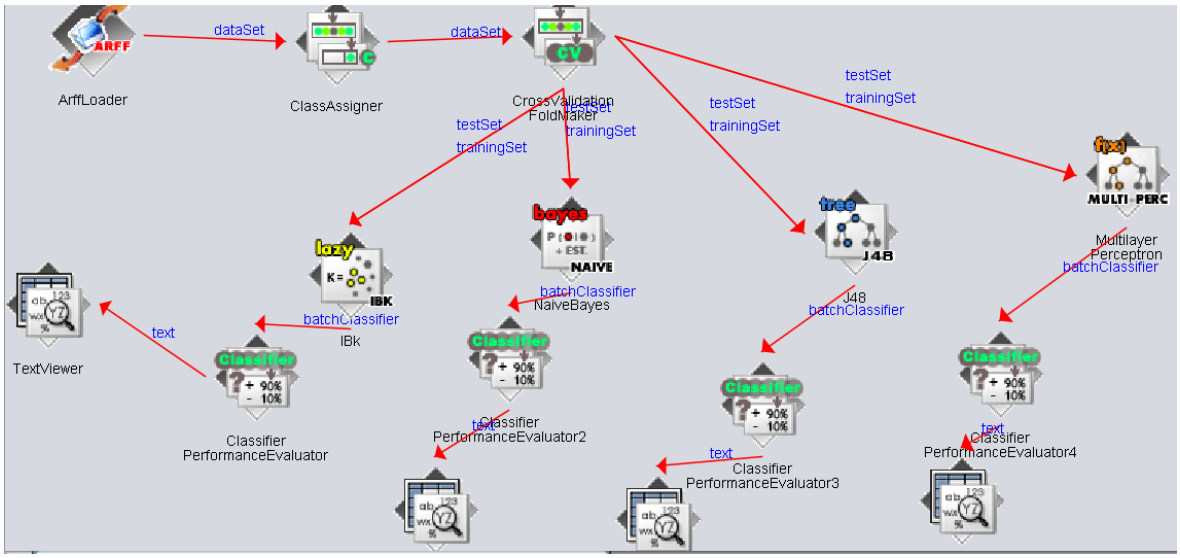
*(Appendix16B, 2019)

Sonuçlar

WEKA programının Knowledge Flow arayüzü kullanılarak aşağıdaki model oluşturulmuş ve analiz sonuçları elde edilmiştir.

Şekil 3

Sınıflama Analizi için WEKA Programında Oluşturulan Model



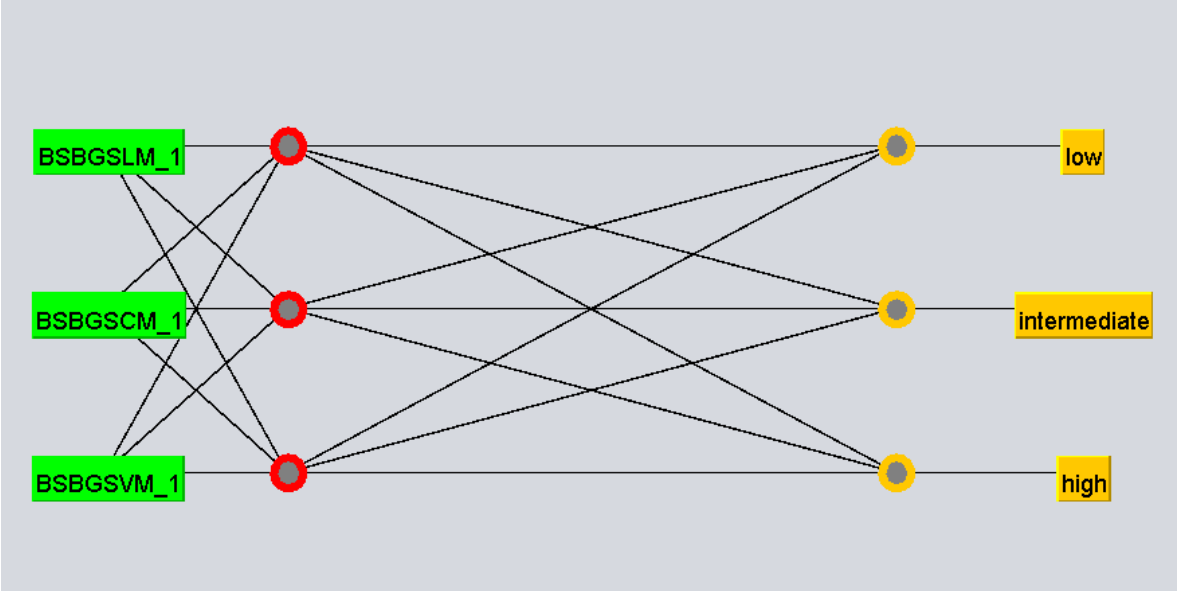
Yukarıdaki modelde K-NN algoritmasında en yakın komşuluk değeri olarak 1, 3, 5, 7, 10, 50, 80 ve 100 değerleri denenmiş ve sırasıyla %57, %60, %64, %67, %67, %70, %69 ve %69 doğru sınıflama yüzdeleri elde edilmiştir. Buna göre en iyi sınıflama yüzdesini veren en yakın 50 komşuluk için elde edilen sonuçlar daha iyi gibi görünse de sınıflama matrislerine bakıldığında en yakın 7 komşuluk için elde edilen matristeki doğru sınıflama veri sayısı daha yüksek olduğundan 7 komşuluk için elde edilen sonuçlar karşılaştırma tablosuna alınmıştır.

Yapay Sinir Ağları algoritması için yapılan analizde ise önce 3 adet gizli katman için analiz çalıştırılmış ve doğru sınıflama yüzdesi %17 bulunmuştur. Gizli katman olarak 5-15 değerleri

seçildiğinde ise bu oran yüzde 52 ye yükselmiştir. Elde edilen katmanlı sinir ağları aşağıda Şekil 4 ve Şekil 5'te gösterilmiştir. Karşılaştırma tablosuna 5-15 gizli katmanlı modelin sınıflama yüzdesi daha yüksek olduğu için bu modelin sonuçları karşılaştırma tablosuna dahil edilmiştir.

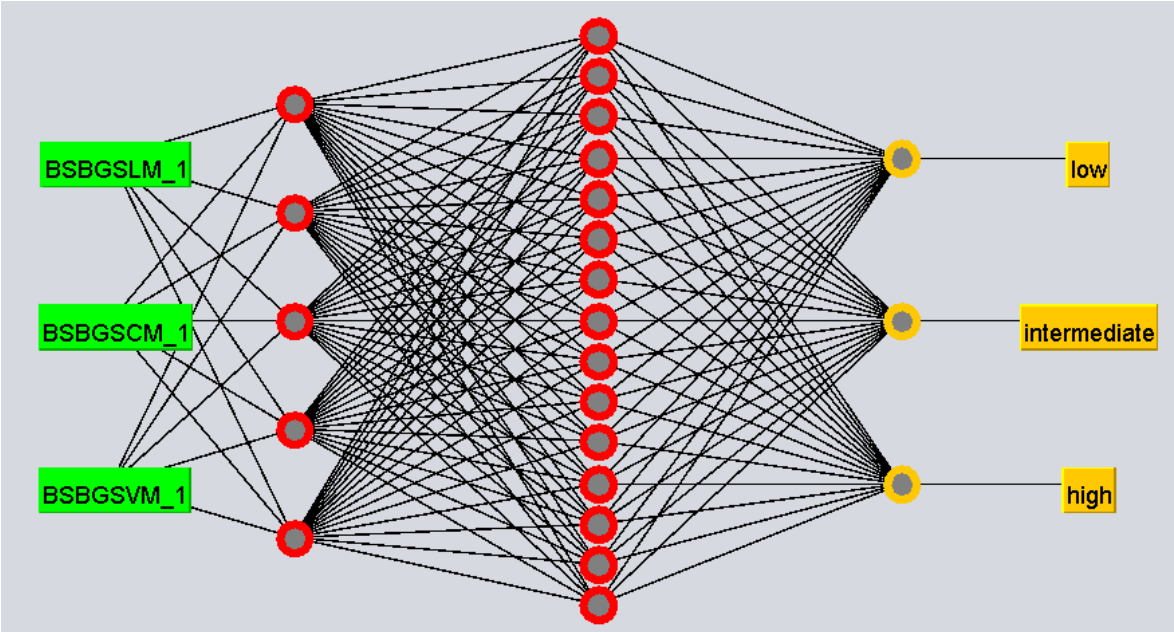
Şekil 4

3 Gizli Katman İçeren Yapay Sinir Ağı



Şekil 5

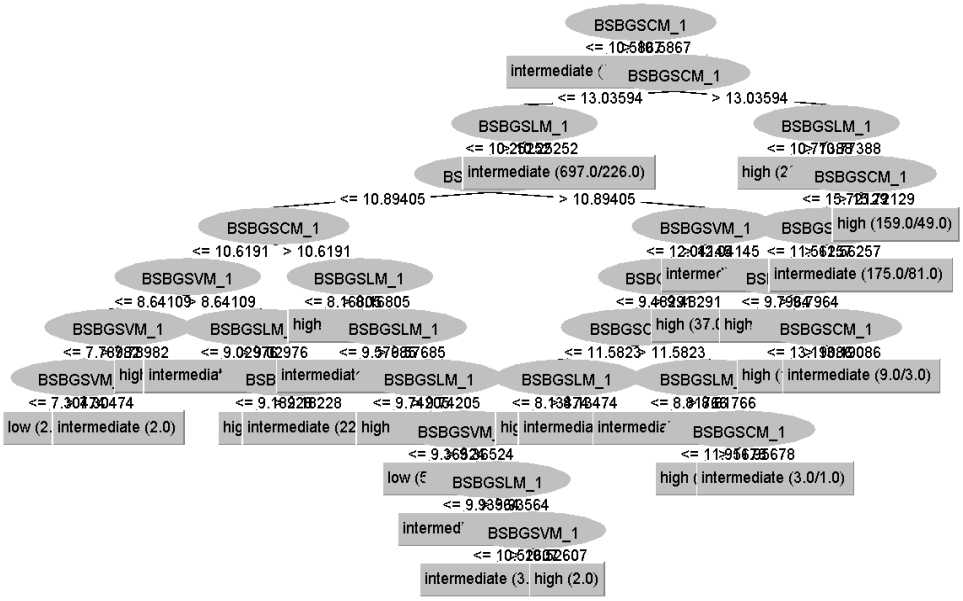
5-15 Gizli katman İçeren Yapay Sinir Ağı



Karar Ağacı (KA-J48) Algoritmasına göre elde edilen karar ağacı Şekil 6'da gösterilmiştir.

Şekil 6

KA-J48 Algoritması Sonucunda Elde Edilen Karar Ağacı Modeli



Veri seti kullanılarak elde edilen karşılaştırma matrisi değerleri Tablo 5'te ve algoritmaların sınıflandırma kriterlerine göre sınıflandırma başarıları Tablo 6'da verilmiştir.

Tablo 5

K-NN, NB, KA-J48 VE YSA Algoritmaları için Karşılaştırma Matrisi Değerleri

	K-NN			NB			KA-J48			YSA		
	a	b	c	a	b	c	a	b	c	a	b	c
a=low	30	175	6	13	639	10	0	658	4	198	462	2
b=intermediate	60	820	56	10	2462	278	2	2642	106	825	1907	18
c=high	7	157	75	0	389	276	1	477	187	198	442	25

Tablo 5'te verilen karşılaştırma matrislerine göre K-NN algoritması öğrenci verisinin 925 (30+820+75) tanesini, NB algoritması 2751(13+2462+276) tanesini, KA-J48 algoritması 2829 (0+2642+187) tanesini ve YSA algoritması ise 2130 (198+1907+25) tanesini doğru sınıflamıştır.

Tablo 6

K-NN, NB, KA-J48 VE YSA Algoritmaları için Sınıflandırma Sonuçları

	DP Oranı	YP Oranı	Precision	F-Ölçütü	κ Statistic	MAE	RMSE	ROC Area	MCC
K-NN	.67	.52	.62	.63	.19	.29	.40	.64	.19
NB	.68	.54	.65	.61	.19	.30	.40	.63	.18
KA-J48	.69	.58	.58	.61	.16	.29	.39	.62	.18
YSA	.52	.51	.58	.51	.01	.43	.46	.49	.03

Tablo 6'daki ROC alanı değerleri tüm algoritmalar için incelendiğinde YSA algoritması hariç diğer algoritmaların ROC alanı değerlerinin 0.62 ve bu değerden daha büyük olduğu görülmektedir. Bu durum algoritmaların iyi sınıflandırma yaptığını gösterir. Algoritmalar karşılaştırıldığında ise en iyi sınıflandırmayı %69 luk bir DP oranı ile KA-J48 Algoritmasının yaptığı görülmektedir. KA-J48 algoritması öğrenci verisinin toplamda 2829 tanesini doğru sınıflamıştır. İkinci en iyi sınıflamayı NB algoritması (DP= 0.68) yapmıştır. K-NN algoritması %67 lük doğru sınıflama oranına sahip olup en yakın 7 komşuluk modeline göre en iyi sınıflama yapan üçüncü algoritma olmuştur.

Sonuç ve Tartışma

Hızla küreselleşen dünyada eğitimi her yönüyle bir bütün olarak ele alınıp değerlendirmek ve eğitimde aksayan durumları tespit etmek ve iyileştirmek öncelikli hedeflerden biridir. Uluslararası ve ulusal çapta yapılan sınavlarda matematik başarısının değerlendirilmesi çalışmalarında başarı ortalamasının diğer derslere göre daha düşük olduğu görülmektedir. Matematikteki bu başarı düşüklüğünün nedenleri arasında öğrencinin matematiğe karşı tutumu, öğretmen, okul ve aile etkileşiminin rolü olduğu bilinmektedir. 2019-TIMMS sekizinci sınıf Türkiye matematik başarı ortalamasının 497 puanla TIMMS ortalaması olan 500 puanın altında kaldığı görülmektedir. Bu araştırmada 2019-TIMMS Türkiye sekizinci sınıf matematik başarı puanlarının öğrencinin matematiğe karşı tutum ölçeği puanlarına göre "düşük", "orta", "yüksek" olarak sınıflandırılmasında veri madenciliği algoritmalarından hangisinin daha başarılı olduğu araştırılmıştır. Bu sınıflandırma sadece öğrenci tutumu ölçeği kullanılmıştır. Daha geniş çaplı bir araştırma olarak okul, aile ve öğretmen anketlerinden elde edilen verilere göre algoritmaların sınıflandırma başarılarının sınanması önerilebilir. Ayrıca bu çalışmada öğrenci başarıları "düşük", "orta", "yüksek" olarak üç kategoride etiketlenmiştir. Başarı durumu iki kategoride "başarılı" ve "başarısız" olarak etiketlendiğinde algoritmaların sınıflama başarılarının daha yüksek olacağı düşünülmektedir.

Çalışma kapsamında 2019-TIMMS Türkiye örnekleminden elde edilen 8. Sınıfların matematik başarı puanları (BSMMAT01) bağımlı değişken olarak ele alınmıştır. BSMMAT01 ortalama matematik puanları önce SPSS programında z-puanlarına dönüştürülmüştür. Daha sonra -1 den küçük olan z-puanları için "low", -1 ile 1 arasındaki z-puanları için "intermediate" ve 1 den büyük z-puanları için "high" olarak değer atanmış ve sınıflama etiketleri belirlenmiştir. Öğrenci anketindeki Matematiğe karşı tutum ölçeğindeki Matematikte öğrencinin özgüveni, (BSBGSCM), Matematik öğrenmeyi sevmesi (BSBGSLM), Öğrencinin Matematiğe verdiği değer (BSBGSVM) değişkenleri ise matematik başarısını etkilediği düşünülen bağımsız değişkenler olarak seçilmiştir.

Matematik başarısını sınıflamada karşılaştırılacak algoritmalardan ise K en yakın komşu (K-NN), Karar Ağacı (KA-J48), Yapay Sinir Ağları (YSA) ve Naive Bayes (NB) algoritmaları kullanılmıştır. Yapılan analizler sonucunda elde edilen karşılaştırma matrislerine göre K-NN algoritması öğrenci verisinin 925 tanesini, NB algoritması 2751 tanesini, KA-J48 algoritması 2829 tanesini ve YSA algoritması ise 2130 tanesini doğru sınıflamıştır. ROC alanı değerleri tüm algoritmalar için

incelendiğinde YSA algoritması hariç (0.49) diğer algoritmaların ROC alanı değerlerinin 0,62 ve bu değerden daha büyük olduğu görülmektedir. Bu durum algoritmaların iyi sınıflandırma yaptığını gösterir. Algoritmalar karşılaştırıldığında ise en iyi sınıflandırmayı %69 luk bir DP oranı ile KA-J48 Algoritmasının yaptığı görülmektedir. KA-J48 algoritması öğrenci verisinin toplamda 2829 tanesini doğru sınıflamıştır. İkinci en iyi sınıflamayı NB algoritması (DP= 0.68) yapmıştır. K-NN algoritması %67 lük doğru sınıflama oranına sahip olup en yakın 7 komşuluk modeline göre en iyi sınıflama yapan üçüncü algoritma olmuştur. Sınıflama başarısındaki oranların gruplama etiketleri 2 kategoriye indiğinde daha da yükselebileceği düşünülmektedir.

Kaynaklar

Amasyalı, M. F., Diri, B. ve Türkoğlu, F. (2006, Haziran). *Farklı özellik vektörleri ile Türkçe dökümanların yazarlarının belirlenmesi.* Türkiye Yapay Sinir Ağları Sempozyumu (TAINN), Muğla, Türkiye. http://www.kemik.yildiz.edu.tr/data/File/publications/Author%20Detection/Farkli%20Ozellik%20Vektorleri%20ile%20Turkce%20Dokumanlarin%20Yazarlarinin%20Belirlenmesi.pdf

Balaban, M. E. ve Kartal, E. (2015). *Veri madenciliği ve makine öğrenmesi: Temel algoritmaları ve R dili uygulamaları.* Çağlayan Kitabevi.

Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145-1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Daş, B. ve Türkoğlu, İ. (2014, 25-27 Eylül). *DNA dizilimlerindeki nükleotit çiftlerinin frekans değerlerine göre farklı sınıflandırma yöntemleri ile karşılaştırılması.* Tıp Teknolojileri Ulusal Kongresi (TIPTEKNO'2014), 191-194. http://www.biyoklinikder.org/tiptekno14/TIPTEKNO_14_bildiriler/DNA%20Dizilimlerindeki%20N%C3%BCkleotit%20%C3%87iftlerinin%20Frekans%20De%C4%9Ferlerine%20G%C3%B6re%20Farkl%C4%B1%20S%C4%B1n%C4%B1fland%C4%B1rma%20Y%C3%B6ntemleri%20ile%20Kar%C5%9F%C4%B1la%C5%9Ft%C4%B1r%C4%B1lmas%C4%B1.pdf

Erpolat, S. ve Öz, E. (2010). *Kanser verilerinin sınıflandırılmasında yapay sinir ağları ile destek vektör makinelerinin karşılaştırılması.* İstanbul Aydın Üniversitesi Fen Bilimleri Dergisi, 2(5), 71-83.

Filiz, E. (2019). *Makine öğrenmesi yöntemleri ve eğitim verisi üzerine bir uygulama: uluslararası matematik ve fen eğilimleri araştırması 2015 Türkiye örneği* (Tez No. 598315) [Doktora tezi, Yıldız Teknik Üniversitesi]. Yükseköğretim Kurumu Tez Merkezi.

Filiz, E. ve Öz, E. (2017). Classification of BIST-100 Index' Changes via machine learning methods. *Marmara Üniversitesi İktisadi ve İdari Bilimler Dergisi, 39*(1), 117-129. https://doi.org/10.14780/muiibd.329913

Filiz, E. ve Öz, E. (2019). Finding the best algorithms and effective factors in classification of turkish science student success. *Journal of Baltic Science Education, 18*(2), 239-253. https://doi.org/10.33225/jbse/19.18.239

Fishbein, B., Foy, P., and Yin, L. (2019). *TIMSS 2019 user guide for the ınternational* database (2nd ed.). IEA. https://timss2019.org/international-database/downloads/TIMSS-2019-User-Guide-for-the-International-Database-2nd-Ed.pdf

Frank, E., Hall, M. A., and Witten, I. H. (2016). *The WEKA workbench: Data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann. https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf

Han, J., Kamber, M., and Pei, J. (2012). *Data mining: Concept and techniques* (3rd ed.). Morgan Kaufmann. http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

Jiang, S., Pang, G., and Wu, M., and Kuang, L. (2012). An improved K-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications, 39*(1), 1503-1509. https://doi.org/10.1016/j.eswa.2011.08.040

John, G. H. ve P. Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. *archivePrefix,* 338-345. https://arxiv.org/ftp/arxiv/papers/1302/1302.4964.pdf

Kılıç-Depren, S., Aşkın, Ö. E., and Öz, E. (2017). Identifying the classification performances of educational data mining methods: A case study for TIMSS. *Educational Sciences: Theory & Practice, 17*(5), 1605-1623. https://doi.org/10.12738/estp.2017.5.0634

Klar, A. D. (1996). The statistical analysis of kappa statistics in multiple samples. *Journal of Clinical Epidemiology, 49*(9), 1053-1058. https://doi.org/10.1016/0895-4356(96)00057-1

L. Breiman, J. F. (1984). *Classification algorithms and regression trees.* Chapman and Hall.

Maimon, O., and Rokach, L. (2005). Decision trees. In L. Rokach, and O. Maimon (Eds.), *Data mining and knowledge discovery handbook* (pp. 165-192). Springer.

MEB. (2021). *Ortaöğretim kurumlarına ilişkin merkezi sınav* (Eğitim Analiz ve Değerlendirme Raporları Serisi, No. 16). MEB. https://cdn.eba.gov.tr/icerik/2021/06/rapor/2021_Ortaogretim_Kurumlarina_Iliskin_Merkezi_Sinav.pdf

Nizam, H. ve Akın, S. S. (2014, 27-29 Kasım). *Sosyal medyada makine öğrenmesi ile duygu analizinde dengeli ve dengesiz veri setlerinin performanslarının karşılaştırılması.* XIX. Türkiye'de İnternet Konferansı. http://inet-tr.org.tr/inetconf19/bildiri/10.pdf

Öz, E., Kurt, S., Asyali, M. H., Kaya, H. ve Yücel, Y. (2016). Feature based quality assessment of DNA sequencing chromatograms. Applied Soft Computing, 41, 420–427.

Savaş, E., Tat, S. ve Duru, A. (2010). Factors affecting students' achievement in mathematics. *İnönü University Journal of the Faculty of Education (INUJFE), 11*(1), 113-115.

Şeker, Ş. E. (2021). https://bilgisayarkavramlari.com/2013/03/31/siniflandirma-classification/.

Şen, Z. (2004). *Yapay sinir ağları ilkeleri.* Su Vakfı Yayınları.

TIMSS (2019). *Creating and interpreting the TIMSS 2019 context questionnaire scales.* TIMSS & PIRLS International Study Center and IEA.

https://timssandpirls.bc.edu/timss2019/methods/pdf/T19_MP_Ch16-context-questionnaire-scales.pdf

Willmott, C. J., and Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research, 30*(1), 79-82. http://www.jstor.org/stable/2486923

Zhang, H. (2004, May 17-19). *The optimality of naive Bayes.* 17th FLAIRS Conference, Florida, USA. https://www.aaai.org/Papers/FLAIRS/2004/Flairs04-097.pdf

Çoktan seçmeli testlerde kısmi puanlama sağlayan sınav sisteminin geliştirilmesi[1]

Ufuk Akbaş, Şeyhmus Aydoğdu, Merve Yıldırım Seheryeli ve Şener Büyüköztürk

Anahtar kelimeler: Çoktan seçmeli test, kısmi puanlama, kısmi bilgi, öğrenme eksikleri

Giriş

Hızlı ve nesnel bir şekilde puanlanabilmesi, kısa sürede cevaplanabildiği için çok sayıda sorunun sorulabilmesi ve geniş bir kapsamın yoklanmasını sağlaması gibi avantajları açısından çoktan seçmeli testler, pek çok alanda yaygın bir şekilde kullanılmakta ve genellikle 1-0 şeklinde puanlanmaktadır. Şans başarısı ve kısmi bilginin ölçülememesi çoktan seçmeli test maddelerin temel sınırlıkları arasındadır (Kaplan ve Atalmış, 2019). Bu noktada, çoktan seçmeli testlerde kısmi puanlama yöntemlerinin kullanımı bu sınırlılıklara çözüm sağlayabilir.

Test maddesiyle ölçülen davranışa sahip olmayan ya da kısmen sahip olan bireylerin doğru yanıtı tahmin ederek bulmaları testin geçerliğini ve güvenirliğini olumsuz etkilemektedir (Baykul, 2000). Şans başarısının miktarı seçenek sayısına göre değişir ve testin 1-0 şeklinde puanlanması halinde gerçek puanı daha düşük olan bireyler için şansa bağlı kazanılabilecek puanın miktarı, gerçek puanı yüksek olanlara göre daha yüksektir (Tan, 2004). Şans başarısını engellemek amacıyla düzeltme formülünün uygulanması, boş bırakılan sorulara belli bir puanın verilmesi gibi yöntemler önerilmiştir (Gulliksen, 1950; Abu-Sayf, 1979). Lord'a (1975) göre düzeltme formülünün, testi cevaplayan tüm bireylerin testteki tüm sorularda hiçbir seçeneği elemeden rastgele seçim yapması gibi savunulamayacak bir varsayımı vardır. Bireyler, ölçülen özellikle ilgili kısmi de olsa bilgiye sahiptir ve çoğu çoktan seçmeli test maddesinde bir ya da daha fazla seçenek bu kısmi bilgiye dayalı olarak elenebilmektedir.

Farklı öğrenim düzeylerinden öğrencilerin yanıtlama davranışlarını inceleyen araştırmalarda da seçenek elemenin yaygın (Dirkzwager, 1996; Umay, 1998; Koçak, 2013) ve hatta en az bir seçenek elemenin, sadece doğru olduğu düşünülen cevabı işaretlemek kadar sık sergilenen bir davranış olduğu ortaya koyulmaktadır (Pehlivan-Tunç ve Kutlu, 2004).

[1] TÜBİTAK-SOBAG tarafından desteklenen "Çoktan Seçmeli Testlerde Test Güvenliğini Arttıran ve Kısmi Puanlama Sağlayan Sınav Sisteminin Geliştirilmesi" başlıklı projeden (120K294) üretilmiştir.

Bir sorudaki çeldiricilerin elenebilmesi, kısmi bilginin varlığına işaret etmektedir. Literatürde, 1-0 puanlamanın duyarsız olması nedeniyle kısmi puanlama yöntemi öneren ve bu yöntemleri karşılaştıran çok sayıda araştırmaya rastlanmaktadır (Coombs ve diğ., 1956; Arnold ve Arnold, 1970; Kansup ve Hakstian, 1975; Waters, 1976; Cross ve Frary, 1977; Çetin, 2005; Gözen-Çıtak 2010; Yurdugül, 2010; Bauer ve diğ., 2011; Wu ve diğ., 2019; Selvi ve Derici-Yıldırım, 2020).

Pek çok araştırmada, farklı puanlama yöntemlerinin puanların psikometrik özellikleri üzerindeki etkilerinin incelendiği görülmekle birlikte; Hoe ve diğ., (2009), Lau ve diğ., (2011) ve Wu ve diğ. (2019) tarafından yürütülen çalışmalarda, öğrencilerin durumunun ayrıntılı bir şekilde ortaya konabildiği ve ulaşılan sonuçlardan izleme-biçimlendirme amacıyla yararlanılabileceği görülmektedir. Temelinde eleme puanlamasının yer aldığı bu araştırmalarda puanlama yöntemlerinde bazı farklılıklar söz konusu olmakla birlikte tam bilgi, farklı düzeylerdeki kısmi bilgi, farklı düzeylerdeki kısmi kavram yanılgısı ve tam kavram yanılgısı arasında ayrım yapılabilmektedir.

Elemeye dayalı doğru cevap (EDDC) yöntemi (*Number Right Elimination Testing-NRET*), 1-0 şeklindeki puanlama ile eleme puanlamasını birleştirmektedir. Cevaplayıcıdan, her bir seçenek için "doğru", "yanlış" ya da "emin değilim" şeklinde görüş bildirmesi istenmektedir. Hoe ve diğ. (2009) tarafından yürütülen araştırmada, öğrencilerden maddedeki herhangi bir seçeneği eleyemeseler bile mutlaka bir doğru cevap belirtmeleri istenmiştir. Kısmi bilginin veya kavram yanılgılarının tespitinde bu yaklaşımın hatalı olduğu; öğrencilerden, doğru cevabı bulamıyorlarsa bile eleyebildikleri seçenekleri belirtmelerini istemenin daha isabetli olacağı düşünülmektedir. Bu bilgiler doğrultusunda bilgi düzeyleri arasındaki farkları daha kapsamlı bir şekilde ortaya koyabileceği düşünülen ve önerilen puanlama yöntemini içeren bilgiler Tablo 1'de sunulmuştur.

Tablo 1

EDDC Yöntemiyle Elde Edilecek Puanlar ve Öğrencilerin Bilgi Düzeyi

Cevap örüntüsü	1-0	Eleme puanı	EDDC puanı	Bilgi düzeyi
Doğru cevabın doğru, üç çeldiricinin yanlış olduğunu belirtme	1	3	4	Tam bilgi
Üç çeldiricinin yanlış olduğunu belirtme	1	3	4	
Doğru cevabın doğru, iki çeldiricinin yanlış olduğunu belirtme	1	2	3	Kısmi bilgi
Doğru cevabın doğru, bir çeldiricinin yanlış olduğunu belirtme	1	1	2	
İki çeldiricinin yanlış olduğunu belirtme	0	2	2	
Sadece doğru cevabın doğru olduğunu belirtme	1	0	1	
Sadece bir çeldiricnin yanlış olduğunu belirtme	0	1	1	
Doğru cevabın ve iki çeldiricnin yanlış olduğunu belirtme	0	1	1	Kısmi bilgi / kısmi kavram yanılgısı
Doğru cevabın ve bir çeldiricnin yanlış olduğunu belirtme	0	0	0	

(devam ediyor)

Tablo 1 (devam)

Cevap örüntüsü	1-0	Eleme puanı	EDDC puanı	Bilgi düzeyi
Sadece doğru cevabın yanlış olduğunu belirtme	0	-1	-1	Kısmi kavram yanılgısı
Çeldiricilerden birinin doğru, diğer çeldiricilerin ve doğru cevabın yanlış olduğunu belirtme	0	-2	-2	
Doğru cevabın ve bir çeldiricinin yanlış, bir çeldiricnin doğru olduğunu belirtme	0	-3	-3	
Sadece bir çeldiricinin doğru olduğunu belirtme	0	-3	-3	
Doğru cevabın yanlış ve bir çeldiricinin doğru olduğunu belirtme	0	-4	-4	Tam kavram yanılgısı

Puanlamanın bu şekilde yapılmasındaki temel amaç, öğrencileri çeldirici elemeye teşvik etmek veya yanlış cevapları için cezalandırmak değil; öğrenme düzeyini veya öğrenme eksiklerini daha ayrıntılı bir şekilde ortaya koyabilmektir. Böylece, öğrencilere toplu veya bireysel geribildirimlerin sağlanması; öğrencilerin yönlendirilebileceği ek çalışma, etkinlik ve kaynakların belirlenmesi; uygulanabilecek telafi ya da yetiştirme programlarının hazırlanması gibi amaçlarla yararlanılabilir.

Yöntem

Araştırmada, Tip 1 gelişimsel araştırma yöntemi benimsenmiştir. Tip 1 gelişimsel araştırmalarda, belirli bir duruma yönelik ürünün geliştirme sürecini tanımlama, analiz etme ve nihai ürünü değerlendirme süreçlerine yer verilmektedir (Richey ve diğ., 2004). Sırasıyla analiz, tasarım, geliştirme, uygulama ve değerlendirme aşamaları izlenerek kısmi puanlamaya imkan sağlayan bir çevrimiçi sınav sistemi geliştirilmiştir.

Araştırma, özel eğitim öğretmenliği lisans programına devam eden 74 öğrenci üzerinde yürütülmüştür. Katılımcılar 2. ve 4. sınıf düzeyinde olup bilimsel araştırma yöntemleri dersi almaktadır. Öğrencilerin bilimsel araştırma yöntemleri dersindeki başarı düzeylerini belirlemek amacıyla 20 madde içeren çoktan seçmeli bir test geliştirilmiştir. Nitelikli çeldirici üretmenin zorluğu ve bir bütün olarak testin ve testteki her bir maddenin yüksek psikometrik özelliklere sahip olması için üç ya da dört seçeneğin yeterli olabileceği belirtilmektedir (Loudon ve Macias-Munoz, 2018; Rodriguez, 2005). Güçlü çeldiricilere yer verildiği takdirde, tam bilgi ile kısmi bilgi ve kavram yanılgısı arasındaki farkların daha iyi ortaya koyulabileceğinden hareketle testlerin dört seçenek (bir doğru cevap, üç çeldirici) içermesinin uygun olacağı düşünülmüştür. Geliştirilen testlerin kapsam geçerliğinin incelenmesi amacıyla, lisans düzeyinde bilimsel araştırma yöntemleri ve ölçme ve değerlendirme derslerini yürütmüş ölçme ve değerlendirme alan uzmanlarından görüş alınmıştır. Uzmanlardan, hazırlanan soruları kapsamı temsil etme, bilimsel doğruluk ve ölçme ve değerlendirme ilkelerine uygunluk açısından incelemeleri ve görüşlerini belirtmeleri istenmiştir. Uzmanlardan gelen eleştiri ve öneriler doğrultusunda sorular üzerinde gerekli değişiklik ve iyileştirmeler yapılmış ve böylece teste son hali verilmiştir. Elde edilen veriler

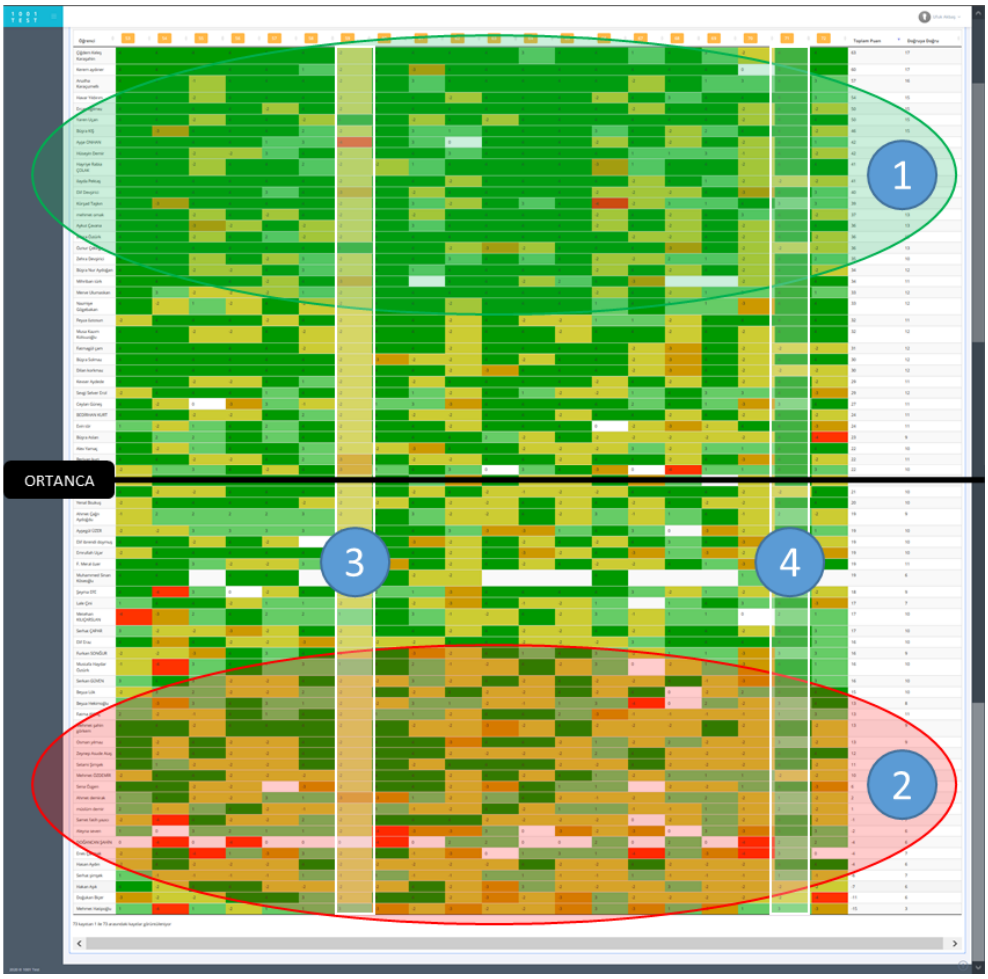
üzerinden betimsel istatistikler, iç tutarlılık katsayıları, madde istatistikleri hesaplanmış ve ek olarak öğrencilerin her bir sorudan aldığı puanın bir bütün olarak görülmesini sağlayan ve "geribildirim haritası" şeklinde adlandırılması uygun görülen özel bir tablo hazırlanmıştır.

Sonuçlar

EDDC yöntemiyle alınan puanların büyükten küçüğe sıralanmasıyla oluşturulan geribildirim haritası Şekil 2'de sunulmuştur.

Şekil 1

Geribildirim Haritası



Şekil 1'in, "1" ile gösterilen bölgesinde yer alan öğrenciler testteki soruların çoğundan pozitif ve büyük ölçüde tam puan almışlardır. Diğer bir ifade ile bu bölgedeki öğrenciler, tam bilgiye yakın öğrenme düzeyine sahiptir. "2" ile gösterilen bölgesinde yer alan öğrencilerin ise testteki soruların

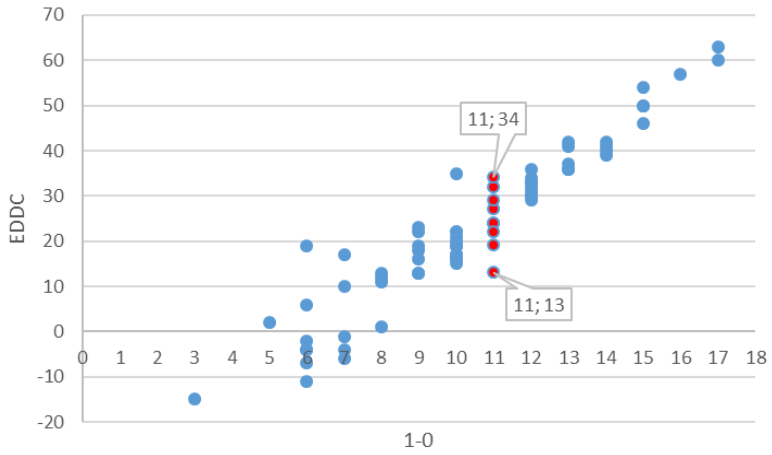
çoğundan negatif ve büyük ölçüde düşük puanlar aldıkları görülmektedir. Bu bölgedeki öğrencilerin, testte yer alan konularda kısmi ya da tam kavram yanılgısı yaşadığı söylenebilir.

Geribildirim haritası, öğretim elemanına testteki maddeler hakkında da bilgi vermektedir. Örneğin, "3" numaralı sütun, testteki 7 numaralı maddeden alınan puanları içermektedir. Bu madde ile ölçülen özellik hakkında tam bilgiye sahip olan öğrenci sayısı oldukça az, kavram yanılgısı yaşayan öğrenci sayısı ise fazladır. Diğer bir örnek ise "4" numaralı sütunda bulunan 19 numaralı madde için verilebilir Bu maddede ölçülen özelliğe yönelik olarak öğrencilerin büyük bir kısmının tam ya da kısmi bilgiye sahip olduğu görülmektedir.

EDDC yöntemi ile yapılan puanlamanın 1 – 0 yöntemiyle yapılan puanlamaya göre üstünlüğü, iki yöntemle alınan puanlar saçılma grafiği üzerinde karşılaştırıldığında açıkça görülmektedir (Şekil 3).

Şekil 2

EDDC ve 1-0 Yöntemleriyle Elde Edilen Puanlara İlişkin Saçılma Grafiği



EDDC ve 1 – 0 yöntemleri ile yapılan puanlamalar ile elde edilen ölçümler arasındaki korelasyonun oldukça yüksek olduğu görülmüştür, r= .95, p< .05. Şekil 2'de kırmızı renk ile belirtilen noktalar, testin 1 – 0 yöntemiyle puanlanması halinde 11 puan alan öğrencileri temsil etmektedir. Testin EDDC yöntemiyle puanlanması halinde bu öğrencilerin alacakları puanların 13-34 aralığında değişeceği görülmektedir. Sonuç olarak EDDC yöntemiyle yapılan puanlamanın öğrenme düzeyi hakkında ayrıntılı bilgi sağlayabileceği görülmüştür.

Kaynaklar

Abu-Sayf, F. (1979). The scoring of multiple-choice tests: A closer look. *Educational Technology, 19*(6), 5-15. http://www.jstor.org/stable/44421466

Arnold, J. & Arnold, P. (1970). On scoring multiple choice exams allowing for partial knowledge. *The Journal of Experimental Education, 39*(1), 8-13. http://www.jstor.org/stable/20157146

Bauer, D., Holzer, M., Kopp, V. & Fischer, M. R. (2011). Pick-n multiple choice-exams: a comparison of scoring algorithms. *Advances in health sciences education, 16*(2), 211-221. https://doi.org/10.1007/s10459-010-9256-1

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması*. ÖSYM Yayınları.

Coombs, C. H., Milholland, J. E. & Womer, F. B. (1956). The assessment of partial knowledge. *Educational and Psychological Measurement, 16*(1), 13-37. https://doi.org/10.1177/001316445601600102

Cross, L. H. & Frary, R. B. (1977). An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests, *Journal of Educational Measurement*, 313-321. https://doi.org/10.1111/j.1745-3984.1977.tb00047.x

Çetin, B. (2005). *Geleneksel yöntemle ve eleme yöntemi ile puanlanan çoktan seçmeli testlerin psikometrik özelliklerinin incelenmesi.* (Tez No. 159929) [doktora tezi, Hacettepe Üniversitesi]. Yükseköğretim Kururmu Tez Merkezi.

Dirkzwager, A. (1996). Testing with personal probabilities: 11-year-olds can correctly estimate their personal probabilities, *Educational and Psychological Measurement, 56*(6), 957-971.

Gözen Çıtak, G. (2010). *Klasik test ve madde tepki kuramlarına göre çoktan seçmeli testlerde farklı puanlama yöntemlerinin karşılaştırılması* (Tez No. 234256) [Doktora tezi, Ankara Üniversitesi]. [Yükseköğretim Kururmu Tez Merkezi]

Gulliksen, H. (1950). *Theory of mental tests*. John Wiley & Sons, Inc.

Hoe, L. S., Kiong, L. N., Sam, H. K., & Usop, H. B. (2009). Improving educational assessment: A computer-adaptive multiple choice assessment using NRET as the scoring method, *US-China Education Review, 6*(5), 51-60.

Kansup, W., & Hakstian, A. R. (1975). A comparison of several methods of assessing partial knowledge in multiple-choice tests: I. Scoring procedures, *Journal of Educational Measurement*, 219-230.

Kaplan, M., & Atalmış, E. H. (2019). Examining gender bias in multiple choice item formats violating item writing guidelines, *International Online Journal of Educational Sciences, 11*(1), 214–229.

Koçak, D. (2013). *Farklı yönergelerle verilen çoktan seçmeli testlerde yanıtlama davranışlarının incelenmesi.* (Tez No. 342461) [Yüksek lisans tezi, Ankara Üniversitesi]. Yükseköğretim Kurumu Tez Merkezi.

Lau, P. N. K., Lau, S. H., Hong, K. S., & Usop, H. (2011). Guessing, partial knowledge, and misconceptions in multiple-choice tests, *Journal of Educational Technology & Society, 14*(4), 99-110.

Lord, F. M. (1975). Formula scoring and number-right scoring, *The Journal of Educational Measurement, 12*, 7-11.

Loudon, C. & Macias-Munoz, A. (2018). IItem statistics derived from three-option versions of multiple-choice questions are usually as robust as four- or five-option versions: Implications for exam design, *Advances in Physiology Education, 42*, 565-575.

Pehlivan Tunç, E. B. ve Kutlu, Ö. (2014). Türkçe test maddelerinde yanıtlama davranışlarının incelenmesi, *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 5*(1), 61-71.

Richey, R. C., Klein, J. D., & Nelson, W. A. (2004). Developmental Research: Studies of Instructional Design and Development. In D. H. Jonassen (Ed.), *Handbook of research on educational communications and technology* (2nd ed.). Lawrence Erlbaum Associates, Inc.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research. *Educational Measurement: Issues and Practices, 24*(2), 3-13.

Selvi, H. ve Derici Yıldırım, D. (2020). Örtük sınıf analizinin farklı puanlama durumlarında incelenmesi. *Bolu Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 20*(2), 754-766.

Tan, Ş. (2004). Çoktan seçmeli testlerde şans başarısını gidermede ölçmenin standart hatasının kullanımı. *Celal Bayar Üniversitesi Sosyal Bilimler Dergisi, 2*(2), 123-131. https://dergipark.org.tr/en/download/article-file/1142570

Umay, A. (1998). Seçmeli derslerde yanıtlayıcı davranışları ve şans başarısının elimine edilmesi işlemlerine ilişkin bazı öneriler. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 14*(14).

Waters, B. K. (1976). The measurement of partial knowledge a comparison between two empirical option-weighting methods and rights-only scoring. *The Journal of Educational Research, 69*(7), 256-260. https://www.jstor.org/stable/27536896

Wu, Q., De Laet, T., & Janssen, R. (2019). Modeling partial knowledge on multiple-choice items using elimination testing. *Journal of Educational Measurement, 56*(2), 391-414. https://doi.org/10.1111/jedm.12213

Yurdugül, H. (2010). Farklı madde puanlama yöntemlerinin ve farklı test puanlama yöntemlerinin karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 1*(1), 1-8.

TIMSS 2019 Matematik maddelerinin uygulama ortamına göre değişen madde fonksiyonunun belirlenmesi

Ahmet Yıldırım ve Burcu Parlak

Anahtar kelimeler: TIMSS, uygulama ortamı, bilgisayar tabanlı değerlendirme, kâğıt-kalem testleri, değişen madde fonksiyonu

Giriş

Son yıllarda, psikometristler ve araştırmacılar eğitimde ve psikolojide kullanılan testlerin farklı gruplar için farklı şekilde çalışıp çalışmadığı ile ilgilenmektedir. Yani bir testte yer alan soruların belli bir alt grup için (örn. erkekler, farklı etnik kökene sahip olanlar vb.) yanlılık gösterip göstermediği veya ilgili gruba avantaj sağlayıp sağlamadığı araştırma konusu olmaktadır (Embretson ve Reise, 2000). Bir maddenin testi alan bir *alt örneklem* için farklı çalışıp çalışmadığı ile ilgili çalışmalar *değişen madde fonksiyonu* (DMF) çalışmalarını gündeme getirmektedir.

DMF çalışmalarında temel olarak iki alt grup (örn. erkekler ve kadınlar) yer almakta ve gruplardan biri referans, diğeri ise odak grup olarak ele alınmaktadır. DMF çalışmalarında araştırmacılar, testte yer alan maddelerin referans ve odak gruplarda eş değer bir biçimde çalışıp çalışmadığını incelemektedir (Golia, 2012). Bir maddenin DMF göstermesi durumunda, farklı alt grupların (referans ve odak gruplar) aynı yetenek düzeyindeki bireylerinin testte yer alan ilgili maddeyi yanıtlama olasılıklarının aynı olmadığı ileri sürülebilir. Bir başka deyişle, grup üyeliğine bağlı olarak farklı gruplarda yer alan bireylerin maddeyi yanıtlama olasılıkları farklılık göstermektedir (Embretson ve Reise, 2000). Yani DMF, bir maddenin iki alt grup için farklı şekillerde çalışmasına bağlı olarak ortaya çıkan psikometrik farklılığa karşılık gelmektedir. Eğer bir testte yer alan bir madde farklı alt gruplarda aynı "madde yanıt fonksiyonu"na sahipse, farklı gruplarda yer alan aynı yetenek düzeyindeki bireylerin grup üyeliğinden bağımsız olacak şekilde ilgili maddeyi doğru yanıtlama olasılıkları aynı olacaktır (Dorans ve Schmitt, 1991).

Tek biçimli DMF ve tek biçimli olmayan DMF olmak üzere DMF'nin iki türü mevcuttur. Tek biçimli DMF'de bir madde tutarlı bir biçimde, bütün yetenek düzeylerinde bir grup lehine çalışmakta ve ilgili gruba bütün yetenek düzeylerinde avantaj sağlamaktadır. Tek biçimli olmayan DMF'de ise madde, bazı yetenek düzeylerinde referans grup lehine bazı yetenek düzeylerinde ise odak grup lehine çalışmaktadır. Bir başka deyişle, tek biçimli olmayan DMF'de maddeye ilişkin performans koşullu bir

biçimde grup üyeliğine bağımlılık göstermekte ve yetenek düzeyi ile grup üyeliği arasında bir etkileşim etkisi ortaya çıkmaktadır (Golia, 2012).

DMF yöntemleri alt gruplar arasında performans farklılığına yol açan maddelerin belirlenmesine dayanmaktadır. Bu çalışmalarda farklı alt grupların bir *eşleştirme değişkenine* bağlı olarak eşleştirilmesi gerekmektedir. DMF çalışmalarında bu eşleştirme değişkeni genellikle toplam puan veya yetenek düzeyi olmaktadır (Dorans ve Schmitt, 1991).

Teknolojinin ilerlemesiyle birlikte ölçme ve değerlendirme uygulamalarının hem ulusal hem de uluslararası değerlendirme çalışmalarında (örn. PISA ve TIMSS) bilgisayar ortamında gerçekleştirilmesine yönelik eğilimin arttığı gözlenmektedir. Bu nedenle kâğıt-kalem testleri ile bilgisayar tabanlı testlerin karşılaştırılabilirliğinin ele alınması bir mecburiyet doğurmaktadır (McClelland ve Cuaves, 2020; Wang ve diğ., 2008). Çünkü önemli olan öğrencilerin bilgisayar yeterliklerinden ziyade alan bilgisine ilişkin yeterliklerinin karşılaştırma konusu edilmesidir. Bir sınavın bilgisayar ortamında ve kâğıt-kalem ortamında uygulanmasından kaynaklanan birçok farklılık söz konusu olabilmektedir. Bu durumların alt gruplar (bilgisayar ortamında test alanlar ve kâğıt-kalem ortamında test alanlar) için farklılık gösterebileceği düşünülmektedir. McClelland ve Cuaves (2020) inceledikleri farklı çalışmalarda kâğıt-kalem ortamında uygulanan testler ile bilgisayar ortamında uygulanan testler arasındaki farklılığın genel olarak test için harcanan zamana, uygulamadan kaynaklanan faktörlere, ölçmenin yapıldığı alana, bilgisayara aşinalığa bağlı olarak farklılaştığını ortaya koymuşlardır.

Türkiye'nin de katıldığı uluslararası bir değerlendirme çalışması olan TIMSS uygulaması, 2019 yılında elektronik ortamda gerçekleştirilmiştir. Ancak elektronik ortamda yapılan uygulamaya ek olarak bir grup öğrenci TIMSS kapsamında uygulanan testleri kâğıt-kalem ortamında almıştır. Öğrencilerin teknolojik gelişmeler bağlamında daha teknoloji yönelimli olması ve kullanılan maddelerin doğası gereği elektronik ortamda uygulanmaya daha müsait olması nedeniyle (interaktif ve yenilikçi maddeler) bu maddelerin testi elektronik ortamda alan bireyler ile kâğıt-kalem ortamında alan bireyler için farklı bir biçimde çalışabileceği düşünülmektedir. Bu nedenle, uygulama ortamlarına göre maddelerin değişen madde fonksiyonu gösterip göstermediğinin belirlenmesi amacıyla bu çalışmanın yapılmasına ihtiyaç duyulmuştur. Bu araştırmanın amacı, TIMSS 2019 matematik maddelerinin uygulama ortamına göre değişen madde fonksiyonu gösterip göstermediğinin belirlenmesidir.

Yöntem

Bu araştırmada, TIMSS 2019 matematik maddelerinin uygulama ortamına göre DMF gösterip göstermediğinin belirlenmesi amaçlanmaktadır. Bu kapsamda ilgili maddeleri farklı uygulama ortamlarında alan gruplara göre maddelerin farklı bir şekilde çalışıp çalışmadığı incelenmiş dolayısıyla ilişkisel tarama modelinde bir araştırma yürütülmüştür. İlişkisel tarama modeli, iki veya daha fazla değişken arasında ilişki olup olmadığını ve/veya ilişkinin derecesini ortaya koymayı amaçlayan bir araştırma modelidir (Karasar, 2008).

Araştırma grubunu, TIMSS 2019 uygulamasına Türkiye'den katılan sekizinci sınıf öğrencileri oluşturmaktadır. Araştırma kapsamında TIMSS 2019 uygulamasına Türkiye'den katılan sekizinci sınıf öğrencilerin matematik testine ait verileri kullanılmıştır. Uygulamaya Türkiye'den sekizinci sınıf düzeyinde elektronik ortamda 4077 öğrenci, kâğıt-kalem ortamında ise 1819 öğrenci katılmıştır (MEB, 2020). Veri analizi çalışmaları elektronik ortamda birinci kitapçıkta, kâğıt-kalem ortamında ise 22. kitapçıkta kullanılan (Mullis ve Martin, 2017) 16 madde üzerinde gerçekleştirilmiştir. Bu maddeleri elektronik ortamda alan öğrencilerin sayısı 287 iken aynı maddeleri kâğıt-kalem ortamında alan öğrencilerin sayısı 226'dır. Bu ortak maddeler belirlendikten sonra maddelerin uygulandığı ortama göre (elektronik ortam ve kâğıt-kalem ortamı) DMF gösterip göstermediğini belirlemek amacıyla DMF analizleri gerçekleştirilmiştir.

Değişen madde fonksiyonunu incelemede tutarlı sonuçların elde edilebilmesi için en az iki yöntemin kullanılması gerektiği ifade edilmektedir (Sireci ve Rios, 2013). TIMSS 2019 matematik maddelerinin uygulama ortamına göre DMF gösterip göstermediğini belirlemek amacıyla biri Klasik Test Kuramı (KTK) diğeri ise Madde Tepki Kuramına (MTK) dayalı iki yöntem kullanılmıştır. Bu yöntemler: MTK olabilirlik oran yöntemi ile lojistik regresyon yöntemleridir.

Sonuçlar

TIMSS 2019 matematik testlerinde yer alan maddelerin uygulama ortamına göre değişen madde fonksiyonu gösterip göstermediğini belirlemek amacıyla yapılan MTK olabilirlik oran testinden elde edilen sonuçlar, 16 madde içinden dördünün DMF gösterdiğini ortaya koymaktadır. Bu maddelerin üç tanesi ihmal edilebilir düzeyde DMF göstermekte iken bir madde ise orta düzeyde DMF göstermektedir. İlgili dört maddeden üçünün uygulamayı kâğıt-kalem ortamında alan bireylerin lehine, bir maddenin ise uygulamayı elektronik ortamda alan bireylerine lehine olduğu ortaya konulmuştur. Lojistik regresyon yöntemine göre gerçekleştirilen değişen madde fonksiyonu analizi sonucunda ise 16 madde içinden ikisinin DMF gösterdiği tespit edilmiştir. DMF gösteren maddelerin DMF büyüklüklerinin ihmal edilebilir düzeyde olduğu ortaya konulmuştur.

Kaynaklar

Dorans, N. J. and Schmitt, A. P. (1991). *Constructed response and differential item functioning: a pragmatic approach.* Educational Testing Service.

Embretson, S. E., and Reise, S. P. (2000). *Item response theory for psychologists.* Lawrence Erlbaum Associates.

Golia, S. (2012). Differential item functioning classification for polytomously scored items. *Electronic Journal of Applied Statistical Analysis, 5*(3), 367-373.

Karasar, N. (2008). *Bilimsel araştırma yöntemi.* Nobel Akademik Yayıncılık.

McClelland, T., and Cuaves, J. (2020). A comparison of computer based testing and paper and pencil testing in mathematics assessment. *The Online Journal of New Horizons in Education, 10*(2), 78-89.

MEB (2020). *TIMSS 2019 Türkiye ön raporu* (Eğitim Analiz ve Değerlendirme Raporları Serisi No. 15). Milli Eğitim Bakanlığı. http://www.meb.gov.tr/meb_iys_dosyalar/2020_12/08202713_No15_TIMSS_2019_Turkiye_On_Raporu.pdf

Mullis, I. V. S., and Martin, M. O. (2017). *TIMSS 2019 assessment frameworks*. TIMSS & PIRLS International Study Center and IEA. https://timss2019.org/wp-content/uploads/frameworks/T19-Assessment-Frameworks.pdf

Sireci, S. G., and Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation: An International Journal on Theory and Practice*, *19*(2-3), 170-187. https://doi.org/10.1080/13803611.2013.767621

Wang, S., Jiao, H., Young, M. J., Brooks, T., and Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement*, *68*(1), 5-24. https://doi.org/10.1177/0013164407305592

ABİDE Puanları ile ortak yazılı sınav puanları arasındaki ilişkinin kanonik korelasyon analizi ile incelenmesi

Burcu Parlak ve Ahmet Yıldırım

Anahtar kelimeler: ABİDE, kanonik korelasyon, ortak yazılı sınav

Giriş

Kanonik korelasyon analizi, çok değişkenli bir istatistiksel modeldir. Bu istatistiksel model, birden fazla bağımlı ve bağımsız değişkenin yer aldığı veri setleri arasındaki ilişkilerin incelenmesini temel almaktadır. Yani kanonik korelasyon analizi, her biri iki veya daha fazla değişkenden oluşan iki değişken kümesi arasındaki ilişkiyi modellemektedir (Dattalo, 2014; Yin, 2004). Kanonik korelasyon analizinin mantığı, iki değişken seti arasındaki ilişkiyi maksimize edecek şekilde iki değişken setinin her birinden doğrusal bir değişken kombinasyonu elde etmektir (Chacko, 1986; Dattalo, 2014). Kanonik fonksiyonlar elde etme, faktör analizinde işe koşulan sürece benzemektedir. Faktör analizinde çıkartılan ilk faktör, değişken setindeki en fazla varyansı açıklayan faktördür. Birinci faktör tarafından açıklanamayan varyansı mümkün olduğunca açıklayabilmesi için ikinci faktör çıkartılır. Kanonik korelasyon analizinde de benzer işlem adımları takip edilir ancak kanonik korelasyon analizi, tek bir değişken setindeki değişkenlerden ziyade iki değişken seti arasındaki maksimum ilişkiyi açıklamaya çalışır. Bu yüzden iki değişken seti arasındaki olası en yüksek ilişkiyi ortaya koyacak şekilde ilk kanonik fonksiyon çıkartılır. İkinci kanonik fonksiyon ise ilk kanonik fonksiyon tarafından açıklanamayan iki değişken seti arasındaki maksimum ilişkiyi gösterir. Elde edilebilecek maksimum kanonik fonksiyon sayısı, değişken sayısı daha az olan değişken kümesindeki sayıya eşittir (Chacko, 1986; Dattalo, 2014). Yani iki bağımlı, dört bağımsız değişkenin yer aldığı iki değişken kümesindeki ilişkilerin incelendiği bir kanonik korelasyon analizinde kanonik fonksiyon sayısı ikidir. Kanonik korelasyon analizi, *genel doğrusal modelin* en genel durumu olup diğer bütün parametrik testler (örn. çoklu doğrusal regresyon, MANOVA, ANOVA gibi), kanonik korelasyon analizinin özel durumları olarak değerlendirilebilir. Özellikle çok sayıda değişkenin birbiri ile ilişki ve etkileşim gösterdiği sosyal bilimler ve eğitim bilimleri alanlarında ortak bağımlılık gösteren değişkenler arasındaki ilişkinin incelenmesi için kanonik korelasyon analizi büyük bir önem arz etmektedir (Chacko, 1986; Dattalo, 2014).

Kanonik korelasyon analizi, Pearson momentler çarpımı korelasyon katsayısının genelleştirilmiş halidir (Dattalo, 2014). Yin (2004) kanonik korelasyon analizini, çoklu doğrusal regresyon analizinin bir uzantısı olarak değerlendirmektedir.

Kanonik korelasyon analizinin yapıldığı çalışmalar (İlhan ve diğ., 2013; Özdemir ve Gelbal, 2014; Sayın ve diğ., 2012) incelendiğinde her iki değişken setinin de ulusal düzeyde uygulanan geniş ölçekli testlere ait olmadığı ve böyle bir araştırmaya ihtiyaç duyulduğu görülmektedir. ABİDE araştırması, Milli Eğitim Bakanlığı (MEB) tarafından ulusal düzeyde yürütülen geniş ölçekli bir izleme araştırmasıdır. ABİDE araştırmasının genel amacı, öğrencilerin akademik becerilerinin ortaya konulması ve bu becerilerle ilişkili öğrenci, öğretmen ve okul özelliklerinin belirlenmesidir. ABİDE 2016 araştırması sekizinci sınıf öğrencilerinin Türkçe, matematik, fen bilimleri ve sosyal bilgiler alanlarındaki akademik becerilerini ölçecek şekilde gerçekleştirilmiştir (MEB, 2017). Öğrencilerin kazanım odaklı ortak yazılı sınavlardan aldıkları puanların ABİDE testlerinden aldıkları puanlar ile ne düzeyde ilişki gösterdiğini ortaya koymak amacıyla bu araştırmanın yapılmasına gereksinim duyulmuştur. Bu amaçla bu araştırmada geniş ölçekli bir izleme araştırması olan ABİDE'den elde edilen puanlar ile yine merkezi olarak yapılan ortak yazılı sınavlara ait puanlar arasındaki ilişkinin incelenmesinin önemli olduğu düşünülmektedir. Bu araştırmanın amacı, MEB tarafından yürütülmekte olan ABİDE araştırmasında öğrencilerin elde ettikleri Türkçe, matematik ve fen bilimleri puanları ile MEB tarafından sekizinci sınıf öğrencilere uygulanan ortak yazılı sınav puanları arasındaki ilişkinin incelenmesidir.

Yöntem

Bu araştırma, her iki değişken kümesinde üç değişkenin yer aldığı ve bu değişkenler arasındaki ilişkilerin incelenmesinin amaçlandığı bir araştırmadır. Bu amaca uygun olacak şekilde bu araştırmada ilişkisel tarama modeli kullanılmıştır. İlişkisel tarama modeli, iki veya daha fazla değişken arasında ilişki olup olmadığını ve/veya ilişkinin derecesini ortaya koymayı amaçlayan bir araştırma modelidir (Karasar, 2008). Araştırma grubunu, 2015-2016 eğitim öğretim yılında sekizinci sınıfta olup ABİDE araştırmasına ve ortak yazılı sınavlara katılan öğrenciler oluşturmaktadır.

Araştırma kapsamında ABİDE 2016 uygulamasına ait Türkçe, matematik ve fen bilimleri testlerine ait puanlar ile aynı yıl MEB tarafından yapılan ortak yazılı sınavlara giren öğrencilerin Türkçe, matematik ve fen bilimleri puanları kullanılmıştır. Öncelikle 2016 yılında hem ABİDE araştırmasına hem de ortak yazılı sınavlara katılan öğrencilerin verileri eşleştirilerek toplam 34772 öğrenciye ait veri seti elde edilmiştir. İlgili veri setinden tesadüfi olarak yaklaşık 1000 gözleme sahip (veri setinin yaklaşık %3'ü) üç ayrı veri seti çekilmiş ve bu üç veri seti üzerinde kanonik korelasyon analizleri yürütülmüştür. Elde edilen sonuçların birbiriyle tutarlılık gösterdiğinin ortaya konulması üzerine tesadüfi olarak belirlenen veri setlerinden birine ait bulgular raporlaştırılmıştır. Sonuç olarak analizler 995 gözleme ait bir veri seti üzerinde yürütülmüştür.

ABİDE araştırmasına ait Türkçe, matematik ve fen bilimleri puanlarından oluşan değişken kümesi ile öğrencilerin ortak yazılı sınavlarına ait Türkçe, matematik ve fen bilimlerinden oluşan değişken

kümesi arasındaki ilişkinin incelenmesi amacıyla kanonik korelasyon analizi kullanılmıştır. Kanonik korelasyon analizi öncesinde varsayımlar (çok değişkenli normal dağılım, doğrusallık, eşvaryanslılık ve çoklu doğrusal bağlantı) test edilmiştir (Tabachnick ve Fidell, 2018).

Sonuçlar

ABİDE araştırmasına ait Türkçe, matematik ve fen bilimleri puanlarından oluşan değişken kümesi ile öğrencilerin ortak yazılı sınavlarına ait Türkçe, matematik ve fen bilimleri puanlarından oluşan değişken kümesi arasındaki ilişkinin incelenmesi amacıyla yapılan kanonik korelasyon analizi sonucunda üç farklı kanonik korelasyon katsayısı ve kanonik değişken çifti elde edilmiştir. Wilk's Lambda ve ki-kare değerleri, hesaplanan kanonik korelasyon değerlerinin anlamlılık düzeyleri ile ilgili bilgi vermektedir. Buna göre, ilk iki kanonik korelasyon katsayısının istatistiksel olarak anlamlı olduğu ($p< .05$) belirlenmiştir. Birinci kanonik korelasyon çiftinin ortak varyansın %71'ini, ikinci kanonik korelasyon çiftinin ise ortak varyansın %7'sini açıkladığı görülmektedir. İlk iki kanonik değişken çifti anlamlı bulunmuş olmasına rağmen açıklanan ortak varyans yüzdeleri incelendiğinde birinci kanonik korelasyon çiftinin ikinci kanonik korelasyon çiftine göre daha anlamlı sonuçlar verdiği ve varyansın önemli bir bölümünü açıkladığı görülmektedir.

Kaynaklar

Chacko, H. E. (1986, November 19-21). *An example of the use of canonical correlation analysis.* [Paper Presentation]. The Annual Meeting of the Mid-South Educational Research Association. Memphis, Tennessee, USA.

Dattalo, P. V. (2014). *A demonstration of canonical correlation analysis with orthogonal rotation to facilitate ıinterpretation*. Unpublished manuscript, School of Social Work, Virginia Commonwealth University, Richmond, Virginia.

İlhan, M. Çetin, B., Öner-Sünkür, M. ve Yılmaz, F. (2013). Ders çalışma becerileri ile akademik risk alma arasındaki ilişkinin kanonik korelasyon ile incelenmesi. *Eğitim Bilimleri Araştırmaları Dergisi*, *3*(2), 123-146. https://dergipark.org.tr/tr/pub/ebader/issue/44712/555593

Karasar, N. (2008). *Bilimsel araştırma yöntemi*. Nobel Akademik Yayıncılık.

MEB (2017). *Akademik becerilerin izlenmesi ve değerlendirilmesi (ABİDE) 8. sınıflar raporu*. Ölçme, değerlendirme ve sınav hizmetleri genel müdürlüğü. http://edirne.meb.gov.tr/meb_iys_dosyalar/2018_06/08104327_ABYDE_Turkiye.pdf

Özdemir, B. ve Gelbal, S. (2014). PISA 2009 sonuçlarına göre öğrenci başarısını etkileyen faktörlerin kanonik ortak etki analizi ile incelenmesi. *Eğitim ve Bilim*, *39*(175), 41-57.

Sayın, A., Koğar, H. ve Çakan, M. (2012). Aşamalı dersler arasındaki ilişkilerin kanonik korelasyon tekniğiyle incelenmesi: Sınıf öğretmenliği örneği. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *3*(1), 210-220. https://dergipark.org.tr/tr/pub/epod/issue/5803/77222

Tabachnick, B. G., and Fidell, L. S. (2018). *Using multivariate statistics* (7th ed.). Pearson.

Yin, X. (2004). Canonical correlation analysis based on information theory. *Journal of Multivariate Analysis*, 91, 161-176. https://doi.org/10.1016/S0047-259X(03)00129-5

Matematik başarısını etkileyen duyuşsal özelliklerin MARS yöntemiyle incelenmesi

Çağla Kuddar ve Sevda Çetin

Anahtar kelimeler: Veri madenciliği, R programı, SPM, MARS.

Giriş

MARS (Multivariate Adaptive Regression Splines) tekniği ilk olarak 1991 yılında Stanford 'da fizikçi Jerame Friedman tarafından geliştirilmiştir. MARS modeli veri madenciliğinde oldukça popüler bir model haline gelmeye başlamıştır. MARS modelinin araştırmacılara sağladığı pek çok avantaj vardır. MARS modeli sebep ve sonuç değişkenleri arasında bir varsayım gerektirmez ve herhangi bir matematiksel ilişki aramaz. Genel olarak, doğrular birbirleriyle ilişki kurar ve bu doğrular, temel fonksiyonlar olarak da bilinir, doğrusal ve doğrusal olmayan davranış örüntüleri ve değişkenler arasında çok yönlü ve esnek bir ilişki kurar (Zhang ve Goh, 2016). MARS modeli adımsal bir regresyon yöntemidir (Özfalcı, 2008). Adımsal regresyon analizi, ileriye doğru seçme yönteminin gelişmiş bir yöntemi olarak düşünülebilir (Anıl, 2010). Bu metot ile bağımlı değişken ve bağımsız değişkenler arasındaki korelasyona göre tahmin modeline en çok katkısı olabilecek değişkenler seçilir, önemsiz olanlar elenir. Böylece modeldeki sapmalar azaltılarak tahmin doğruluğu yüksek bir model elde edilir. Bağımlı ve bağımsız değişkenler arasındaki korelasyon katsayına göre, korelasyon katsayısı en yüksek olan bağımsız değişken ilk olarak modele alınır. Adımsal regresyon modeli ile en yüksek doğruluklu en az hatalı tahmin modeli elde edilir (Zateroğlu ve Kandırmaz, 2018).

Uluslarası sınavlarda ya da veri toplama yöntemleri ile elde ettiğimiz veriler gereken varsayımları sağlamama eğiliminde olabilir. Bu varsayımlar sağlanmadığında çeşitli istatistiksel yöntemler kullanılamamaktadır. Bu durumda şu söylenebilir: istatistiksel olarak araştırmacılar kısıtlılık yaşayabilmektedir. Bu durum bir problem olarak karşımıza çıkabilir. Veri çağı olarak adlandırılan çağımızda veri analiz yöntemleri ne kadar gelişmiş olursa veriden yararlanma, veriyi anlama, veriden bilgi çıkarma o kadar kolay olur. Teknoloji ile birlikte gelişen çağımızda veri güçtür. Bu gücü etkili kullanmak veri analiz yöntemlerini geliştirmek ile mümkündür. Parametrik olmayan verilerde durum daha farklıdır. Bu yüzden bu sayıltıların sağlanmadığı durumlarda yeni parametrik olmayan yöntemler geliştirmenin önemli olduğu düşünülmektedir. Eğitim ve sosyal bilimler alanında gerekli varsayımları sağlamayan verilerde veri madenciliği yöntemleri uygulanabilir. Veri madenciliği, yüzlerce değişken arasında sayısız

olası ilişkiyi araştırmak için hesaplama gücünün yardımıyla anlamlı yeni bilgileri keşfetmek için tekrarlanan bir süreçtir (Yoon ve ve diğ., 2016). Bu yönü nedeniyle veri madenciliği yöntemleri eğitim ve sosyal bilimler alanında kullanıldığında da gerçeğe en yakın tahminler verilebilir. Alan yazında istatistiksel analiz karşılaştırmalarının az da olsa olduğu görülmüştür. Bu çalışmada MARS modelinin incelenmesi planlanmaktadır.

Yöntem

Çalışma, TIMSS 2019 değerlendirmesinden faydalanarak karşılaştırmalı veri analizi yapmayı amaçladığı ve bunu yaparken de hazır paket programlardan yararlandığı için temel araştırma özelliği göstermektedir. Temel araştırma, görünürde özel herhangi bir uygulaması veya kullanımı bulunmayan ve öncelikle olgu ve gözlemlenebilir olayların temellerine ait yeni bilgiler elde etmek için yürütülen deneysel veya teorik çalışmalardır (Anlağan, 2011). Temel araştırmalar sonuç yerine sürece odaklanan bir bilgiyi keşfetmeye yardımcı olan ve keşfetme amacı olan; dünyayı daha iyi anlamamızı ve anlamlandırmamızı sağlayan deneysel ya da teorik çalışmalardır.

Türkiye, TIMMS 2019 döngüsüne 4. Sınıf düzeyinde 180 okul ve 4.028 öğrenci ile katılım göstermiştir. 8. Sınıf düzeyinde ise uygulama 181 okuldaki 4.077 öğrenci katılımı ile gerçekleşmiştir. Bu çalışmanın örneklemi TIMMS 2019 8. Sınıf Matematik değerlendirmesine katılan 4.077 öğrencinin duyuşsal anketlerine verdiği cevaplardan seçilen maddeler ile başarı düzeyini (plausible values) gösteren BSMMAT01-05 değişkenleridir (MEB, 2019).

TIMMS 2019 uygulaması Matematik ve Fen bilimleri alanında dördüncü ve sekizinci sınıf öğrencileri ile yürütülmüştür. Sınavda başarı testleri ve duyuşsal anketler yer almaktadır. Sınavda başarı testi olarak matematik ve fen konuları ile ilişkili sorular sorulmuştur. Duyuşsal anketleri ile de öğrencinin sosyo-ekonomik düzeyi, öğretmene yönelik tutum, öğrencinin derse ilgi ve motivasyonunu, uğradığı zorbalık düzeyini ölçen likert ölçek türünde olan anketlerdir. Bu duyuşsal anketler yalnızca öğrenciler için değil; öğretmen ve okul yöneticileri için de hazırlanmıştır.

Bu çalışma kapsamında TIMMS 2019 sekizinci sınıf matematik öğrenci duyuşsal anketlerinden seçilen maddeler ile başarı puanları veri olarak seçilmiştir. Çalışmaya başlamadan önce Hacettepe Üniversitesi Etik Komisyonunda Etik izni alınmıştır. Çalışma için kullanılan tüm veriler TIMMS sınavının resmî sitesi olan (https://timss2019.org/international-database/) sitesinden alınmıştır. Seçilen maddelerin ayrıntıları şöyledir çalışmanın diğer kısımlarında kodlarıyla verilecektir.

Verilerin analizi üç adımda gerçekleşmiştir. İlk olarak verilerin analize uygunluğu test edilmiş ve veriler analize uygun hale getirilmiştir. Sonrasında ise çeşitli programlar ile MARS veri madenciliği modeli kurulmuştur.

Sonuçlar

Mars ile yapılan analiz sonucunda zorbalık değişkenin başarıyı anlamlı düzeyde etkilediği sonucuna varılmıştır. Öğrencinin okula yönelik olumlu ya da olumsuz tutumunun başarıyı anlamlı

düzeyde etkilediği sonucuna varılmıştır. Ancak hipotezin yönü belirli olmadığı için yönü hakkında yorum yapılamamıştır. Öğretmene yönelik tutum başarıyı anlamlı bir düzeyde etkilese de pozitif yönde olması durumu nedeniyle hipotez kabul edilmemiştir. Öğretmene yönelik tutum başarıyı etkilemektedir sonucuna varılmıştır. Matematiğe yönelik ilgi başarıyı etkilediği sonucuna varılsa da pozitif yönde olup olmadığı hakkında yorum yapılamamıştır. Yani derse ilgili olmak Mars modeli sonucuna göre başarıyı yordamaktadır sonucuna varılmıştır. Ancak yönü hakkında yorum yapılamayacağı sonucuna varılmıştır. Okula yönelik tutum ve öğretmene yönelik tutum arasında matematiğe yönelik ilgi aracılığında başarı üzerinde anlamlı bir etki bulunmuştur. Yani okula yönelik tutum ve öğretmene yönelik tutum arasındaki ilişkide derse ilgi durumu başarıyı yordamaktadır sonucuna varılmıştır. Okula yönelik tutum ile öğretmene yönelik tutum arasındaki ilişkide zorbalık değişkeni analize girdiğinde başarı önemli ölçüde azaldığı sonucuna varılmıştır. Zorbalık değişkeninin başarıyı belirtilen koşullarda bile olumsuz etkilediği sonucuna varılmıştır. Okula yönelik tutum ve öğretmene yönelik tutum ilişkisinde ilgi değişkeninin pek etkisi olmasa da zorbalık değişkeni analize girdiğinde başarı düzeyinin etkilendiği sonucuna varılmıştır. Burada zorbalık değişkeni başarı üzerinde güçlü bir etkisi olan değişken olduğu sonucuna varılmıştır.

Kaynaklar

Anlağan, Ö. (2011). Temel ar-ge ve yenilik kavramları. Ar-Ge, Yenilik ve Teknoloji Politikaları Forumu (AYTEP), TÜBİTAK. https://www.emo.org.tr/ekler/16f6ef8160d5168_ek.pdf

MEB. (2019). *TIMSS 2019 ulusal matematik ve fen ön raporu*. Eğitim Analiz ve Değerlendirme Raporları Serisi, No. 15). Milli Eğitim Bakanlığı. https://odsgm.meb.gov.tr/meb_iys_dosyalar/2020_12/10175514_TIMSS_2019_Turkiye_On_Raporu_.pdf

Özfalcı, Y (2008). *Çok değişkenli uygulanabilir regresyon kesitleri: MARS* (Tez No. 233937) [Yüksek lisans tezi, Gazi Üniversitesi]. Yükseköğretim Kurumu Tez Merkezi.

Yılmaz V. & Çelik E. (2009). *LISREL ile yapısal eşitlik modellemesi* (1. Baskı). Pegem Akademi.

Yoon, S., Co, M. C., Jr, Suero-Tejeda, N., & Bakken, S. (2016). A data mining approach for exploring correlates of self-reported comparative physical activity levels of urban latinos. *Studies in health technology and informatics*, *225*, 553–557.

Zateroğlu, M. T., & Kandırmaz, H. M. (2018). Türkiye için güneşlenme süresi değişiminin izlenmesi, değerlendirilmesi ve bazı meteorolojik verilerle ilişkisinin belirlenmesi. *Ç.Ü Fen ve Mühendislik Bilimleri Dergisi, 35*(3), 105-114. https://fbe.cu.edu.tr/storage/fbeyedek/makaleler/2017/T%C3%9CRK%C4%B0YE%20%C4%B0%C3%87%C4%B0N%20G%C3%9CNE%C5%9ELENME.pdf

Zhang, W., & Goh, A. T. (2016). Multivariate adaptive regression splines and neural network models for prediction of pile drivability. *Geoscience Frontiers*, *7*(1), 45-52.

Doğrusal ve doğrusal olmayan regresyon yöntemlerinin eğitim verileri üzerinde modellenmesi: ÇDR-MARS

Hikmet Şevgin

Anahtar kelimeler: Lineer regresyon, non-lineer regresyon, MARS, ÇDR, PISA

Öz

Bu çalışmada regresyon yöntemlerinden Çoklu Doğrusal Regresyon yöntemi (ÇDR) ile MARS yöntemine ait RMSE, MSE, MAPE ve R^2 metrikleri karşılaştırılarak PISA 2018 veri seti arasından Türk öğrencilere ait veriler üzerinde uygulama yapılmış ve bu iki regresyon yönteminin performanslarını karşılaştırmak amaçlanmıştır. Çalışmada kullanılan veri seti PISA 2018 uygulaması içinden Türk öğrencilere ait okuma becerileri puanları ile öğrencilere ait çeşitli demografik değişkenlerden oluşmaktadır. Çalışma grubunda yer alan 6890 öğrenci sayısı, kayıp veri silme atama işlemleri sonucu 6508 olarak elde edilmiştir. Çalışmada kullanılan verilerin yapısı itibariyle, kategorik yapıya sahip bağımsız değişkenler dummy kodlanarak sürekli hale getirilmiştir. Araştırmada ÇDR yönteminin gerektirdiği varsayımların sınanması adına normallik, doğrusallık ve homojenlik için Kolmogorov-Smirnov testi, histogram grafiği ve standardize edilmiş artık değerler incelenmiştir. Yöntemlerin analizleri için SPSS ve SPM 7.0 programlarından yararlanılmıştır. RMSE, MSE ve MAPE metriklerinde değerlendirme kriteri, regresyon yöntemlerinden düşük değer alan, yüksek değer alana göre daha iyi performans sergiler. RMSE, MSE ve MAPE için MARS yönteminin ÇDR yönteminden daha düşük olduğu görülmüştür. Dolayısıyla MARS yöntemi daha iyi performans sergilemiştir. Yine bağımsız değişkenlerin bağımlı değişkende açıkladıkları varyans miktarı bakımından MARS yönteminin açıkladığı varyans miktarının (% 28.4) ÇDR yönteminden (% 17.5) daha yüksek olduğu görülmüştür. MARS doğrusal olmayan modelleri, doğrusal parçacıklara bölüp her parçacıkta parametre kestirimlerini ayrı bir şekilde yapma özelliğinden dolayı hata oranlarının düşmesinde ve açıklanan varyansın artmasında etkili olduğu söylenebilir. MARS yöntemi normallik doğrusallık ve homojenlik gibi varsayımlara gereksinim duymayan güçlü robust bir yöntemdir. MARS yöntemi bu tür verilerin incelenebilmesine olanak tanımasından ötürü Eğitim Bilimleri çalışmalarında kullanılması önerilmektedir.

Madde tepki kuramı varsayımlarının incelenmesi: Bir doküman analizi

Mahmut Sami Yiğiter ve Erdem Boduroğlu

Anahtar kelimeler: Madde tepki kuramı, varsayım, boyutluluk, yerel bağımsızlık, model veri uyumu

Giriş

Madde Tepki Kuramı (MTK), ölçmelerden elde edilen yanıt örüntülerini modellemek için kullanılan en popüler metodolojilerden biridir. MTK, testte yer alan maddelere bireylerin verdiği yanıtlardan yola çıkarak maddelerin ve bireylerin özelliklerini belirlemek amacıyla kullanılan güçlü bir ölçekleme yöntemidir (Embretson ve Reise, 2000). MTK modelleri, bireyin örtük yapı üzerindeki yeteneğini veya puanını kestirmeyi ve bu örtük yapıyı ölçmek için kullanılan maddelerin özelliklerini eleştirel olarak değerlendirmeyi mümkün kılar. MTK'nın önemli özelliklerinden biri hem bireyleri hem de maddeleri aynı ölçeğe yerleştirmesidir. Bir birey yüksek veya düşük bir yetenek düzeyine sahip olabileceği gibi, bir madde de yüksek veya düşük güçlüğe sahip olabilir ve aynı ölçekte yer alabilir. Bireyler ve maddeler için ortak bir ölçeğe sahip olmak, maddelerin örtük yapı açısından sağladığı bilgi miktarını değerlendirmeyi ve maddeleri teste giren bireyin yetenek düzeyiyle uyumlu bir şekilde eşleştirmeyi mümkün kılar (Van der Linden ve Glas, 2010). MTK, sunduğu avantajlardan dolayı bireyselleştirilmiş bilgisayarlı testler, test eşitleme, değişen madde fonksiyonu, bilişsel tanı modeli, ölçek geliştirme uygulamalarında aktif olarak kullanılmaktadır.

MTK'nın avantajlarından yararlanmak için kullanılan MTK modeli ile verinin uyumlu olması gerekir (Hambleton ve Swaminathan, 1985). Veri ile kullanılan MTK modeli uyumsuzken kestirilecek parametreler sistematik hatalar içerecektir. Dolayısıyla model-veri uyumunun sağlanamaması sonucu elde edilecek madde ve yetenek parametrelerinin geçerliği şüpheli hale gelecektir. Madde Tepki Kuramı'nın bir diğer avantajı değişmezlik özelliğidir. Değişmezlik özelliği, MTK'nın Klasik Test Kuramı'na (KTK) tercih edilmesinin nedenidir. Bu özellik, yetenek parametrelerinin bireyin aldığı testten bağımsız olması ve madde parametrelerinin ise bireyin içerisinde yer aldığı gruptan bağımsız olması gerektiğini ifade etmektedir (Lord, 1953; Baker, 2001; Wells ve Hambleton, 2016).

MTK'da madde ve yetenek parametrelerinin kestirimi için büyük örneklemlere ihtiyaç duyulmaktadır. Bu durum MTK'nın bir sınırlılığı olarak görülmektedir. Dolayısıyla MTK ile parametre

kestirimleri için örneklem büyüklüğü önemli bir unsurdur. Fakat pek çok MTK uygulaması küçük örneklemlerle gerçekleştirilmektedir. MTK uygulamalarında parametrelerin doğru bir şekilde kestirilmesi için kullanılan MTK modeline göre minimum örneklem büyüklüğü farklılık göstermektedir. Kullanılan MTK modeli karmaşıklaştıkça daha büyük örneklemler gerekmektedir (Sireci, 1991).

MTK'nın sunduğu avantajlardan faydalanmak için modelin varsayımlarının sınanması ve sağlanması gerekir. Madde Tepki Kuramı'nın varsayımları pek çok farklı kaynakta ele alınmıştır. Lord ve Novick (1968) (a) tek boyutluluk, (b) tüm maddelerin normal ogive modele uyumlu olması ve (c) yerel bağımsızlık ve (d) yetenek (theta) dağılımının normal olması olarak dört başlık altında ele almaktadır. Hambleton ve Swaminathan (1985), MTK'nın varsayımlarını (a) boyutluluk, (b) yerel bağımsızlık, (c) madde karakteristik eğrisi uyumu, (d) hız testi olmaması olmak üzere dört başlık altında ele almaktadır. Crocker ve Algina (1986) MTK'nın varsayımlarını (a) tek boyutluluk, (b) yerel bağımsızlık olmak üzere iki alt başlıkta ele almaktadır. Embretson ve Reise (2000), MTK'nın madde karakteristik eğrisi uyumu ve yerel bağımsızlık olmak üzere iki temel varsayımı olduğunu belirtmektedir. Demars (2010) ise (a) tek boyutluluk, (b) yerel bağımsızlık ve (c) uyum (fit) başlıkları altında ele almaktadır. MTK varsayımlarının sağlanamaması durumunda psikometrik olarak bazı sorunlar yaşanacaktır. Örneğin tek boyutluluk ihlal edilirse, örtük yetenek uzayının çok boyutlu yapısı ile tek boyutlu MTK modeli birebir eşleme yapmayacaktır. Dolayısıyla tek boyutlu MTK ile elde edilen sonuçların bireyler açısından yanlı olabileceği anlamına gelir (Reckase, 2009).

Yerel bağımsızlık varsayımı ise tek boyutluluk ile yakından ilişkilidir (Lord, 1980). Tek boyutluluğun saf bir şekilde sağlanamaması yerel bağımsızlık varsayımının sağlanıp sağlanamadığına şüphe düşürmektedir. Bu şüpheyi ortadan kaldırmak için yerel bağımsızlığın sınanması gerekir. Fakat günümüzde gelişen bilgisayar teknolojisine rağmen araştırmacıların yerel bağımsızlığı sınamak yerine testin tek boyutlu olduğuna atıf vererek bu varsayım sınanmamaktadır.

Bu araştırmada 1995-2021 yılları arasında MTK üzerine yazılmış ve YÖKTEZ arşivinde yer alan yüksek lisans ve doktora tezlerinin model varsayımlarının ve model veri uyumlarının sınanıp sınanmadığının belirlenmesi amaçlanmıştır. Bu bağlamda ele alınan tezler örneklem büyüklüğü, normallik, boyutluluk, yerel bağımsızlık, model veri uyumu, madde uyumu, değişmezlik ve hız testi kriterlerine göre incelenmiştir.

Yöntem

Bu araştırmada literatürde yer alan çalışmalar incelendiğinden doküman analizi yöntemi kullanılmıştır. Doküman analizi; araştırmanın amacı doğrultusunda ele alınan yazılı unsurlardaki bilgilerin ve içeriklerin analiz edilmesine olanak sağlayan sistematik bir prosedürdür (Ary ve diğ., 2010). Bu bağlamda; Türkiye'de 1995-2021 yılları arasında Madde Tepki Kuramı üzerine yazılmış yüksek lisans ve doktora tezleri incelemek için doküman analizi tercih edilmiştir.

Araştırma kapsamında YÖKTEZ arşivinin sitesinde "madde tepki kuramı", "MTK", "item response theory", "IRT" anahtar kelimeleri kullanılarak arama yapılmıştır. Arama sonucunda 1995-2021

yılları arasında yazılan elde edilen 144 tez çalışması çalışma grubu olarak belirlenmiştir. 18 tez çalışması erişimin kapalı olmasından veya sistemde paylaşılmadığından çalışma grubuna alınmamıştır. Ancak bu tezlerin erişime açık olan özet kısımlarından yola çıkarak örneklem büyüklüklerine erişilmiş ve çalışmaya sadece örneklem büyüklüğü bazında dahil edilmiştir. Ayrıntılı olarak incelenen 126 tez çalışması gerçek veriye ve simülatif veriye dayalı olarak iki gruba ayrılmıştır. Simülatif çalışmaların verileri manipülatif olarak üretildiğinden bu çalışmaların varsayımlarının incelenmesi kapsam dışı bırakılmıştır. Araştırmacılar tarafından yapılan literatür taraması sonucunda Madde Tepki Kuramı için önemli görülen sekiz başlık kapsamında (1) örneklem, (2) normallik, (3) boyutluluk, (4) yerel bağımsızlık, (5) model veri uyumu, (6) madde uyumu, (7) değişmezlik ve (8) hız testi olmaması kriterlerine göre incelenmiştir. Dokümanların araştırmacılar tarafından belirtilen kapsamlarda incelendikten sonra veriler betimsel analiz kullanılarak çözümlenmiştir. Elde edilen bulgular frekans ve yüzde kullanılarak Microsoft Excel programı aracılığı ile raporlanmıştır.

Sonuçlar

MTK'nın büyük örneklemler gerektirdiği göz önüne alınarak örneklem büyüklükleri incelendiğinde çalışmaların %24.5'inde küçük örneklemler ile çalışıldığı (n<500) görülmüştür. Tüm varsayımlar arasında en yüksek oranda sınanan varsayımın tek boyutluluk varsayımı olduğu görülmüştür (%92). Yerel bağımsızlık varsayımı çalışmaların %58'inde sınanmamıştır. Birçok çalışmada tek boyutluluk varsayımının sağlanmasının yerel bağımsızlık için kanıt olacağı belirtilmiş ve araştırmacılar tarafından ayrıca bir sınama yapılmaya gerek duyulmamıştır. Model-veri uyumu, tek boyutluluk varsayımından sonra araştırmacıların en çok incelediği koşul olarak göze çarpmaktadır (%80). Normallik (%18), madde uyumu (%19), değişmezlik (%20), ve hız testi (%11), varsayımlarının ise daha düşük oranlarda sınandığı sonucuna ulaşılmıştır. Araştırma sonuçları tablolar halinde aşağıda sunulmuştur.

Tablo 1

Tez Türü Dağılımı

		f	%
Tez türü	Yüksek lisans	45	31.25
	Doktora	99	68.75
Veri	Gerçek	98	68.10
	Simülatif	46	31.90

Tablo 2

Örneklem Büyüklüğü Dağılımı

Örneklem Büyüklüğü	f	%
1-500	24	24.5
501-1000	19	19.4
1001-1500	10	10.2
1501-2000	11	11.2
2001 ve üzeri	34	34.7

Tablo 3

Normallik Dağılımı

Sınanma Durumu	f	%
Evet	15	18.8
Hayır	65	81.2
Toplam	80	100

Tablo 4

Boyutluluk Varsayımı Dağılımı

Sınanma Durumu	f	%
Evet	74	92.5
Hayır	6	7.5
Toplam	80	100

Tablo 5

Yerel Bağımsızlık Varsayımı Dağılımı

Sınanma Durumu	f	%
Evet	33	41.3
Hayır	47	58.7
Toplam	80	100

Tablo 6

Model Veri Uyumunun Sınanma Durumu Dağılımı

Sınanma Durumu	f	%
Evet	64	80.0
Hayır	16	20.0
Toplam	80	100

Tablo 7

Madde Uyumunun Sınanma Durumu Dağılımı

Sınanma Durumu	f	%
Evet	14	17.5
Hayır	66	82.5
Toplam	80	100

Tablo 8

Değişmezlik Özelliğinin Sınanma Durumu Dağılımı

Sınanma Durumu	f	%
Evet	16	20.0
Hayır	64	80.0
Toplam	80	100

Tablo 9

Hız Testi Varsayımının Sınanma Durumu Dağılımı

Sınanma Durumu	f	%
Evet	9	11.3
Hayır	71	88.8
Toplam	80	100

Kaynaklar

Ary, D., Jacobs, L. C., Sorensen, C., & Razavieh, A. (2010). *Introduction to research in education* (8th ed.). Wadsworth Cengage Learning.

Baker, F. B. (2001). *The basics of item response theory.* ERIC Clearinghouse on Assessment and Evaluation. http://ericae. net/irt/baker.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Cengage Learning.

DeMars, C. (2010). *Item response theory.* Oxford University Press.

Embretson, S. E., and Reise, S. P. (2000). *Item response theory for psychologists.* Psychology Press.

Hambleton, R. K., & Swaminathan, H. (1985). A look at psychometrics in the Netherlands. *Nederlands Tijdschrift voor de Psychologie en haar Grensgebieden, 40*(7), 446–451.

Hambleton, R. K., Shavelson, R. J., Webb, N. M., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and psychological measurement, 13*(4), 517-549. https://doi.org/10.1177/001316445301300401

Lord, F.M., Novick, M.R., & Birnbaum, A. (1968). *Statistical theories of mental test scores.* Addison-Wesley.

Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional item response theory* (pp. 79-112). Springer.

Sireci, S. G. (1991, October). *"Sample-Independent" Item Parameters? An investigation of the stability of IRT item parameters estimated from small data sets.* [Paper presentation]. At the Annual Conference of the Northeastern Eduational Research Association, Ellenville, New York. https://files.eric.ed.gov/fulltext/ED338707.pdf

van der Linden, W. J., and Glas, C. A. W. (Eds.) (2010). *Elements of adaptive testing.* Springer.

Wells, C. S., & Hambleton, R. K. (2016). Model fit with residual analyses. In W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 2) (pp. 395-413). Cahpman & Hall.

Açımlayıcı faktör analizi bağlamında bir ayırt edicilik katsayısı önerisi ve diğer madde ayırt edicilik indeksleri ile karşılaştırılması

Eren Can Aybek

Giriş

Test geliştirme süreci, ölçülmek istenilen özelliğin tanımlanmasıyla başlayan ve geliştirilen testle güvenilir ve geçerli ölçmeler yapılıp yapılmadığına dair kanıtlar toplanan bir süreçtir. Bu sürecin ayrılmaz parçalarından biri de madde analizidir. Klasik test kuramı bağlamında sınır yeterliklerin ölçülmesinin amaçlandığı araçlarda madde güçlüğü ve madde ayırt ediciliği gibi istatistikler hesaplanırken, tipik tepkilerin ölçülmesinde madde güçlüğü, yerini madde onaylanma oranına bırakmakta ve madde ayırt ediciliği ile birlikte hesaplanmaktadır. Madde ayırt ediciliği, bir maddenin ölçülmek istenen özellik bakımından bu özelliğe az ya da çok sahip olan bireyleri ne derece ayırdığına odaklanmaktadır. Bu amaçla da alt – üst %27'lik gruplar arasında t-testi, madde ve toplam puan ya da madde ve kalan madde toplam puanları arasında hesaplanan korelasyon katsayılarından yararlanılmaktadır (Crocker ve Algina, 2008). Buna göre toplam puanı düşük olan bireylerin maddeye yanlış yanıt ya da daha olumsuz bir tepki vermesi beklenirken; toplam puanı yüksek olan bireylerin maddeye doğru yanıt ya da daha olumlu bir tepki vermesi beklenmektedir. Ancak farklı madde yanıt ve toplam puan örüntüleri ile aynı korelasyon katsayısının elde edilmesi mümkündür. Bu durumda, ayırt edici olmayan maddelerin de faktör analizine dahil edilme olasılığı ortaya çıkmaktadır. Bu araştırma kapsamında çok kategorili puanlanan maddeler (ör: Likert tipi) için madde – toplam puan saçılım grafiği ile elde edilen doğrunun eğiminin hesaplanmasına yönelik yeni bir ayırt edicilik katsayısının diğer madde ayırt edicilik katsayılarıyla karşılaştırılması ve bu katsayılara göre faktör analizine dahil edilecek maddelere karar verilmesi halinde kaç maddenin analize dahil edildiği ve açıklanan toplam varyansın karşılaştırılması amaçlanmıştır.

Yöntem

Araştırma için veriler R (R Core Team, 2021) yazılımında *catR* (Magis ve Barrada, 2017; Magis ve Raiche, 2012) paketine ait *genPolyMatrix* fonksiyonu kullanılarak üretilmiştir. catR paketi, verileri Madde Tepki Kuramı (MTK)'na dayalı olarak üretmektedir. Her ne kadar bu araştırma Klasik Test Kuramı (KTK)'na göre yürütülmüş olsa da madde havuzu ve yanıt örüntüsü üretiminde sağladığı kolaylık nedeniyle bu paket tercih edilmiştir.

Veri üretiminde hem madde havuzu hem de örneklem büyüklüğü kontrol edilmiştir. Buna göre madde havuzu büyüklüğü sırasıyla 10, 30, 50 ve 100; örneklem büyüklüğü ise sırasıyla 50, 100, 250, 500 ve 1000 olarak atanmıştır. Buna göre dört farklı madde havuzu ve beş farklı örneklem olmak üzere simülatif olarak toplam 20 farklı yanıt örüntüsü üretilmiştir. Yanıt kategorisi ise 5 olarak seçilmiştir. catR ile madde parametreleri MTK'ya göre üretildiği için maddelerin ayırt edicilik indeksleri oldukça yüksek olmaktadır. Bunun önüne geçmek için üretilen maddelerin %30'una ait ayırt edicilik *a* parametreleri düzeltilmiştir. Bunun için ortalaması 0.3 ve standart sapması 0.1 olan bir dağılımdan tesadüfi olarak bir *a* parametresi atanmıştır. Daha sonra her bir madde havuzu için belirtilen örneklem büyüklüğünde yanıt örüntüsü *genPattern* fonksiyonu kullanılarak üretilmiştir. Sonuç olarak, araştırmada farklı madde havuzu ve örneklem büyüklüklerine sahip toplam 20 veri dosyası ile çalışılmıştır.

Araştırmada, simülatif verilerle çalışılmış ve madde ayırt ediciliği için beş farklı katsayı hesaplanmıştır. Bu ayırt edicilik katsayıları: Madde – toplam, madde – kalan korelasyon katsayıları, alt-üst %27 gruplar yöntemi ve eğim katsayısıdır. Araştırmanın özünü oluşturan eğim katsayısı *s* son kategorinin sayısal değeri, *i* ilk kategorinin sayısal değeri olmak üzere $(\bar{X}_s - \bar{X}_i) / (s - i)$ formülü ile hesaplanmıştır. Eğim katsayısı dışında kalan indeksler *ShinyItemAnalysis* (Martinkova ve Drabinova, 2018) paketi içerisinde yer alan *ItemAnalysis* fonksiyonu kullanılarak hesaplanmıştır. Elde edilen bulguların grafik haline getirilmesinde *ggplot2* (Wickham, 2016) paketinden yararlanılmış, ayrıca *psych* (Revelle, 2020) paketi yardımıyla açımlayıcı faktör analizi gerçekleştirilmiştir.

Sonuçlar

Her bir madde havuzundaki maddelerin %30'unun madde ayırt edicilik indeksleri bilinçli olarak düşürülmüştü. Bu durumun yansıması grafiklerde de görülebilmektedir. Bunun yanında, maddelerin %70'i için tüm madde havuzu büyüklüklerinde madde-toplam ve madde-kalan korelasyon katsayılarının .40'tan büyük olduğu görülmüştür. Alt-üst %27 gruplar yöntemiyle de çoğu durumda .40'tan büyük değerler elde edilmiştir. Öte yandan eğim katsayısı kullanıldığında, diğer yöntemlerle .40 üzeri elde edilen madde ayırt edicilik indeksleri .30 - .40 arasında bulunmaktadır. Madde ayırt edicilik indeksi düşük olması beklenen son %30'luk gruptaki maddeler içinse tüm yöntemlerin benzer şekilde düşük madde ayırt edicilik indeksi bulduğu, ancak bu durumda da eğim katsayısının diğerlerinden daha düşük sonuçlar verdiği görülmüştür.

Üretilen veri setlerindeki her bir madde için madde ayırt edicilik indeksleri dört yöntemle hesaplanmıştır. Daha sonra sırasıyla .20, .35 ve .40 ölçütleri kullanılarak, bu ölçütün altında madde ayırt edicilik indeksine sahip olan maddeler testten çıkarılarak kalan maddeler ile açımlayıcı faktör analizi yapılmıştır ve kalan madde sayıları ile tek faktörün açıkladığı varyans miktarları her bir durum için incelenmiştir. Buna göre eğim katsayısına göre daha fazla sayıda maddenin faktör analizinden çıkarıldığı, buna karşın açıklanan toplam varyansın %8-9 civarında daha fazla olduğu bulunmuştur.

Kaynaklar

Crocker, L. ve Algina, J. (2008). *Introduction to classical and modern test theory.* Cengage Learning.

Magis, D., and Barrada, J.R. (2017). Computerized Adaptive Testing with R: Recent Updates of the Package catR. *Journal of Statistical Software, Code Snippets*, *76*(1), 1-19. https://.doi.org/10.18637/jss.v076.c01

Magis, D. ve Raiche, G. (2012). Random Generation of Response Patterns under Computerized Adaptive Testing with the R Package catR. *Journal of Statistical Software*, *48*(8), 1-31. https://doi.org/10.18637/jss.v048.i08

Martinková, P., & Drabinová, A. (2018). ShinyItemAnalysis for Teaching Psychometrics and to Enforce Routine Analysis of Educational Tests. *The R Journal*, *10*(2), 503-515. https://doi.org/10.32614/RJ-2018-074

R Core Team (2021). R: A language and environment for statistical computing. https://www.R-project.org/.

Revelle, W. (2020) *psych: Procedures for personality and psychological research (version 2.1.3).* https://cran.r-project.org/package=psych

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag.

PISA 2018 okuma becerileri testinde değişen madde fonksiyonunun incelenmesi

Evrim Yalçın, Şerife Zeybekoğlu ve Ayşe Bilicioğlu

Anahtar kelimeler: PISA, Mantel-Haenzsel, lojistik regresyon, SIBTEST, Lord ki-kare ve Raju'nun alan ölçümleri

Giriş

Yetenek düzeyi aynı olan fakat farklı alt gruplarda yer alan bireylerin bir maddeye ya da teste verdikleri tepkiler farklılaşabilmektedir. Bu durum, maddenin ya da testin farklı alt gruplardaki bireylerde, farklı fonksiyon göstermesine neden olmaktadır. Öte yandan testlerin tüm gruplarda eşdeğer bir biçimde çalıştığından emin olunması oldukça önemlidir. Aksi bir durum, test puanlarının belli bir gruba karşı sistematik hata içermesi olarak tanımlanan yanlılık kavramına karşılık gelmektedir (Camilli ve Shepard, 1994). Zumbo (1999), yanlılığın ölçme aracından elde edilen sonuçların geçerliğini etkileyen önemli tehditlerden biri olduğunu ifade etmiştir. Bir maddenin yanlı olması halinde, maddenin doğru yanıtlanması ölçülen yetenekten ziyade herhangi bir gruba ait olmaya bağlı olmaktadır (Osterlind, 1983). Bu bağlamda, ülkemizin de dahil olduğu PISA, TIMSS ve PIRLS gibi geniş ölçekli testler göz önünde bulundurulduğunda, yanlılık çalışmalarının yürütülmesi oldukça kritik görünmektedir. Sonuçları ile ülkelerin eğitim sistemlerine yön vermek amacıyla uygulanan uluslararası sınavların en düşük düzeyde hata barındırması beklenmektedir. Bu yönüyle PISA uygulamalarında yer alan maddeler üzerinde yürütülecek yanlılık çalışmaları, çıktıların değerlendirilmesine büyük ölçüde katkı sağlayacaktır. PISA, 15 yaş grubundaki öğrencilerin sadece okulda öğrendiklerini ne ölçüde hatırlayabildiğini değil, bunları günlük yaşamlarında kullanabilme düzeylerini, yeni bir durumla karşılaştıklarında sorunları çözebilmek, bilmedikleri bir konuda tahmin yürütebilmek ve yargıda bulunabilmek için bilgi ve becerilerini ne ölçüde kullanabildiğini belirlemeyi amaçlayan her üç yılda bir gerçekleştirilen bir geniş ölçekli sınavdır. Her döngüsünde okuma becerileri, matematik okuryazarlığı ve fen okuryazarlığı alanlarından biri ağırlıklı alan olarak belirlenmektedir. Son PISA uygulaması 2018 yılında gerçekleştirilmiş olup okuma becerileri ağırlıklı alan olarak seçilmiştir. Bu nedenle araştırma kapsamında PISA 2018 uygulamasında Okuma Becerileri testinde yer alan maddelere ilişkin yanlılık çalışmalarının yürütülmesi planlanmıştır.

Yanlılık çalışmaları, değişen madde fonksiyonu (DMF) belirleme yöntemlerinin kullanılmasıyla başlayan ve uzman görüşünün alınması ile devam eden süreci içerir. Yanlılığın istatistiksel olarak

manidarlığının görülebilmesi amacıyla DMF belirleme yöntemleri kullanılır. Ardından ise DMF içeren madde ya da maddelerdeki farklı fonksiyonlaşmanın kaynağının madde etkisi olarak adlandırılan gruplar arası gerçek farktan mı; yanlılıktan mı kaynaklandığının belirlenmesi amacıyla uzman görüşlerine başvurulur.

Literatürde, DMF'nin belirlenmesi amacıyla kullanılabilecek çok sayıda yöntem bulunmaktadır. İki ya da çok kategorili verilerin kullanılması; iki ya da daha fazla sayıda grubun bulunması halinde kullanılacak DMF yöntemleri farklılaşmaktadır. Ancak yöntemleri en genel haliyle Klasik Test Kuramına (KTK) ve Madde Tepki Kuramına (MTK) dayalı olarak iki sınıfta toplamak mümkündür. KTK'ya dayalı olarak geliştirilen DMF belirleme yöntemlerinin sık kullanılanlarından bazıları; Mantel-Haenszel (MH), Varyans Analizi, Dönüştürülmüş Madde Güçlüğü, SIBTEST, Lojistik Regresyon (LR) olarak sıralanabilir. MTK'ya dayalı tekniklerin sık kullanılanlarından bazıları ise; Olabilirlik Oran Testi, Raju'nun Alan Ölçümleri ve Lord'un χ^2 'dir (Camilli ve Shepard, 1994; Hambleton ve diğ., 1991). Çalışma kapsamında ise iki gruplu veriler için MH, LR, SIBTEST ve Raju'nun Alan Ölçümleri yöntemlerinin kullanılması planlanmıştır. Öte yandan üç gruplu değişkenler için ise Genelleştirilmiş MH, Genelleştirilmiş LR ve Genelleştirilmiş Lord'un χ^2 yöntemlerinin kullanılması uygun bulunmuştur.

Bu çalışma kapsamında, PISA 2018 uygulaması Türkiye örneklemi okuma becerileri testinde yer alan maddeler üzerinde farklı yöntemler kullanılarak DMF çalışmalarının yürütülmesi amaçlanmaktadır.

Bu amaç doğrultusunda ise aşağıdaki alt problemlere yanıt aranmıştır:

1. PISA 2018 uygulamasındaki okuma becerileri testinde yer alan maddeler, cinsiyete göre MH, LR, SIBTEST ve Raju'nun Alan Ölçümleri yöntemlerine göre DMF göstermekte midir?
2. PISA 2018 uygulamasındaki okuma becerileri testinde yer alan maddeler, sosyoekonomik düzeye göre MH, LR, SIBTEST ve Raju'nun Alan Ölçümleri yöntemlerine göre DMF göstermekte midir?
3. PISA 2018 uygulamasındaki okuma becerileri testinde yer alan maddeler, okulun bulunduğu yerleşim bölgesine göre Genelleştirilmiş MH, Genelleştirilmiş LR ve Genelleştirilmiş Lord'un χ^2 yöntemlerine göre DMF göstermekte midir?

Yöntem

PISA 2018 uygulaması Türkiye örneklemi okuma becerileri testinde yer alan maddeler üzerinde farklı yöntemler kullanılarak DMF belirleme çalışmalarının yürütülmesinin amaçlandığı bu araştırma, betimsel bir araştırma özelliği göstermektedir.

Araştırmada, 2018 PISA uygulamasına katılan 6890 Türk öğrenciye ait okuma becerileri testi ile öğrenci ve okul anketlerinden elde edilen veriler kullanılmıştır. Kullanılan verilere OECD PISA internet sitesinden erişilmiştir. Verilerin düzenlenmesine, Türkiye örneklemi dışındaki verilerin setten çıkarılmasıyla başlanmıştır. Okuma becerileri testinde bulunan ortak köklü maddelerden, kısmi

puanlanan maddeler, araştırmanın dışında tutulmuştur. PISA 2018 uygulamasında, öğrenci başarısını daha doğru bir şekilde ölçmek amacıyla bireyselleştirilmiş testler kullanılmıştır. PISA 2015 ve önceki uygulamalarda kullanılan kitapçıklardaki sorular sabit bir yapıya sahip; başka bir ifadeyle kitapçıklardaki soruların yeri önceden belirlenmişken bu uygulamasında, öğrencilerin önceki sorulara vermiş olduğu tepkilerin doğruluğuna göre belirlenen dinamik bir yapı kullanılmıştır (OECD, 2019). Bu bağlamda, veriler üzerinde düzenleme yapılırken kitapçıkların ve öğrencilerin karşılaştıkları ortak maddeler tespit edilerek 161 öğrencinin cevapladığı 25 madde üzerinden araştırma yürütülmüştür. Verilerin analizinde, ilk olarak kuramlara ait varsayımlar sınanmıştır. Bu bağlamda her ne kadar KTK'ya dayalı DMF yöntemleri için tek boyutluluk gereğinin sınanması yeterli görünse de MTK'ya dayalı yöntemler düşünüldüğünde, tek boyutluluğun yanı sıra yerel bağımsızlığın ve model- veri uyumunun da sınanması gerekmektedir. Yapının tek boyutluluğu faktör analizi ile sınanırken; yerel bağımsızlık varsayımı için Yen'in Q_3 istatistiğinden yararlanılmıştır. 0.20'nin altında çıkan değerler, yerel bağımsızlığın sağlandığının bir göstergesi olarak kabul edilmiştir (DeMars, 2010). Model- veri uyumunun sınanmasında ise modellerin oluşturulmasının ardından bir sonraki adım olan modeller arası farkın incelenmesine ve model- veri uyumunun değerlendirilmesine geçilmiştir. İkili model karşılaştırılmaları ANOVA istatistiği kullanılarak yapılmıştır. Anlamlı fark bulunan modellere ilişkin Akaike'nin bilgi ölçütü, Bayes bilgi ölçütü ve -2Log-likelihood değerleri incelenmiştir. Değerlendirmelerin sonucunda ise her bir ölçme aracından elde edilen verilerin 2PL modele uygun olduğu sonucuna ulaşılmıştır. Varsayımların sağlanmasının ardından araştırma problemlerine yanıt oluşturulmaya çalışılarak R yazılımı aracılığıyla, iki gruplu veriler için MH, LR, SIBTEST ve Raju'nun Alan Ölçümleri yöntemleriyle; üç gruplu değişkenler için ise Genelleştirilmiş MH, Genelleştirilmiş LR ve Genelleştirilmiş Lord'un χ^2 yöntemleri kullanılarak DMF analizleri yapılmıştır.

Sonuçlar

Yapılan araştırma sonucunda, cinsiyet değişkeni bağlamında KTK tabanlı yöntemlerden MH ve LR çıktıları benzerlik göstererek aynı maddede DMF bulunmuştur. İlgili madde, MH ile yüksek düzeyde (C) DMF gösterirken; LR ile orta düzeyde (B) DMF göstermiştir. SIBTEST yöntemine göre yapılan analizler sonucunda ise farklı bir maddenin cinsiyet değişkenine göre yüksek düzeyde (C) DMF içerdiği belirlenmiştir. MTK tabanlı yöntemlerden biri olan Raju'nun Alan Ölçümleri yöntemi sonucunda ise diğer yöntemlerin aksine çok sayıda maddenin DMF gösterdiğine ulaşılmıştır.

Araştırmada kullanılan sosyoekonomik düzey değişkenine göre yapılan analizler sonucunda ise SIBTEST, MH ve LR yöntemleri aynı maddede DMF göstermiştir. DMF'nin düzeyleri incelendiğinde, MH ve SIBTEST yöntemine göre yüksek düzeyde (C); LR yöntemine göre orta düzeyde (B) DMF içerdiğine ulaşılmıştır. Ek olarak, cinsiyet değişkenine benzer şekilde burada da Raju'nun Alan Ölçümleri yöntemi ile çok sayıda maddede DMF tespit edilmiştir. Okulun bulunduğu bölge değişkenine yönelik yapılan analizler sonucunda ise Genelleştirilmiş MH ile Genelleştirilmiş LR yöntemi bulguları benzerlik göstererek aynı maddede DMF göstermiştir.

Kaynaklar

Camilli, G. & Shepard, L.A. (1994*). Methods for identifying biased test items.* Sage Publications.

DeMars, C. (2010). *Item Response Theory.* Oxford.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.

OECD (2019). *PISA 2018 results volume I: What students know and can do.* OECD Publishing. https://www.oecd-ilibrary.org/docserver/5f07c754-en.pdf?expires=1644826655&id=id&accname=guest&checksum=1D54472F42236BC07BC78288792EF427

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential ıtem functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores.* Directorate of Human Resources Research and Evaluation, Department of National Defense.

Çok kategorili puanlanan maddelerde değişen adım fonksiyonu belirleme yöntemlerinin incelenmesi[1]

Yasemin Kuzu ve Selahattin Gelbal

Anahtar kelimeler: Değişen adım fonksiyonu, ardışık yaklaşım, kümülatif yaklaşım, AC-LOR, CU-LOR

Giriş

Teknoloji ve iletişim araçlarının hızla gelişmesiyle dünya bir küreselleşme sürecine girmiş ve ülkeler arasında siyasal, toplumsal ve kültürel etkileşim artmıştır. Bu etkileşim modern çağın gerektirdiği becerilere ayak uyduran bireyler yetiştirmeyi daha da önemli kılmaktadır. Ülkelerin aynı ölçüte dayanarak, özellikle eğitim alanında dünyadaki konumunu görme, eğitim politikalarını değerlendirme ve bu politikalara şekil verme amacıyla uluslararası geniş ölçekli sınavlara (PIRLS, PISA, TIMSS vb.) katılım her geçen gün artmaktadır. Bu sınavların ve sınavlardan elde edilen puanların amacına hizmet etmesi büyük önem taşımaktadır. Testler, ölçülen yapıyla ilgisi olmayan değişkenleri dikkate almaksızın tüm bireyler için aynı doğrulukla ölçme yapmalıdır (Sireci ve Rios, 2013). Yanlılık, testteki bir maddeye aynı yetenek düzeyindeki bireyler tarafından verilen yanıtların ölçülmek istenen özellikten farklı sebeplere bağlı olarak değişkenlik göstermesidir. Bu durum ölçme sonuçlarına ölçülmek istenen özellik dışında başka değişkenlerin (cinsiyet, din, ırk vb.) karıştığı anlamına gelmektedir. Psikolojik ölçmelerde bu unsurların tamamen kontrol altına alınması imkânsızdır. Ancak bazı istatistiksel yöntemler yardımıyla belirlenen gruplardan herhangi birine avantaj ya da dezavantaj oluşturmasının önüne geçmek mümkündür.

Bir testte yanlılık içeren maddelerin tespitinde ilk olarak maddelerde değişen madde fonksiyonu (DMF) olup olmadığı belirlenmelidir. DMF, farklı alt gruplarda yer alan aynı yetenek düzeyindeki bireylerin bir maddeyi doğru cevaplama olasılığının gruplara göre farklılık göstermesidir (Embretson ve Reise, 2000; Hambleton ve diğ., 1991). Çok kategorili maddelerde yanıt kategorilerinin sayısı nedeniyle DMF farklı biçimler alabilmektedir. Çok kategorili bir maddede yanıt kategorisi sayısının bir eksiği kadar adım fonksiyonu tanımlanmaktadır. Belirli bir adımda ölçülen özelliklerde görülen gruplar arası fark ise değişen adım fonksiyonu (DAF) olarak tanımlanmaktadır (Penfield, 2007).

[1] Bu çalışma birinci yazarın ikinci yazar danışmanlığında yürüttüğü doktora tezinden üretilmiştir.

DAF analizleri, çok kategorili maddelerde kapsamlı bir DMF analizi için önemli bir bileşendir. Son yıllarda araştırmacılar çok kategorili maddelerde değişmezlik formunu incelerken birçok sebep göstererek, tek bir toplam puan düzeyi yerine her bir puan düzeyinin dikkate alınması gerektiğini belirtmektedirler (Gattamorta ve Penfield, 2012). Buna göre her adım için DAF'nin hesaplanması önemli bilgilerin fark edilmesini sağlayacaktır. Ayrıca böyle bir yaklaşımla değişmezliğin ihlalinden hangi puan düzeylerinin sorumlu olduğu anlaşılabilecek, dolayısıyla DMF'nin olası nedenleri hakkında bilgi sahibi olunacaktır.

Bu çalışmanın amacı çok kategorili maddelerde DAF belirleme yöntemlerini farklılaşan koşullarda karşılaştırmaktır. Bu bağlamda PISA 2018 kapsamında öğrencilere yönelik hazırlanan "Bilgi ve İletişim Teknolojilerine Aşinalık Anketi" içerisindeki okulda dijital cihazları kullanma sıklığıyla ilgili maddelerin (IC011) yanıtlandığı Kazakistan, Türkiye ve ABD verileri üzerinde çalışılmıştır. Ülkelerin seçiminde öncelikle PISA 2018 sonuçlarına göre ağırlıklı alanda ülke sıralamaları incelenmiş ve ülkeler düşük, orta ve yüksek düzey olarak üç gruba ayrılmıştır. Aynı zamanda ülkelerin ekonomik durumları da göz önünde bulundurularak, her iki durum (başarı ve ekonomik düzey) için de düzeyi düşük olan gruptan ilgili anketi yanıtlayan Kazakistan, orta düzey gruptan Türkiye ve üst düzey gruptan ABD tercih edilmiştir. DAF analizleri için Ardışık Kategori Log Odds Oranı (AC-LOR) ve Kümülatif Kategori Log Odds Oranı (CU-LOR) yöntemleri kullanılmıştır.

Araştırmanın alt problemleri şu şekildedir:

1. Odak grup örneklem büyüklüğü 200 (küçük) olduğunda, ülke karşılaştırmalarında DAF yöntemleriyle elde edilen DAF miktarları; değişen kategori birleştirme kuralı ve odak grup:referans grup örneklem oranlarında farklılık göstermekte midir?

2. Odak grup örneklem büyüklüğü 1000 (büyük) olduğunda, ülke karşılaştırmalarında DAF yöntemleriyle elde edilen DAF miktarları; değişen kategori birleştirme kuralı ve odak grup:referans grup örneklem oranlarında farklılık göstermekte midir?

3. Ülke karşılaştırmalarında; değişen kategori birleştirme kuralı ve odak grup:referans grup örneklem oranlarında DAF yöntemleriyle elde edilen DMF miktarları; odak grup örneklem büyüklüğüne göre farklılık göstermekte midir?

Yöntem

Çok kategorili DAF belirleme yöntemlerinin çeşitli koşullarda karşılaştırıldığı bu araştırma ilişkisel tarama modelinde bir araştırmadır. Çalışmada PISA 2018 kapsamında öğrencilere uygulanan "Bilgi ve İletişim Teknolojilerine Aşinalık Anketi" içerisindeki okulda dijital cihazları kullanma sıklığıyla ilgili maddelere yanıt veren Türkiye, Kazakistan ve Amerika Birleşik Devletleri (ABD) örneklemleri üzerinde çalışılmıştır. İlgili ülkelerden PISA 2018 uygulamasına katılan birey sayısı Kazakistan için 19.507, Türkiye için 6.890 ve ABD için 4838 olduğu görülmektedir (OECD, 2019) Tüm veri temizleme

işlemleri sonucunda toplamda Kazakistan verisinde 10.991 (%60,63), Türkiye verisinde 3.997 (%22,05) ve ABD verisinde 3.140 (%17,32) kişi üzerinde çalışılmıştır. İncelenen koşullar aşağıda açıklanmıştır.

Kategori birleştirme kuralı. Madde kategori sayısını değiştirmek amaçlandığından kategori birleştirme yoluna gidilmiştir. Birleştirilen kategorilerin birbirine yakın (ardışık gelen) kategoriler olmasına dikkat edilerek kategori birleştirmede olabilecek tüm kombinasyonlar ele alınmış olup bu bağlamda sekiz koşul elde edilmiştir.

Odak grup – Referans grup. Çalışmada iki farklı odak grup- referans grup oluşturulmuştur. Birincisi Türkiye-ABD, ikincisi Türkiye-Kazakistan şeklindedir.

Örneklem büyüklüğü ve odak grup referans grup örneklem oranı. Bu çalışmada odak grup örneklem büyüklüğü 200 (küçük) ve 1000 (büyük) olmak üzere iki farklı büyüklükte ele alınmıştır. Bununla birlikte (odak grup):(referans grup) örneklem oranı 2:1, 1:1, 1:2 ve 1:3 olacak şekilde dört koşulda incelenmiştir.

DAF belirleme yöntemleri. AC-LOR ve CU-LOR

Verilerin analizinde DIFAS 5.0 programından faydalanılmıştır. Her bir maddenin her adımı için hesaplanan DAF değerleri incelenmiştir. Bu bağlamda analiz çıktılarından elde edilen $\hat{\lambda}_j$ değerleri yorumlanarak adımları yüksek DAF gösteren maddeler her iki yöntem için ayrı ayrı işaretlenmiştir. Yüksek DAF gösteren maddelerin işaretlenmesinde $|\hat{\lambda}_j| > 0{,}64$ ölçütü dikkate alınmıştır (Penfield, 2007; Penfield, Alvarez ve diğ., 2008).

DAF analizleri için kritik değer $\pm$0,64 olup, [-(0,64), (0,64)] aralığı dışında kalan değerler madde adımının ilgili koşullarda yüksek düzey DAF sergilediğini göstermektedir. DAF grafiklerinde kritik değerlerden uzaklaşıldıkça DAF düzeyi artmakta olup bu uzaklaşma negatif yönde ise DAF odak grup lehine, pozitif yönde ise referans grup lehinedir.

Sonuçlar

AC-LOR ve CU-LOR yöntemleri kullanılarak yapılan DAF analizi sonuçlarında bazı durumlarda CU-LOR yönteminden elde edilen DAF değerlerinin; bazı durumlarda ise AC-LOR yönteminden elde edilen DAF değerlerinin daha yüksek olduğu görülmüştür. DAF belirleme yöntemleri örneklem büyüklüklerine göre incelendiğinde örneklemin küçük olduğu durumlarda her iki yöntemden elde edilen DAF miktarlarının büyük örnekleme kıyasla daha yüksek olduğu görülmüştür.

Örneklem büyüklüğü arttıkça kümülatif yaklaşım altında kullanılan CU-LOR ve ardışık kategoriler yaklaşımında kullanılan AC-LOR yöntemlerinin, madde adımlarını DAF açısından sınıflamadaki benzerlik oranları artmaktadır. DAF belirleme yöntemlerine ilişkin sonuçlar örneklem büyüklüğü oranları ve koşullar bazında incelendiğinde; aynı örneklem büyüklüğü oranında bazı maddelere ilişkin DAF miktarlarında artış görülürken bazı maddelerin DAF miktarlarında azalma

görülmektedir. Dolayısıyla örneklem büyüklüğü oranlarının DAF sonuçlarına önemli bir etkisi olmadığı düşünülmektedir.

Kategori birleştirme kuralının DAF analizleri üzerinde sistematik bir etkisi görülmemiştir. Bununla birlikte işaretlenme sıklıkları yüksek olan kategorilerin birleştirmesiyle oluşturulan koşullardan elde edilen sonuçların diğer koşullardan elde edilen sonuçlardan farklı olduğu söylenebilir.

Kaynaklar

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates, Inc.

Gattamorta, K. A., & Penfield, R. D. (2012). A comparison of adjacent categories and cumulative differential step functioning effect estimators. *Applied Measurement in Education, 25*(2), 142–161.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.

Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of educational measurement, 44*(3), 187–210.

Penfield, R. D., Alvarez, K., & Lee, O. (2008). Using a taxonomy of differential step functioning to improve the interpretation of DIF in polytomous items: An illustration. *Applied Measurement in Education, 22*(1), 61–78.

Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation, 19*(2-3), 170-187.

Kernel eşitleme yöntemlerinin karşılaştırılması: TIMSS 2019 Fen testi örneği

Şeyma Nur Özsoy ve Sevilay Kilmen

Anahtar kelimeler: Test eşitleme, kernel eşitleme, NEAT deseni, NEC deseni, ortak değişken

Giriş

Ülkemizde ve uluslararası çapta (TIMSS ve PISA gibi) yapılan sınavlarda birden çok kitapçığa sahip testler ya da yılda birkaç kez yapılan sınavların (ALES ve YDS gibi) her birinde farklı kitapçıklardan oluşan testler yer almaktadır. Mevcut araştırmada, bu testlerin farklı kitapçıklara sahip olması nedeniyle ortaya çıkan olumsuz psikolojik etkiler ve adaletsizlik gibi sorunları önlemek amacıyla Kernel eşitleme yöntemleri kullanılmıştır. Çünkü gerçek puan eşitleme, örneklem büyüklüğü, yerel bağımsızlık gibi varsayımları içerir ve pratikte bu koşulları sağlamak zordur. Henüz yeni bir yaklaşım olan Kernel eşitleme ise diğer yöntemlere göre daha gerçekçi varsayımlara sahiptir (Godfrey, 2007). Diğer yandan, test eşitleme çalışmaları, genelde ortak madde kullanımını gerektirmektedir. Bu durumda çalışma, NEAT deseninde ilerlemektedir. Ortak madde olmadığında ise ulaşımı daha kolay olan ortak değişkenlerin de kullanımıyla NEC deseninde eşitleme yapılabilir. Ortak değişkenlerle Kernel eşitleme üzerine literatürde çok az araştırma vardır (örneğin Wiberg ve Branberg, 2015; Albano ve Wiberg, 2019). Yapılan araştırmalar incelendiğinde yaş, cinsiyet, okul türü, eğitim durumu gibi değişkenlerin ele alındığı görülmektedir. Bu çalışmalardan bazılarında benzer sonuçlara ulaşılması nedeniyle ortak değişkenlerin ortak maddeler yerine kullanılabileceği, bazılarında ortak maddelerin daha iyi sonuçlar verdiği ve bazılarında da ortak değişken kullanımının daha iyi sonuçlar verdiği sonucuna ulaşılmıştır. Hatta ortak madde ve ortak değişkenin birlikte kullanımına yönelik çalışmalar da vardır. Ancak görüldüğü gibi birlik sağlanamamıştır. Mevcut çalışmada, her testte ortak madde bulunmaması durumu da göz önünde bulundurulmuş olup ortak madde ve ortak değişkenler (cinsiyet ve bilgisayar/tablet sahibi olma) kullanılarak Kernel eşitleme yöntemleriyle yapılan eşitlemelerin incelenmesi amaçlanmış olup aşağıdaki sorulara cevap aranmıştır.

1) Test formları NEAT desenine göre Kernel sontabakalama eşit yüzdelikli, Kernel zincirleme eşit yüzdelikli, Kernel sontabakalama doğrusal ve Kernel zincirleme doğrusal eşitleme yöntemleriyle eşitlendiğinde eşitleme hatası nasıl değişmektedir?

2) Test formları NEC desenine göre ortak değişken kullanılarak Kernel eşit yüzdelikli ve Kernel doğrusal eşitleme yöntemleriyle eşitlendiğinde eşitleme hatası nasıl değişmektedir?
3) NEAT ve NEC desenleri kullanılarak yapılan eşitlemeler sonucunda elde edilen ham puanlar ile eşitlenmiş puanların farkı ve eşitleme hataları arasında farklılık var mıdır?

Yöntem

Çalışmada TIMSS 2019 Türkiye örnekleminde, sekizinci sınıf fen testinin 1 ve 14 numaralı kitapçıkları kullanılmıştır. Verilen kitapçıkları sırasıyla 288 ve 295 öğrenci cevaplandırmıştır. Ancak bu çalışmada, değişkenlerle eşitleme yapıldığından mevcut değişkenlere ait maddeleri cevaplandırmayan öğrenciler analizden çıkarılmıştır. Dolayısıyla çalışma grubunu, sırasıyla 284 ve 293 olmak üzere 577 öğrenci oluşturmaktadır.

Mevcut çalışmada NEAT ve NEC desenleri kullanılmıştır. Eşitleme için NEAT deseninde, 17 iç ortak madde kullanılırken NEC deseninde, cinsiyet ve bilgisayar/tablet sahibi olma değişkenleri kullanılmıştır. Diğer yandan, çalışmada ele alınan 1.kitapçığı 43 madde ve 14.kitapçığı 39 madde oluşturmaktadır. Veriler, 1-0 şeklinde ikili puanlanan maddeler haline getirilmiştir. Bunun için doğru (correct), kısmi kredi (partial credit) ve tam kredi (full credit) cevapları 1 puan olarak kodlanırken boş bırakılan ya da yanlış cevaplar 0 puan olarak kodlanmıştır. Bu çalışmada, güvenirlik katsayıları ve istatistiksel değerlerin hesaplanması amacıyla SPSS programı kullanılmıştır. Kernel eşitleme için de R programı (R Core Team, 2013) aracılığıyla "kequate" paketi (Andersson ve diğ., 2013) kullanılmıştır. Kernel eşitleme süreci, NEAT ve NEC desenleri için benzerdir.

İlk aşamada loglineer modellerle öndüzgünleştirme yapılmıştır. İkinci aşamada, düzgünleştirilmiş puan dağılımları kullanılarak puan olasılık dağılımları kestirilmiştir. Bu aşamada NEAT ve NEC desenleri için farklılıklar bulunmaktadır. NEAT deseninde, zincirleme ve sontabakalama eşitleme yapılabilmekteyken NEC deseninde yapılamamaktadır. Ayrıca sontabakalama ve zincirleme eşitleme için de farklılıklar vardır. Zincirleme eşitlemede iki ayrı tek grup deseni oluşturulup, birinci test formu ortak maddelere, ortak maddeler de diğer test formuna bağlanmaktadır. Sontabakalama eşitlemede ise iki grup birleştirilerek hedef evren oluşturulmakta ve marjinal dağılımlara ulaşılmaktadır. Üçüncü aşamada, kesikli puan dağılımlarını süreklileştirmek amacıyla Gauss Kernel kullanılmıştır. Dördüncü aşamada ise sürekli puan dağılımları arasında h parametresinin ideal ya da geniş olmasına bağlı olarak doğrusal ve eşit yüzdelikli eşitleme yapılmıştır. Bu çalışmada, h parametresi "kequate" paketi tarafından belirlenmiştir. Son olarak ise SEE değeri hesaplanmıştır.

Sonuçlar

Bu çalışmada eşitlemenin standart hatası (SEE) incelenerek hangi Kernel eşitleme yönteminin daha az hata ürettiği incelenmiş olup aşağıdaki sonuçlara ulaşılmıştır.

1. NEAT deseninde en düşük hata, Kernel sontabakalama eşit yüzdelikli eşitleme yöntemlerinden elde edilmiştir.

2. NEC deseninde en düşük hata, bilgisayar/tablet sahibi olma değişkeni aracılığıyla yapılan Kernel doğrusal eşitleme yönteminden elde edilmiştir.
3. NEC deseni altında cinsiyet ve bilgisayar/tablet sahibi olma değişkenlerinin ayrı ayrı kullanıldığı Kernel eşitleme sonuçları, birlikte kullanıldığı Kernel eşitleme sonuçlarına kıyasla daha az hata vermiştir.
4. Genel olarak, Kernel doğrusal eşitleme sonuçları, Kernel eşit yüzdelikli eşitleme sonuçlarına kıyasla daha az hata ortaya koymuştur.
5. NEAT ve NEC desenleri genel olarak karşılaştırıldığında, NEAT desenindeki Kernel eşitleme sonuçları, NEC desenindeki Kernel eşitleme sonuçlarına kıyasla daha düşük hata vermiştir.

Araştırmanın sonuçları göz önünde bulundurulduğunda NEC deseninden elde edilen SEE değerlerinin NEAT desenindeki SEE değerlerine göre daha yüksek olduğu görülmüştür. Bu durumda araştırmanın genel sonucu olarak şunlar söylenebilir: Ortak değişkenler, ortak maddeler yerine kullanılamamaktadır. Ortak maddeye ulaşılamaması durumunda ortak değişkenlerle eşitleme yapılabilir. Ancak en iyi eşitleme yolu, ortak madde kullanılarak yapılan eşitlemedir.

Kaynaklar

Albano, A. D., & Wiberg, M. (2019). Linking with external covariates: examining accuracy by anchor type, test length, ability difference, and sample size. *Applied Psychological Measurement*, *43*(8), 597-610. https://doi.org/10.1177/0146621618824855

Andersson, B., Branberg, K., & Wiberg, M. (2013). Performing the kernel method of test equating with the package kequate. *Journal of Statistical Software*, *55*(6), 1-25. https://doi.org/10.18637/jss.v055.i06

Godfrey, K. E. (2007). *A comparison of kernel equating and IRT true score equating methods* (Publication No. 3273329) [Doctoral Dissertation, The University of North Carolina]. ProQuest Dissertations & Theses Global.

R Core Team. (2013). *R: A language and environment for statistical computing* (versiyon 4.0.3) [Computer software]. R Foundation for Statistical Computing.

Wiberg, M., & Branberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement, 39*(5), 349-361. https://doi.org/10.1177/0146621614567939

Pozitif ve negatif ruminasyon ölçeğinin Türk kültürüne uyarlanması ve geçerlik güvenirlik çalışması

Tuğba Aksoy ve Sevilay Kilmen

Anahtar kelimeler: Pozitif ruminasyon, negatif ruminasyon, geçerlik, güvenirlik

Giriş

Kişinin duygu düzenlemesi için kullandığı olumlu ve olumsuz stratejik tepkiler olarak karşımıza çıkan ruminasyon kavramı, bireyin kendini gerçekleştirebilmesi için bir bütün olarak gelişimini ele alan eğitim stratejilerinde, bireyin psikolojik uyumu açısından son derece önemlidir. Gençler arasında yaşama karşı doyumsuzluk, depresyon gibi duygular giderek arttığı için kişinin sadece olumsuz yönlerine değil pozitif özelliklerine de yönelerek, potansiyelini yükseltmek ve psikolojik uyum sağlaması için pozitif ve negatif düşüncelerinin ölçülebilmesi gerekmektedir. Bu sebeple Yang ve diğ. (2018) tarafından geliştirilen Pozitif ve Negatif Ruminasyon Ölçeği'nin (PANRS) Türkçe formunun geçerlik ve güvenirlik çalışmasının yapılması amaçlanmıştır. Yang ve diğ. (2020) ruminasyonu en iyi şekilde açıklamak için hem olumlu hem de olumsuz yönlerini dikkate alarak çok boyutlu ruminasyon ölçümü yapılmasını savunmuşlardır.

Alan yazın incelendiğinde kaygı, endişe, stres, üzüntü, öfke, obsesyon, mükemmeliyetçilik, erteleme davranışı, olumsuz otomatik düşünceler ve depresyon gibi kavramlarla ilişkilendirilerek ruminasyon kavramının psikolojik uyumda önemli bir role sahip olduğu görülmektedir. Bu sebeple ruminasyonu ölçmek için birçok ölçek geliştirilmiştir. Bu ölçeklerden Türk kültürüne uyarlananlara bakıldığında; Kişinin ruminasyonu ile ilgili olumlu inanışlarını ölçmeyi amaçlayan Rumi-Olumlu ve Rumi-Olumsuz Ölçekleri depresyonla ilgili üstbiliş çalışmalarına katkı sağlamaktadır (Papageorgiou ve Wells, 2001). Ruminatif Düşünce Biçimi Ölçeği genel ruminasyon eğilimini ölçmekte fakat olumlu ya da olumsuz yönleri ayırt etmemektedir (Brinker ve Dozois, 2009).

Kişiler Arası Hataya ilişkin Ruminasyon (Wade ve diğ., 2008) ve Yasa Bağlı Ruminasyon Ölçekleri (Boelen ve diğ., 2003) dışsal nedenlerden kaynaklanan ve durumluluk ile ilgili ruminasyon seviyesini ölçmektedir. Eşli Ruminasyon ölçeğinde (Rose, 2002) iki kişi arasında olumsuz durumlarla ilgili karşılıklı konuşmalarla yapılan ruminasyon ölçülür. Kısa Durum Ruminasyon Envanteri (Marchetti ve diğ., 2018) hem iç hem dış uyaranlara yanıt olarak durumluluk ruminasyon düzeyindeki değişiklikleri ölçmeyi amaçlamaktadır, ölçek olumsuz duygulanım ve depresyonla ilişkilidir. Öz eleştirel ruminasyon ölçeği de

kişinin utanç ya da suçluluk gibi duygularla özüne odaklanarak kendisini sürekli eleştirmesini ölçmeyi amaçlamaktadır.

Ruminatif Tepkiler Ölçeği (RTÖ) içsel nedenlerden kaynaklanan, olumsuz duygularla ilgili sürekli ruminasyonu ölçer (Treynor ve diğ., 2003). Pozitif Duygulanıma Verilen Tepki Ölçeği (PDVTÖ) ise uyumlu veya uyumsuz işlev gören 3 faktöre sahiptir. Sadece olumlu duygulanıma verilen tepkilere odaklanır. Türk kültürüne uyarlaması yapılan ölçekler incelendiğinde ya olumsuz duygulara karşı uyumlu ya da uyumsuz ruminasyonu (RTÖ, RDÖ), ya da olumlu duygulara karşı uyumlu ya da uyumsuz ruminasyonu (PDVTÖ) ölçmeyi amaçlamaktadır. Fakat ikisini birden yapan bir ölçme aracı bulunmamaktadır. PANRS, olumlu duygulanıma karşı pozitif ve negatif ruminasyonu aynı zamanda olumsuz duygulanıma karşı pozitif ve negatif ruminasyonu ölçmeyi amaçlayarak tüm özellikleri tek bir çatı altında toplamaktadır. Ruminasyonun iki duygu durumu içinde olumlu ve olumsuz yönlerini ayırt etmektedir. Aynı zamanda psikolojik uyum açısından uyumsuz ya da işlevselliğini de ortaya koymaktadır. Ölçek ikinci düzey iki faktörden oluşan 23 maddeden oluşmaktadır. Görece az madde ile işlevselliğinin çok olması PANRS'ı olumlu ve olumsuzu ayırmayan diğer ölçeklerin önüne geçirmektedir. Ölçeğin, ülkemizde bu alanda yapılacak kesitsel, boylamsal ya da deneysel çalışmalarda kullanılabileceği düşünülmektedir. Bu sebeple PANRS'ın Türk kültürüne kazandırılması için Türk lise öğrencileri örnekleminde geçerlik güvenirlik çalışmasının yapılması amaçlanmıştır. Bu genel amaç altında belirlenen alt problemler:

1) PANRS Türk kültürü için geçerli bir ölçme aracı mıdır?
2) PANRS güvenilir bir ölçme aracı mıdır?

Yöntem

Çalışma grubunu 2020-2021 eğitim-öğretim yılının bahar döneminde Türkiye' de Batı Karadeniz'de yer alan orta ölçekli bir il merkezinin ortaöğretim kurumlarındaki öğrenciler oluşturmaktadır. Araştırma için gerekli izinler alınarak, uygulama süresi yaklaşık otuz dakika olan üç ölçek katılımcılara çevrimiçi olarak uygulanıp, google-form aracılığı ile veriler toplanmıştır.

Araştırmada Treynor ve diğerleri (2003) tarafından geliştirilen, Erdur-Baker ve Bugay (2012) tarafından Türkçeye uyarlanan RTÖ; Feldman ve diğerleri (2008) tarafından geliştirilen Yüksel (2014) tarafından Türkçeye uyarlanan PDVTÖ, Yang ve diğerleri (2020) tarafından geliştirilen, Hambleton ve Patsula'nın (1998) belirlerdiği adımlar izlenerek Türkçe formu hazırlanan PANRS kullanılmıştır.

Verilerin normal dağılım gösterip göstermediğini test etmek için çarpıklık basıklık katsayıları hesaplanmıştır. Ölçeklerin yapı geçerliklerini ve örnekleme uyumlarını test etmek amacıyla doğrulayıcı faktör analizleri (DFA) yapılmıştır. DFA, korelasyon ve asymptotik kovaryans matrisi kullanılarak ağırlıklandırılmamış en küçük kareler yöntemi ile yapılmıştır. DFA'da faktör yapıları arasındaki uyum dereceleri gözlemlenen verilerle Karşılaştırmalı Uyum İndeksi (CFI ≥ .90), Standartlaştırılmış Ortalama Hataların Karekökü (SRMR ≤ .08) ve Yaklaşık Hataların Ortalama Karekökü (RMSEA ≤ .08), Non-

normed Fit Index (NNFI ≥ .90) aracılığıyla değerlendirilmiştir (Kline, 2015). Gözlenen değişkenlere ait t değerlerinin manidarlık düzeyleri ve hata varyansları kontrol edilmiştir, maddelerin t değerleri 1.96'yı geçiyorsa 0.05, 2.56'yı geçiyorsa 0.001 düzeyinde anlamlı olduğu sonucuna ulaşılır (Şimşek, 2007). Araştırmada kikare (χ^2) değeri, örneklem büyüklüğünden etkilendiği için bilgilendirme amacı ile kullanılmıştır.

Yakınsak ve ayırt edici geçerlik çalışmasında daha önce geçerlik ve güvenilirliği kanıtlanmış ölçme araçları ile ilişkilerine bakılmıştır (Baykul, 2015). Araştırmada ölçeklerin iç tutarlılığını göstermek için alfa ve omega güvenirlik katsayısı kullanılmıştır. Güvenirlik katsayısını değerlendirmek için.60 (Cohen, 1992) ya da madde sayısı az olan ölçekler için .50 (Raines-Eudy, 2000) ölçüt olarak kullanılmıştır. Ölçeklerin iç tutarlıklarına kanıt için madde-toplam korelasyonlarına bakılmış ve değerlendirmek için 0.20 (Ebel, 1965, aktaran, Erkuş, 2016) ölçüt olarak kullanılmıştır.

Sonuçlar

Pozitif ve Negatif Ruminasyon Ölçeği'nin ikinci düzey dfa sonucu oluşan uyum iyiliği değerleri ($\chi^2_{(223)}$ = 1159.79; χ^2/sd = 5.20, CFI = .94; GFI = .97; NFI = .93; NNFI = .94; RMSEA = .064; SRMR = .053) olarak hesaplanmıştır. Model veri uyum değerleri istenen kriterleri karşıladığı (Kline, 2015) için ölçeğin iki ikinci dereceden faktör yapısı doğrulanmış, yapı geçerliği kanıtlanmıştır. Daha önceden geçerliği ve güvenirliği kanıtlanmış ölçeklerle karşılaştırılarak anlamlı ilişkiler elde edilmiş, yakınsak ve ayırt edici geçerliği kanıtlanmıştır.

PANRS' ın alt ölçeklerinin güvenirliği için Cronbach's alfa ve McDonald's ω katsayıları hesaplanmış ve istenen kriterleri sağladığı görülmüştür (Büyüköztürk, 2009; Raines-Eudy, 2000). Maddelere ait hesaplanan madde-toplam puanları, istenen kriterleri karşılayarak iç tutarlılık kanıtı sağlamıştır. Ölçekten elde edilen bulgularla PANRS'ın Türk kültüründe kullanılabilecek geçerlik ve güvenirliğe sahip olduğu bulunmuştur.

Kaynaklar

Baykul, Y. (2015). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması* (3. baskı). Pegem Akademi.

Boelen, P. A., van den Bout, J., & de Keijser, J. (2003). Traumatic grief as a disorder distinct from bereavement-related depression and anxiety: A replication study with bereaved mental health care patients. *American Journal of Psychiatry*, *160*(7), 1339-1341. https://doi.org/10.1176/appi.ajp.160.7.1339

Brinker, J. K., & Dozois, D. J. (2009). Ruminative thought style and depressed mood. *Journal of clinical psychology*, *65*(1), 1-19. https://doi.org/10.1002/jclp.20542

Büyüköztürk, Ş. (2009). *Sosyal bilimler için veri analizi el kitabı* (10. baskı). Pegem Akademi.

Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, *112*(1), 155-159. https://doi.org/10.1037//0033-2909.112.1.155

Erdur-Baker, Ö., & Bugay, A. (2012). The Turkish version of the ruminative response scale: An examination of ıts reliability and validity. *The International Journal of Educational and Psychological Assessment*, *10*(2), 1-16.

Erkuş, A. (2016). *Psikolojide ölçme ve ölçek geliştirme-1 (Temel kavramlar ve işlemler)* (3. baskı). Pegem Akademi.

Feldman, G. C., Joormann, J., & Johnson, S. L. (2008). Responses to positive affect: A self-report measure of rumination and dampening. *Cognitive Therapy and Research*, *32*(4), 507-525. https://doi.org/10.1007/s10608-006-9083-0

Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Social Indicators Research*, *45*(1), 153-171. https://doi.org/10.1023/A:1006941729637

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (40. baskı). Guilford Press.

Marchetti, I., Mor, N., Chiorri, C., & Koster, E. H. W. (2018). The brief state rumination inventory (BSRI): Validation and psychometric evaluation. *Cognitive Therapy and Research*, *42*(4), 447-460. https://doi.org/10.1007/s10608-018-9901-1

Papageorgiou, C., & Wells, A. (2001). Positive beliefs about depressive rumination: Development and preliminary validation of a self-report scale. *Behavior therapy*, *32*(1), 13-26.

Raines-Eudy, R. (2000). Using structural equation modeling to test for differential reliability and validity: An empirical demonstration. *Structural Equation Modeling*, *7*(1), 124-141. https://doi.org/10.1207/S15328007SEM0701_07

Rose, A. J. (2002). Co–Rumination in the friendships of girls and boys. *Child Development*, *73*(6), 1830-1843. https://doi.org/10.1111/1467-8624.00509

Şimşek, Ö. F. (2007). *Yapısal eşitlik modellemesine giriş-Temel ilkeler ve lisrel uygulamaları* (1. Baskı). Ekinoks Yayınevi.

Treynor, W., Gonzalez, R., & Nolen-Hoeksema, S. (2003). Rumination reconsidered: A psychometric analysis. *Cognitive Therapy and Research*, *27*(3), 247-259. https://doi.org/10.1023/A:1023910315561

Wade, N. G., Vogel, D. L., Liao, K. Y.-H., & Goldman, D. B. (2008). Measuring state-specific rumination: Development of the rumination about an ınterpersonal offense scale. *Journal of Counseling Psychology*, *55*(3), 419-426. https://doi.org/10.1037/0022-0167.55.3.419

Yang, H., Wang, Z., Song, J., Lu, J., Huang, X., Zou, Z., & Pan, L. (2020). The positive and negative rumination Scale: Development and preliminary validation. *Current Psychology*, *39*(2), 483-499. https://doi.org/10.1007/s12144-018-9950-3

Yüksel, B. (2014). *Kaygı belirtilerini açıklamada bağlanma, pozitif ve negatif duygu düzenleme ve belirsizliğe tahammülsüzlük arasındaki ilişkiyi bütünleyici model arayışı* (Tez No. 368995) [Yüksek Lisans Tezi, Hacettepe Üniversitesi]. Yükseköğretim Kurumu Tez Merkezi.

A methodological review of problem statement and method sections in PISA studies

Özen Yıldırım and Safiye Bilican Demir

Keywords: PISA, large-scale tests, document analysis, data reporting

Introduction

Programme for International Student Assessment (PISA) is a large-scale testing that has been conducted by the OECD (Organization for Economic Co-operation and Development) for nearly two decades and evaluates 15-year-olds' ability to use their reading, mathematics and science knowledge and skills to meet real-life challenges (OECD, 2019). In each cycle of the testing process, which is carried out in three-year periods, a field is predominantly selected from the math, science and reading and detailed data are collected about the cognitive characteristics of the student, as well as the characteristics such as affective characteristics like family and school environment. Considering the scope of PISA, it offers a comprehensive and rigorous data source that is easily accessible for researchers at national and international level.

There is a large literature on the PISA application. Even a simple academic search with "PISA" key word yields more than one and a half million results. (https://scholar.google.com/scholar?hl=tr&as_sdt=0%2C5&q=PISA&btnG=). This shows the interest of educational researchers in PISA. The fact that PISA data is open access to all researchers increases this interest. While using PISA data, researchers discuss the impact and results of PISA on education policies in detailed reviews and can bring criticism based on them (e.g., Andrews, 2015; Araujo et al. , 2017; Sjøberg & Jenkins , 2020). In addition, there are many technical details about the use of these comprehensive and rich PISA datasets that are shared publicly. It is emphasized in the reports that these technical details should be strictly followed in research (OECD, 2009). For instance, sample selection in large-scale testing applications involves a complex process. Analyzing the data for these reasons requires the expertise of the researcher in method. Traditional statistical analyzes that assume that the data are selected according to the simple random sampling method cannot reach accurate results in PISA. Therefore, before using large scale testing databases, it is necessary to know the design very well and to use the correct quantitative methodology. Incorrect methodologies affect the validity of results (e.g., Rutkowski et al. 2010). In this context, there are research findings (e.g., Özdemir, 2016; Takayama, 2015) that reveal the mistakes made in the secondary analysis of PISA data sets.

In the study, it is aimed to reveal how accurately and effectively the PISA datasets are used in international publications and to determine the current research trends based on the problem situation in these publications. For this purpose, the problem situation and method of the selected articles were examined, and answers were sought to the following questions:

1. What is the number of articles using PISA data in the last five years?
2. What is the problem situation focused on in the articles?
3. Is information about the sample type and sample weights reported in the articles?
4. Is information about plausible values reported in the articles?
5. Which statistical analysis methods are used in the articles, are these techniques suitable for the purpose and sample structure, and have the relevant assumptions been tested before the analysis?
6. If cross-country or intergroup comparisons have been made, has measurement invariance been tested?
7. Were the software used for the analyzes specified by the authors and which software was used?

Method

In the study, a qualitative research process based on document analysis was carried out. A three-stage was followed to find answers to the research questions. First, the articles that were scanned in the WoS (Web of Science) database (Social Sciences Citation Index (SSCI), Emerging Sources Citation Index (ESCI) and Science Citation Index Expanded) between 2015-2021 and included the keyword "PISA" in their titles determined. Articles are required to be in one of the fields of "Education, Educational Research, Education Scientific Disciplines, Social Science, Special Education", to be Open Access, and to be in English or Turkish.

In the second stage, it was checked whether the PISA dataset was used in the articles. Articles such as technical and final reports, reviews and criticisms were excluded from the analysis. At the last stage, the articles were examined in two sections, both in terms of problem status and method, by using the article review form.

During the development of the form, the consistency between the raters at the same and at the different times was checked and the percentage of agreement was calculated as 78.7% and 88.2% Since the reliability was high, the reviewers coded the articles separately. The article review form is divided into categories and subcategories to answer the research questions. The categories are (1) focused problem, (2) distribution of the number of articles by years, (3) sample characteristics, (4) plausible value, (5) data analysis, (6) measurement invariance, (7) statistical package program

Results

A total of 150 articles were reached in the research. Most of these are research articles. While an increase was observed in the number of research article from 2015 to 2012, a decrease was observed in

review articles. The subjects of reviews can be classified as the education policies of PISA, the discussion of PISA results, and the structure of PISA. The research papers were about factors related to students' PISA performance, technical specifications in PISA development stages, comparison of PISA and other large-scale test results. When the research articles were examined methodically, it was observed that more studies were made with mathematical literacy data and mostly focused on quantitative research. In addition, the number of studies based on student data was high. It was determined that the sample was not reported in some studies, and they mostly did not benefit from the weighted sample. More than half of quantitative research had a hierarchical data structure. However, in some analyzes the structure was ignored and inappropriate statistical techniques were used. The use of plausible value was not taken into consideration in the analysis, instead the average or single plausible value was used. In many studies based on cross-country comparisons, it was observed that measurement invariance based on the structure of the data was not tested. R, Mplus and HLM are mostly used as data software programs.

Resources

Andrews, P. (2015). Mathematics, PISA, and culture: An unpredictable relationship. *Journal of Educational Change*, 16, 251-280. https://doi.org/10.1007/s10833-015-9248-2

Araujo, L., Saltelli, A., and Schnepf, S. V. (2017). Do PISA data justify PISA-based education policy? *International Journal of Comparative Education and Development*, *19*(1), 20-34. https://doi.org/10.1108/IJCED-12-2016-0023

OECD (2009). *PISA data analysis manual.* OECD Publishing.

OECD (2019). *PISA 2018 Results (Volume I): What students know and can do?* OECD Publishing. https://doi.org/10.1787/5f07c754-en

Harvey Goldstein (2004). International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education: Principles, Policy & Practice*, *11*(3), 319-330, https://doi.org/10.1080/0969594042000304618

Liou, P.-Y. & Hung, Y.-C. (2015). Statistical techniques utilized in analyzing PISA and TIMSS databases in science education from 1996 to 2013: A methodological review. *International Journal of Science and Mathematics Education*, *13*(6), 1449-1468

Özdemir, C. (2016). OECD PISA Türkiye verisi kullanılarak yapılan araştırmaların metodolojik taraması. *Eğitim Bilim Toplum*, *14*(56), 10-27.

Rutkowski, L., Gonzalez, E., Joncas, M., and von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, 39, 142–151.

Sjøberg, S., & Jenkins, E. W. (2020). PISA: A political project and a research agenda. *Studies in Science Education*, *58*(1), 1-14.

Takayama, K. (2015). *Has PISA helped or hindered?* Re*flections on the ongoing PISA debate* (The Lecture Series). The HEAD Foundation. https://headfoundation.org/wp-content/uploads/2020/11/thf-papers_Has-PISA-helped-or-hindered_Reflections-on-the-ongoing-PISA-debate.pdf

Genellenebilirlik kuramından elde edilen sonuçlar ile Rasch analizine dayalı genellenebilirlik sonuçlarının karşılaştırılması

Mustafa İlhan, Neşe Güler ve Gülşen Taşdelen Teker

Giriş

Diğer bilim alanlarında olduğu gibi ölçme ve değerlendirme bilim dalında yapılan araştırmalar da genel itibariyle iki kategoriye ayrılmaktadır. Yapılan bilimsel çalışmaların bir kısmı öğretmenlerin ölçme ve değerlendirme yeterliklerinin geliştirilmesi, sınıf içi ölçme ve değerlendirme süreçlerinin iyileştirilmesi, geniş ölçekli testlerin eğitim süreçleri üzerindeki olumsuz etkilerinin azaltılmasına yönelik politikalar üretilmesi gibi uygulamadaki problemlere çözüm bulmaya odaklanmaktadır. Araştırmaların bir kısmı ise ölçme alanına kuramsal açıdan katkı getirmeyi amaç edinmektedir. Ölçme ve değerlendirme bilim dalının gelişimine kuramsal açıdan hizmet eden çalışmalar incelendiğinde genellikle farklı ölçme kuramlarının karşılaştırılmasının amaçlandığı görülmektedir. Bu doğrultuda madde tepki kuramı ile klasik test kuramından elde edilen yetenek kestirimleri ve madde parametreleri arasındaki uyumun incelenmesine yönelik birçok çalışma yapılmıştır (örneğin; Akyıldız ve Şahin, 2017; Çelen, 2008; Çelen ve Aybek, 2013, Doğan ve Tezbaşaran, 2003; Hwang, 2002; Kelecioğlu, 2001; Magno, 2009; Özer-Özkan, 2012). Yine genellenebilirlik kuramı (G kuramı) ile klasik test kuramına ilişkin sonuçların karşılaştırılmasına dönük çok sayıda araştırma yürütülmüştür (örneğin; Goodwin, 2001; Güler ve Gelbal, 2010; Güler ve Taşdelen-Teker, 2015; MacMillan, 2000; Pekin ve diğ., 2018; Yelboğa ve Tavşancıl, 2010; Yıldıztekin, 2014). Bu araştırma da benzer bir düşünceyle ortaya çıkmıştır. Çalışma kapsamında G kuramından elde edilen sonuçların Rasch analizi çıktılarına dayalı olarak hesaplanan varyans yüzdeleri ve genellenebilirlik katsayıları ile karşılaştırılması hedeflenmektedir. FACETS paket programının kullanma kılavuzunda Rasch analizi çıktılarından hareketle varyans yüzdelerinin ve genellenebilirlik katsayılarının nasıl hesaplanacağına dair açıklamalar yer almaktadır (Linacre, 2021, s. 199). Ancak alanyazında ne Rasch analizine dayalı genellenebilirlik uygulamasına ne de G kuramında ulaşılan sonuçların Rasch analizi çıktıları esas alınarak hesaplanan varyans yüzdeleri ve genellenebilirlik katsayıları ile karşılaştırılmasına yönelik bir araştırmaya rastlanmamıştır. Bu durumun çalışmaya özgün bir nitelik kazandırdığı düşünülmektedir. Dolayısıyla araştırmanın ölçme ve değerlendirme alanyazınına katkı sağlaması beklenmektedir.

Yöntem

Bu çalışma mevcut kuramların karşılaştırılmasına yönelik bir araştırmadır. Dolayısıyla, temel bir araştırma özelliği taşımaktadır. Temel araştırmalarda, yeni bir kuramın geliştirilmesi veya var olan kuramların test edilmesi yoluyla bilimin kuramsal yönden ilerletilmesi amacına ağırlık verilmektedir (Kaptan, 1998). Uzmanlar ve bilim adamları tarafından yürütülmesi, genel/evrensel problemler ile ilgilenmesi, kuralların veya ilkelerin formüle edilmesine odaklanması, doğrudan uygulamaya yönelme endişe taşımaması ve ulaşılan sonuçların kişilerden, yerden ve zamandan bağımsız bir şekilde genellenebilir olması temel araştırmaların karakteristik özellikleridir (Jha, 2014).

Bu araştırmada, G kuramı sonuçları ile Rasch analizine dayalı olarak hesaplanan varyans yüzdelerinin ve genellenebilirlik katsayılarının karşılaştırılması amaçladığından çalışmada evren ve örneklem tayinine gerek görülmemiştir. Araştırmada, genellenebilirlik ve Rasch analizlerini yürütmek için ölçme sonuçlarını etkileyen değişkenlik kaynakları ile katılımcı sayıları açısından farklılık gösteren iki ayrı veri seti kullanılmıştır. Kullanılan ilk veri seti *birey* ve *madde* şeklinde iki değişkenlik kaynağı içermektedir. Bu veri seti, 10 maddeden oluşan Likert tipi bir ölçeğe 500 öğrencinin beşli derecelemeye göre verdiği yanıtlardan meydana gelmektedir. G kuramı ile Rasch analizine dayalı olarak hesaplanan varyans yüzdeleri ve genellenebilirlik katsayıları arasındaki tutarlılıkta katılımcı sayısının etkili olabileceği düşünülüp bu veri seti üzerinden birey sayısı açısından farklılık gösteren üç ayrı veri dosyası oluşturulmuştur. Tablo 1'de görüldüğü üzere söz konusu veri dosyalarının ilkinde 100, ikincisinde 300 ve üçüncüsünde 500 katılımcı yer almıştır.

Tablo 1

Birinci Veri Setine Ait Bilgiler

	Veri Dosyaları		
	1	2	3
Birey	100	300	500
Madde	10	10	10

Çalışmada kullanılan ikinci veri seti ise *birey, madde* ve *puanlayıcı* olmak üzere ölçme sonuçları üzerinde etkili olabilecek üç değişkenlik kaynağı içermektedir. Bu veri seti; 350 öğrencinin açık uçlu 10 maddeye verdiği cevaplara dört puanlayıcının dörtlü derecelemeye sahip bir rubriğe göre atadıkları puanlardan oluşmaktadır. Bununla birlikte, iki kurama göre hesaplanan genellenebilirlik katsayıları arasındaki tutarlılığın örneklem büyüklüğünden etkilenip etkilenmediğini test etmek için ilgili veri setinden yararlanılarak birey ve puanlayıcı sayısı bakımından değişiklik gösteren altı ayrı veri dosyası oluşturulmuştur. Bu veri dosyalarının özellikleri Tablo 2'de verilmiştir. Görüldüğü gibi veri dosyalarının tümünde 10 madde bulunmaktadır. Bununla beraber, birey ve puanlayıcı sayıları bakımından farklılıklar söz konusudur.

Tablo 2

İkinci Veri Setine Ait Bilgiler

	Veri Dosyaları					
	1	2	3	4	5	6
Birey	100	100	100	350	350	350
Madde	10	10	10	10	10	10
Puanlayıcı	2	3	4	2	3	4

Çalışmada, G kuramı analizleri EduG 6.0 (Cardinet ve diğ., 2010) paket programında yapılmıştır. Rasch analizi ise FACETS paket programında gerçekleştirildikten sonra elde edilen çıktılar üzerinden Linacre'nin (2021) önerdiği formülden yararlanılarak Microsoft Excel'de varyans yüzdeleri ve genellenebilirlik katsayıları hesaplanmıştır.

Sonuçlar

Birey ve madde şeklinde iki değişkenlik kaynağı içeren ilk veri seti için G kuramından elde edilen sonuçlar ile Rasch analizine dayalı genellenebilirlik sonuçları Tablo 3'te verilmiştir.

Tablo 3

Birinci Veri Setine Ait Çıktılar

Örneklem Büyüklüğü	Değişkenlik Kaynakları	Varyans Yüzdesi		G Katsayısı	
		G Kuramı	Rasch Analizi	G Kuramı	Rasch Analizi
100	B	31.00	38.36	0.83	0.83
	M	7.10	6.62		
	B*M	61.90	52.48		
	A		2.54		
300	B	31.50	39.12	0.84	0.86
	M	7.10	6.11		
	B*M	61.40	52.21		
	A		2.56		
500	B	31.10	38.50	0.83	0.84
	M	5.70	4.70		
	B*M	63.20	54.32		
	A		2.56		

Tablo 3'e bakıldığında birinci veri setindeki her üç veri dosyası için de G kuramına ve Rasch analizi çıktılarına göre elde edilen varyans yüzdeleri arasında kısmen farklılıklar bulunduğu görülmektedir. Buna karşın hesaplanan genellenebilirlik katsayılarının oldukça benzer olduğu dikkat çekmektedir. Tablo 3'te dikkat çeken bir detay, değişkenlik kaynaklarının tümünün birbirleriyle etkileşimi ve diğer hata kaynaklarından gelen artık varyans G kuramı analizlerinde birbirinden ayrıştırılmadan kestirilirken Rasch analizine dayalı genellenebilirlik hesaplamalarında etkileşim ve artık varyansın ayrı ayrı kestirilebilmesidir. Örneğin örneklem büyüklüğünün 100

olduğu veri seti için G kuramı çıktılarında birey-madde etkileşimi ve artık varyansın (B*M,A) toplam varyansın %61.90'ını açıkladığı okunurken Rasch analizine dayalı genellenebilirlik çıktılarında birey-madde etkileşiminin (B*M) toplam varyansın % 52.48'ini, artık varyansın ise toplam varyansın %2.54'ü açıkladığı görülmektedir. Birey, madde ve puanlayıcı şeklinde üç değişkenlik kaynağı içeren ikinci veri seti için G kuramından elde edilen sonuçlar ile Rasch analizine dayalı genellenebilirlik sonuçları Tablo 4'te sunulmuştur.

Tablo 4

İkinci Veri Setine Ait Çıktılar

Örneklem büyüklüğü	Değişkenlik kaynakları	Varyans yüzdesi		G katsayısı	
		G kuramı	Rasch analizi	G kuramı	Rasch analizi
100 birey 2 puanlayıcı	B	37.80	41.71	0.92	0.98
	M	17.20	2.89		
	P	0.60	10.43		
	B*M	17.40	21.56		
	B*P	0.40	5.73		
	M*P	1.30	5.44		
	B*M*P	25.30	7.10		
	A		5.13		
100 birey 3 puanlayıcı	B	35.10	38.33	0.94	0.98
	M	11.20	3.67		
	P	0.00	7.87		
	B*M	9.90	17.31		
	B*P	0.00	9.43		
	M*P	6.10	5.38		
	B*M*P	37.70	13.74		
	A		5.28		
100 birey 4 puanlayıcı	B	28.30	31.24	0.94	0.98
	M	20.80	6.01		
	P	0.00	14.97		
	B*M	9.10	13.04		
	B*P	0.00	9.18		
	M*P	8.30	7.10		
	B*M*P	33.60	13.01		
	A		5.47		
350 birey 2 puanlayıcı	B	38.40	43.99	0.92	0.98
	M	18.40	1.88		
	P	0.80	13.11		
	B*M	14.90	21.15		
	B*P	1.20	5.54		
	M*P	0.70	3.07		
	B*M*P	25.60	6.33		
	A		4.93		

(devam ediyor)

Tablo 4 *(devam)*

Örneklem Büyüklüğü	Değişkenlik Kaynakları	Varyans Yüzdesi		G Katsayısı	
		G Kuramı	Rasch Analizi	G Kuramı	Rasch Analizi
350 birey 3 puanlayıcı	B	31.40	35.89	0.93	0.98
	M	17.00	4.14		
	P	0.00	12.06		
	B*M	11.70	15.83		
	B*P	0.00	9.64		
	M*P	7.40	6.42		
	B*M*P	32.50	10.78		
	A		5.22		
350 Birey 4 Puanlayıcı	B	25.80	30.10	0.94	0.96
	M	24.50	5.19		
	P	0.00	18.90		
	B*M	10.00	12.46		
	B*P	0.00	9.37		
	M*P	9.30	7.16		
	B*M*P	30.40	11.48		
	A		5.35		

Birey ve madde değişkenlik kaynaklarını içeren veri seti ile karşılaştırıldığında birey, madde ve puanlayıcı şeklinde üç değişkenlik kaynağından oluşan veri setinde, G kuramına ve Rasch analizi çıktılarına göre ulaşılan varyans yüzdeleri ile genellenebilirlik katsayıları arasında daha belirgin farklar saptanmıştır. Ancak G kuramına ve Rasch analizi çıktılarına göre hesaplanan varyans yüzdeleri ile genellenebilirlik katsayıları arasında veri setlerindeki değişkenlik kaynaklarının içerdiği bileşen sayısı (örneklem büyüklüğü) ile ilişkilendirilebilecek bir örüntü gözlenmemiştir.

İki değişkenlik kaynaklı ilk veri setinde olduğu gibi üç değişkenlik kaynaklı ikinci veri setinde de Tablo 4'te görüleceği üzere etkileşim ve artık varyans değerleri G kuramı analizlerinde ayrıştırılmadan kestirilirken Rasch analizine dayalı genellenebilirlik hesaplamalarında etkileşim ve artık varyansın ayrı ayrı kestirilmiştir. Örneğin örneklem büyüklüğünün 100 puanlayıcı sayısının 2 olduğu veri seti için G kuramı çıktılarında birey-madde-puanlayıcı etkileşimi ve artık varyansın (B*M*P,A) toplam varyansın %25.30'unu açıkladığı okunurken Rasch analizine dayalı genellenebilirlik çıktılarında birey-madde-puanlayıcı etkileşiminin (B*M*P) toplam varyansın % 7.10'unu, artık varyansın ise toplam varyansın %5.13'ünü oluşturduğu görülmektedir.

Sonuçlar

Literatürde G kuramı analizlerinin EduG, GENOVA, SPSS gibi paket programlarda ya da özellikle son yıllarda R gibi platformlarda yapıldığı görülmektedir. Sıralanan bu programların tümü neredeyse aynı matematiksel alt yapıyı kullanarak varyans bileşenlerini ve genellenebilirlik katsayısını hesaplamaktadır. Aralarındaki tek fark negatif varyans çıktığında yürütülen işlemdir. Bazı programlar Cronbach ve diğ. (1972)'nin önerdiği negatif varyansın sıfır alınması ilkesini benimsemektedir. Bu

programlarda negatif varyans sıfır olarak alınmakta ve sonrasındaki bütün hesaplamalarda ilgili varyans değeri sıfır kabul edilip işlemler yapılmaktadır. Bazı programlar ise Brennan'ın (1983) önerdiği yaklaşımı benimseyerek negatif varyansı sıfır almakta fakat devamında diğer bütün varyans bileşenlerinin hesaplanmasında negatif varyans değerini kullanmaktadır. Negatif varyans ile karşılaşıldığında uygulanan ve sonuçları çok fazla etkilemeyen bu değişiklik dışında programlar temelde aynı formülleri işe koşarak analiz sürecini yürütmektedir. Bu programların yanı sıra FACETS paket programından elde edilen Rasch analizi çıktılarına dayalı olarak varyans yüzdelerinin kestirilmesi ve ardından genellenebilirlik katsayılarının hesaplanması mümkündür (Linacre, 2021). Ancak böyle bir yol izlendiğinde matematiksel olarak yukarıda bahsedilen programlardan biraz daha farklı bir süreç söz konusu olmakta, bu da elde edilen sonuçlarda farklılığa yol açmaktadır.

Öncelikle G kuramı analizlerinde değişkenlik kaynaklarının tümünün birbirleriyle etkileşimi ve diğer hata kaynaklarından gelen artık varyansın birbirinden ayrıştırılması mümkün değildir. Buna karşılık Rasch analizine dayalı genellenebilirlik hesaplamalarında etkileşim ve artık varyansın ayrı ayrı kestirilmesi mümkün olmaktadır. Bu durum belki de analiz sonuçlarında karşımıza çıkan en önemli farklılıktır. Hem iki (birey ve madde) hem de üç (birey, madde ve puanlayıcı) değişkenlik kaynağı içeren desenden elde edilen varyans değerleri kıyaslandığında G kuramı analizlerinde birbirinden ayrıştırılmadan kestirilen etkileşim ve artık varyans değerinin Rasch analizine dayalı kestirilen etkileşim ve artık varyans toplamından hep daha büyük olduğu görülmektedir. Bu durum iki farklı yaklaşımla kestirilen diğer varyans değerlerinde de bir miktar farklılığa sebep olmakla birlikte, özellikle iki değişkenlik kaynağı içeren desende kestirilen genellenebilirlik katsayılarının birbirine çok yakın olduğu görülmüştür. Diğer taraftan üç değişkenlik kaynağı bulunduran desende hem kestirilen varyans değerlerinde hem de genellenebilirlik katsayılarında belirgin farklılıklar göze çarpmaktadır. Puanlayıcılar da bir değişkenlik kaynağı olarak analizde yer aldığında G kuramında ulaşılan sonuçlar ile Rasch analizine dayalı genellenebilirlik sonuçları arasında daha belirgin bir fark çıkması; çok yüzeyli Rasch modelinde puanlayıcılar arası farklılıklara düzeltme uygulanması (Abu Kassim, 2007) ve bahsi geçen düzeltme sayesinde puanlayıcı farklılıklarının kısmen kontrol altına alınması (Linacre ve diğ., 1994) ile bağlantılı olabilir.

Çalışmada ayrıca, iki ve üç değişkenlik kaynağı içeren desenlerde madde sayıları sabit tutulup diğer değişkenlik kaynaklarındaki bileşen sayıları değiştirilmiş ve bu durumun kestirilen varyans değerlerine ve genellenebilirlik katsayılarına etkisine bakılmıştır. İki değişkenlik kaynağı içeren desende birey; üç değişkenlik kaynağından oluşan desende ise birey ve puanlayıcı sayısının değiştirilmesinin elde edilen sonuçlarda belirgin bir örüntüye yol açmadığı gözlenmiştir. Sonuç olarak, ulaşılan bulgularda karşılaşılan farklılıkların programların kullandığı matematiksel süreçlerdeki farklılıktan kaynaklandığı düşünülmektedir. Çalışma kapsamında ele alınan iki veri seti de tümüyle çaprazlanmış desenlerden oluşmaktadır. Ayrıca ele alınan yüzeylerde tesadüfidir. Bu sebeple benzer bir çalışmanın yuvalanmış ve/ya sabit yüzey içeren desenler üzerinde yürütülmesi ileri araştırma önerisi olarak araştırmacılarla paylaşılabilir. Bunun yanı sıra görece daha küçük ve büyük örneklemler üzerinde analizler tekrarlanabilir.

Kaynaklar

Abu Kassim, N. L. (2007, 14–16 June). *Exploring rater judging behaviour using the many-facet Rasch model* [Conference presentation]. Second Biennial International Conference on Teaching and Learning of English in Asia: Exploring New Frontiers (TELiA2), Faculty of Communication and Modern Languages, Universiti Utara Malaysia, Sintok. http://repo.uum.edu.my/3212/1/Noor1.pdf

Akyıldız, M, ve Şahin, M. D. (2017). Açıköğretimde kullanılan sınavlardan klasik test kuramına ve madde tepki kuramına göre elde edilen yetenek ölçülerinin karşılaştırılması. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi*, *3*(4), 141–159.

Brennan, R. L. (1983). *Elements of generalizability theory.* ACT.

Cardinet, J., Johnson, S. & Pini, G. (2010). *Applying generalizability theory using EduG.* Routledge.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* Wiley.

Çelen, Ü. (2008). Klasik test kuramı ve madde tepki kuramı yöntemleriyle geliştirilen iki testin geçerlilik ve güvenilirliğinin karşılaştırılması. *İlköğretim Online*, *7*(3), 758–768.

Çelen, Ü. ve Aybek, E. C. (2013). Öğrenci başarısının öğretmen yapımı bir testle klasik test kuramı ve madde tepki kuramı yöntemleriyle elde edilen puanlara göre karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *4*(2), 64–75.

Doğan, N. ve Tezbaşaran, A. (2003). Klasik test kuramı ve örtük özellikler kuramının örneklemler bağlamında karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 25*, 58–67.

Goodwin, L.D. (2001). Interrater agreement and reliability. *Measurement in Psychical Education and Exercises Science*, *5*(1), 13–34. http://dx.doi.org/10.1207/S15327841MPEE0501_2

Güler, N. ve Gelbal, S. (2010). Açık uçlu matematik sorularının güvenirliğinin klasik test kuramı ve genellenebilirlik kuramına göre incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri*, *10*(2), 989–1019.

Güler N. ve Taşdelen Teker, G. (2015). Açık uçlu maddelerde farklı yaklaşımlarla elde edilen puanlayıcılar arası güvenirliğin değerlendirilmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *6*(1), 12–24. https://doi.org/10.21031/epod.63041

Hwang, D. Y. (2002, 14–16 February). *Classical test theory and ıtem response theory: Analytical and empirical comparisons* [Conference presentation]. Annual Meeting of the Southwest Educational Research Association, Austin, TX.

Jha, A. S. (2014). *Social research methods.* McGraw Hill Education.

Kaptan, S. (1998). *Bilimsel araştırma ve istatistik teknikleri.* Bilim Yayınları.

Kelecioğlu, H. (2001). Örtük özellikler teorisindeki b ve a parametreleri ile klasik test teorisindeki p ve r parametreleri arasındaki ilişki. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *20*, 104–110.

Linacre, J. M. (2021). *A user's guide to FACETS: Rasch-model computer programs.* https://www.winsteps.com/a/Facets-Manual.pdf

Linacre, J. M., Engelhard, G. Jr., Tatum, D. S., & Myford, C. M. (1994). Measurement with judges: Many-faceted conjoint measurement. *International Journal of Educational Research*, *21*(6), 569–577. https://doi.org/10.1016/0883-0355(94)90011-6

MacMillan, P. D. (2000). Classical, generalizability and multifaceted Rasch detection of interrater variability in large sparse data sets. *The Journal of Experimental Education*, *68*(2), 167–190. http://dx.doi.org/10.1080/00220970009598501

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment*, *1*(1), 1–11.

Özer-Özkan, Y. (2012). *Öğrenci başarılarının belirlenmesi sınavından (öbbs) klasik test kuramı, tek boyutlu ve çok boyutlu madde tepki kuramı modelleri ile kestirilen başarı puanlarının karşılaştırılması* (Tez No. 311753). [Doktora Tezi, Ankara Üniversitesi]. YÖK Ulusal Tez Merkezi.

Pekin, Z., Çetin, S. ve Güler, N. (2018). Otizm sosyal beceriler profili ölçeğinde puanlayıcılar arası güvenirliğin farklı kuramlara göre karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 9*(2), 202–215. https://doi.org/10.21031/epod.388590

Yelboğa, A. ve Tavşancıl, E. (2010). Klasik test ve genellenebilirlik kuramına göre güvenirliğin bir iş performansı ölçeği üzerinde incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri*, *10*(3), 1825–1854.

Yıldıztekin, B. (2014). *Klasik test kuramı ve genellenebilirlik kuramından puanlayıcılar arası tutarlılığın farklı yöntemlere göre karşılaştırılması* (Tez No. 363202). [Yüksek Lisans Tezi, Hacettepe Üniversitesi]. YÖK Ulusal Tez Merkezi.

Ölçme ve değerlendirme uzmanlarının iş doyumlarının ve mesleki deneyimlerinin incelenmesi

Selda Örs Özdil ve Esra Kınay Çiçek

Anahtar kelimeler: Ölçme ve değerlendirme uzmanı, iş doyumu, karma yöntem araştırması

Giriş

Bir eğitim sisteminin başarısı hakkında bilgi veren en önemli gösterge öğrenci davranışlarıdır. Eğitim ortamında bir öğrencinin güçlü ya da zayıf yanlarının ya da öğrenme eksikliklerinin belirlenmesi; sınıfın bir ders ya da üniteyle ilgili hangi konuları öğrenemediği, hangi davranışlarının geliştirilmesi gerektiği; sınıftaki çeşitli gruplar arasında çeşitli değişkenler açısından fark olup olmadığı; sınıf düzeyindeki gerekli kazanımların kazanılıp kazanılmadığı ve başarı oranları; öğrencilerin tek tek veya sınıf olarak duyuşsal özelliklerinin durumu ve süreç içindeki değişimi; sınıflar arası başarı, okulun hedeflerine ulaşılıp ulaşılmadığının belirlenmesi gibi durumlarda ölçme ve değerlendirme (ÖD) gereklidir (Erkuş, 2014). Bu nedenle ÖD eğitim sürecinin vazgeçilmez bir öğesidir. ÖD etkinliği ile eğitim sistemin kontrolü sağlanmakta, ilgili kişilere gerekli geri bildirimler verilebilmekte, öğretim süreciyle ilgili güçlükleri çözebilecek önlemler alınarak eğitimin kalitesi artırılmaya çalışılmaktadır. ÖD süreci okullarda öğretmenler tarafından yürütülmekte ancak bazı özel okullar ÖD uzmanı istihdam ederek bu sürecin uzmanlar tarafından koordine edilmesini sağlamaktadır.

İster eğitim alanında ister diğer alanlarda çalışanlar olsun bireyin yaşamının önemli bir kısmını iş hayatı oluşturmaktadır. Çalışanın işteki çevresiyle uyum sağlaması, çalıştığı kurumdan beklentilerinin karşılanması, işini severek yapması, kişinin mutlu olması ve yaşamından haz alması açısından oldukça önemlidir. Başka bir deyişle, bireyin iş doyumunun yüksek olması gerek iş gerek iş dışı yaşamında mutlu ve huzurlu olmasını sağlamaktadır. Spector (1997), basit olarak iş doyumunu insanların işlerinden hoşlanma derecesi, iş doyumsuzluğunu da insanların işlerinden hoşlanmama derecesi olarak tanımlamaktadır. İş doyumu bir çalışanın işinden aldığı zevkin derecesiyle ilgilidir. Doğal olarak iş doyumunun, işin özellikleriyle çalışanın isteklerinin birbirine uyduğu durumda gerçekleştiği ifade edilmektedir (Fidan ve diğ., 2000). Kişinin aldığı ücretten memnun olması, yükselme fırsatı olduğuna inanması, yöneticileri ile olumlu ilişkiler kurması, kurum tarafından sağlanan hak ve ödeneklerden memnun olması, çabalarının takdir gördüğünü düşünmesi, iş yerindeki kuralların gerekliliğine inanması,

iş arkadaşları ile olumlu ilişkiler geliştirmesi, yaptığı işten keyif alması, kurum içinde bilgi akışının iyi ve yeterli olduğuna inanması kişinin yüksek iş doyumunun göstergeleri arasındadır (Spector, 1985).

Türkiye'deki alanyazın incelendiğinde eğitim alanındaki meslek ve iş doyumu araştırmalarının daha çok öğretmen, öğretim elemanı ve okul yöneticisi gruplarıyla yapıldığı (Akın ve Koçak, 2007; Akkamış, 2010; Bayrı, 2006; Cerit, 2014; Gündüz, 2008; İnandı ve diğ., 2013; Koca, 2016; Kocayörük, 2000; Tuzgöl-Dost ve Cenkseven, 2008), okullarda çalışan ÖD uzmanlarıyla ilgili herhangi bir araştırma yapılmadığı belirlenmiştir. Bu bağlamda okullarda, yapılan ÖD çalışmalarında önemli bir etkiye sahip olması gereken, öğretmenlere bu konuda yol gösterebilecek, eğitimin kalitesinin arttırılmasında köprü görevi görebilecek ÖD uzmanlarının iş doyumlarının yüksek olması, okulların eğitim kalitesini yükseltmek açısından önem taşımaktadır. Bu durum göz önüne alındığında, ölçme ve değerlendirme uzmanlarının iş doyumlarını belirlemek, iş doyumu ile doğrudan ilgili olan çalışma koşulları, görev ve sorumlulukları, işteki iletişim gibi faktörleri ortaya koymak önemli görülmektedir. Bu araştırmadan elde edilen sonuçların, özel okullarda çalışan ölçme ve değerlendirme uzmanının görev ve sorumluluklarının belirlenmesi, çalışma koşullarının iyileştirilmesi, meslek saygınlığının sağlanması gibi konularda yapılacak çalışmalara ışık tutması beklenmektedir. Bu araştırmanın amacı, özel okullarda çalışmış ya da çalışmakta olan ölçme ve değerlendirme uzmanlarının iş doyumlarını cinsiyet, yaş, eğitim durumu, iş deneyimi değişkenlerine göre incelenmek ve iş doyumları ile ilgili olduğu düşünülen mesleki deneyim ve çalışma koşullarını ortaya koymaktır.

Yöntem

Bu araştırmada, nitel ve nicel yöntemlerin birlikte kullanıldığı karma yöntem araştırması kullanılmıştır. Karma yöntem araştırma problemine ilişkin daha kapsamlı çözümlemelerin yapılmasına olanak sağlamaktadır (Creswell ve Plano-Clark, 2007). Bu araştırmada karma yöntem desenlerinden "Açıklayıcı Desen" kullanılmıştır. Öncelikle uzmanların iş doyumlarını belirlemek için Spector (1985) tarafından geliştirilen, Yelboğa (2009) tarafından Türk kültürüne uyarlanan İş Doyum Ölçeği uygulanmıştır. İş Doyum Ölçeği, toplam 9 alt boyut ve 36 maddeden oluşmaktadır. Nitel verilerin toplanması amacıyla araştırmacılar tarafından ölçme ve değerlendirme uzmanlarının mesleki deneyim ve çalışma koşulları anketi geliştirilmiştir. İş doyum ölçeğinin alt boyutları da dikkate alınarak ölçme ve değerlendirme uzmanlarının kurumlardaki farklı görev ve sorumluluklarının, yürütülen çalışmaların çeşitliliğinin, özlük haklarının, çalışma koşullarının, varsa yaşadıkları sorunların belirlenmesini sağlayacak anket soruları yazılmıştır. Bu anketin online formu hazırlanarak uzmanlara uygulanmıştır. Bu araştırmanın evrenini Türkiye'de özel okullarda çalışan ya da çalışmış olan ölçme ve değerlendirme uzmanları oluşturmaktadır. Araştırmada, verilerin toplandığı grubun belirlenmesi için amaçsal örnekleme yöntemlerinden kolay ulaşılabilir durum örneklemesi yöntemi kullanılmıştır. Araştırmanın nicel verilerinin toplandığı iş doyum ölçeği 44 uzman, nitel verilerin toplandığı anket ise 22 uzman tarafından doldurmuştur. İş doyum ölçeğinden elde edilen verilere t testi ve varyans analizi yapılarak cinsiyet, yaş, eğitim durumu ve iş deneyimi değişkenlerine göre ölçek puanlarının farklılaşıp farklılaşmadığı incelenmiştir. Ankette verilen yanıtlara ise içerik analizi yapılmıştır. İçerik analizinde

verilerin kodlanması, kategorilerin ve temaların oluşturulması ve bu kategorilere giren verilerin tanımlanıp yorumlanması aşamaları yer almaktadır (Yıldırım ve Şimşek, 2009). İçerik analizinin ilk aşaması verilerin kodlanmasıdır. Bu araştırmada veriler araştırmacılar tarafından oluşturulan kod listesine göre ayrı ayrı kodlanmıştır. Ardından kategoriler ve temalar oluşturulmuştur. Nicel ve nitel çözümlemelerden elde edilen bulgular birbirleriyle ilişkilendirilerek yorumlanmıştır.

Sonuçlar

ÖD uzmanlarının iş doyum ölçeği, "Yükselme olanakları" alt boyutunda memnuniyetsiz oldukları, "Ücret", "Denetim", "Sosyal haklar", "Performansa dayalı ödüllendirme", "İletişim" alt boyutlarında orta düzeyde memnuniyet gösterdikleri; "Çalışma arkadaşları" ve "İşin yapısı" alt boyutunda ise memnun oldukları belirlenmiştir. Ayrıca bu durumun cinsiyet, yaş, eğitim durumu, kıdem değişkenlerine göre farklılaşmadığı belirlenmiştir. İçerik analizi sonucunda, uzmanların yaptıkları işlerin ve organizasyon şemasındaki konumunun kurumdan kuruma değişiklik gösterdiği, uzmanlık alanı görev ve sorumluluklarının bir standardının olmadığı, uzmanların kurumlarda uzmanlık gerektiren işlerin yanı sıra uzmanlık gerektirmeyen ve öğretmen, müdür yardımcısı, memur ve diğer eğitim personelinin işlerinde de görevlendirildikleri, kurumlarda çoğunlukla sınav odaklı çalışmalar yapıldığı, öğretmenlerin soru yazma, üst düzey becerilerin ölçülmesi, test ve madde istatistiklerini yorumlama gibi konularda eksiklikleri olduğu, uzman görüşlerini önemseyen ve çalışmalarını destekleyen yöneticilerin varlığının yanı sıra alanı uzmanlık alanı olarak görmeyen, ölçme ve değerlendirmeyi yalnızca sınav sonucu olarak gören yöneticilerin de bulunduğu belirlenmiştir. Araştırma bulguları, Nartgün (1998) tarafından yürütülen araştırmadan elde edilen, ÖD uzmanlarının görev tanımlarının belirsizliği, yetersiz personel sayısı, personel yeterliklerinin düşük olması, öğretmenlerin ÖD alanında yeterli bilgiye sahip olmamaları ve öğretmenler ile uzmanlar arasında yeterli işbirliğinin bulunmaması bulguları ile uyum göstermektedir.

Kaynaklar

Akın, U. ve Koçak, R. (2007). Öğretmenlerin sınıf yönetimi becerileri ile iş doyumları arasındaki ilişki. *Kuram ve Uygulamada Eğitim Yönetimi*, *51*, 353-370.

Akkamış, O. (2010). *İlköğretim I. ve II. kademe öğretmenlerinin iş tatmini üzerine bir değerlendirme* (Tez No. 249051) [Yüksek lisans tezi, Yeditepe Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Bayrı, H. (2006). *Ortaöğretim kurumlarında çalışan psikolojik danışman/rehber öğretmenlerin iş doyumuna ilişkin görüşlerinin değerlendirilmesi Güneydoğu Anadolu Bölgesi örneği* (Tez No. 204365) [Yüksek lisans tezi, Dicle Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Cerit, Y. (2014). Sınıf öğretmenlerinin iş doyumunun örgütsel kolektivizm ve bireyselcilik ile ilişkisi. *Eğitim ve Bilim*, *39*(173), 55-66.

Creswell, J. W., and Plano Clark, V. L. (2007). *Designing and conducting mixed method research.* Thousand Oaks, CA: Sage Publications.

Erkuş, A. (2014). *Psikolojide ölçme ve ölçek geliştirme-I: Temel kavramlar ve işlemler* (2. baskı). Pegem Akademi.

İnandı, Y., Tunç, B. ve Uslu, F. (2013). Eğitim fakültesi öğretim elemanlarının kariyer engelleri ile iş doyumları arasındaki ilişki. *Eğitim Bilimleri Araştırmaları Dergisi*, *3*(1), 219-238.

Fidan, Y., Ercan, S., Yılmazer, A., ve Şehirli, M. (2016). The effect of gender on job satisfaction: A study on civil servants. *Business & Management Studies: An International Journal, 4*(1), 110-124. https://doi.org/10.15295/bmij.v4i1.149

Gündüz, H. (2008). *İlköğretim okullarında örgütsel iklim ile öğretmenlerin iş doyumu arasındaki ilişki (Gaziantep ili örneği).* (Tez No. 219823) [Yüksek lisans tezi, Gaziantep Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Koca, E. (2016). *Okul yöneticilerinin kişilik özellikleri ile mesleki doyum düzeyleri arasındaki ilişki* (Tez No. 437056) [Yüksek lisans tezi, Marmara Üniversitesi] Yükseköğretim Kurulu Tez Merkezi.

Kocayörük, E. (2000). *Çeşitli değişkenlere göre rehber öğretmenlerin meslek doyumlarının karşılaştırılması.* (Tez No. 94698) [Yüksek lisans tezi, Ankara Üniversitesi] Yükseköğretim Kurulu Tez Merkezi.

Nartgün, Z. (1998). *Özel dershanelerdeki ölçme ve değerlendirme servislerinde çalışan elemanların görevleri ve hizmetlerin etkililiğinin değerlendirilmesi.* (Tez No. 72527) [Yüksek lisans tezi, Ankara Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Spector, P. E. (1985). Measurement of human service staff satisfaction: Development of the job satisfaction survey. *American Journal of Community Psychology, 13,* 693-713.

Spector, P. E. (1997). *Job satisfaction: Application, assessment, causes, and consequences.* SAGE Publications, Inc., https://dx.doi.org/10.4135/9781452231549

Tuzgöl-Dost, M. ve Cenkseven, F. (2008). Öğretim elemanlarının sosyodemografik değişkenlere ve üniversitelerini değerlendirmelerine göre iş doyumları. *Eğitim ve Bilim, 33*(148), 28-39. http://egitimvebilim.ted.org.tr/index.php/EB/article/viewFile/708/124

Yelboğa, A. (2009). Validity and reliability of the Turkish version of the job satisfaction survey (JSS). *World Applied Sciences Journal, 6*(8), 1066-1072. https://www.idosi.org/wasj/wasj6(8)/9.pdf

Yıldırım, A. ve Şimşek, H. (2009). *Sosyal bilimlerde nitel araştırma yöntemleri.* Seçkin Yayıncılık.

Birey uyum indekslerine göre anormal yanıt örüntülerinin belirlenmesi PISA 2018 uygulaması: Amerika Birleşik Devletleri, Birleşik Krallık, Çin, Kazakistan, Kore, Rusya, Türkiye ve Japonya örneği

Esra Kamacı ve Dilara Bakan Kalaycıoğlu

Anahtar kelimeler: Birey uyum indeksleri, anormal yanıt örüntüleri, PISA

Giriş

Birey uyumu, testler veya anketlerde olağandışı yanıtları tespit etmeyi amaçlayan bir dizi teknikten oluşur. Birey uyum indeksleri; uygulayıcıların ayrı ayrı madde yanıt vektörlerinin önceden belirlenmiş bir madde yanıt teorisine uygun olup olmadığını değerlendirmelerine yardımcı olan, anormal (olağandışı) yanıtları tespit etmeyi amaçlayan istatistiklerdir (Tendeiro ve diğ., 2016). Geçerlik, ölçme aracından elde edilen ölçümlerin kullanımlarının ve önerilen yorumlarının uygunluğunun, kuram ve kanıt ile desteklenme derecesidir (Bademci, 2019). Ölçüm raporlarını makul şekilde destekleyebilmek için teknik olarak doğru olmalı ve yapılacak çıkarımları destekleyici açıklamalar içermelidir (Ferrara ve DeMauro, 2006). Veriler anketler veya testler yoluyla toplanırken katılımcılar bazı nedenlerden dolayı geçersiz yanıtlar verebilir bu nedenle eğitimsel ve psikolojik testlerle ilgili çeşitli kılavuzlarda ve yönergelerde, kişi düzeyinde veri kalitesinin kontrol edilmesi önerilir (Olson ve Fremer, 2013). Anormal madde yanıt örüntüsü gösteren bireylere rastlandığında araştırmacı toplam puanlar hakkında yorum yaparken dikkatli olmalıdır. Birey uyum indeksleri, test puanlarının geçerliliğini kontrol etmek için kullanılabilir.

PISA, OECD tarafından 15 yaş grubundaki öğrencilerin belirli alanlarda kazandıkları bilgi ve becerileri değerlendiren ve üçer yıllık döngülerle yapılan uluslararası bir araştırmadır (MEB, 2019). Öğrencilerin öğrenme çıktılarının bugüne kadarki en kapsamlı uluslararası değerlendirmesi olarak kabul edilen PISA uygulaması testi alan öğrencilerin yanıt örüntülerini de içermektedir. 2018 PISA uygulaması matematik alt testi 24 ayrı form üzerinden gerçekleştirilmiştir. Sorular ve soruların formlardaki sıralamaları aynı olup bu çalışmada ilk 5 form analize dâhil edilecektir. Belirlenen ülkelerdeki öğrencilerin yanıt örüntülerinden çeşitli birey uyum indeksleri hesaplanacak ve anormal yanıt örüntülerinin olası nedenleri açıklanmaya çalışılacaktır. Meijer ve Sitjtsma (2001) ve Rupp (2013)'a göre birey uyum analizleri beş adımda gerçekleştirilmektedir;

1. İstatistiksel tespit adımı; kullanılan yanıt verileri için en az bir birey uyum istatistiği hesaplanır. Çalışmada PISA 2018 verilerinden belirlenen ülkelerde testi alan bireylerin yanıt verileri için birey uyum istatistiklerinden H^t, U3 ve C^* puanları hesaplanacaktır.

2. Sayısal tablo adımı; hesaplanan istatistikler sonucu tespit edilen birey sayısının ülkelere göre tablosu yapılarak bütünsel bakış sağlanır.

3. Grafiksel keşif adımı; Anormal yanıt örüntüsü tespit edilen bireylerin ve normal yanıt örüntüsü gösteren bireylerin madde yanıt grafikleri çizilerek yanıt davranışları görselleştirilir.

4. Nicel keşif adımı; Anormal yanıt örüntüsü gösteren bireyler ile görüşmeler yapılarak sesli düşünme protokolleri uygulanır.

5. Nitel açıklama adımı; literatür incelendiğinde bireylerin anormal yanıt örüntüsü sergilemelerinin nedenleri; hile yapma, testi motive bir şekilde almama, dikkatsiz yanıt verme (doğru yanıt verebileceği halde bireyin dikkatsiz davranması sonucu maddeyi yanlış yanıtlaması), okuma becerilerinin eksikliği ve yorumlama sorunları, şanslı tahmin (maddenin yanıtını bilmeden maddenin doğru yanıtlaması), yaratıcı yanıt verme (özellikle beceri düzeyleri yüksek bireylerin kolay sayılabilecek bir maddeyi yaratıcı bir şekilde farklı yorumlayarak yanlış yanıtlamaları), rastgele yanıt verme (bireyin seçeneklerden birini rastgele seçmesi) ve geride kalmadır (bireyin testi vaktinde tamamlayamaması). Bu tür davranışlar genellikle anormal, tutarsız veya beklenmeyen madde yanıt örüntülerine neden olur (Meijer, Niessen ve Tendeiro, 2016).

Nicel keşif adımı ve nitel açıklama adımı uygulamanın türüne bağlı olarak gerçekleşmesi olası olduğu durumlarda tercih edilirken bu çalışma için OECD'nin uygulamayı alan öğrencilerin kimliklerini paylaşmamasından dolayı gerçekleştirilemeyecektir.

1. Farklı ülkelerden PISA 2018 matematik testini alan öğrencilerin anormal yanıt örüntüsü gösterme miktarı ülkelere göre değişiklik göstermekte midir?
2. Hesaplanan birey uyum indeks değerlerinin ortalamaları ülkelere göre farklılık göstermekte midir?
3. Hesaplanan indeks değerlerinin sonuçları birbiriyle uyumlu mudur?

Yöntem

Araştırma, gerçek veri setinden elde edilen bulgulara dayalı olarak mevcut durumun belirlenmesi yönüyle betimsel araştırma özelliği taşımaktadır. Gruba dayalı parametrik olmayan birey uyum indeksleri, genellikle bireylerin doğru cevaplanma oranı sıfıra yakın olan maddeleri doğru cevaplandırmaya eğilimli ve doğru cevaplanma oranı bire yakın olan maddeleri yanlış cevaplandırmaya eğilimli olduğu durumlarda cevap örüntüleri anormal şeklinde nitelendirir (Meijer ve diğ., 2016) Araştırmada Guttman ölçeğine dayalı parametrik olmayan indekslerden ikili puanlama yapılmış maddelerde kullanılan ve çeşitli simülasyon çalışmalarında iyi performans gösteren indekslerden H^t, U3

ve C* indeksleri (Karabatsos, 2003; Sijtsma, 1986; Sijtsma Meijer, 1992) hesaplanacaktır. Hesaplanan her bir indeksin sonuçlarına göre anormal cevap örüntüsüne sahip bireylerin oranı ülke bazında belirlenecektir.

PISA 2018 araştırmasına katılan 79 ülke oluşturmaktadır. Örneklem olarak PISA 2018 araştırmasına katılan Amerika Birleşik Devletleri, Birleşik Krallık, Çin, Kazakistan, Kore, Rusya, Türkiye ve Japonya olmak üzere 8 ülke belirlenmiştir.

PISA 2018 matematik uygulaması 79 ülke katılmış olup bu çalışma belirlenen sekiz ülke ile sınırlıdır. Ayrıca uygulama 24 ayrı form üzerinden gerçekleştirilmiştir ancak bu çalışmada ilk 5 form analize dâhil edilecektir. Çalışmada kullanılan H^t, U3 ve C* indeksleri ikili puanlanan maddelere uygulanmakta olup incelenen cevap örüntüleri sadece iki puanlanan maddelerle sınırlıdır.

Öncelikle belirlenen sekiz ülkeye uygulanan beş ayrı forma ait ikili puanlanan maddeler seçilerek cevap verileri R programına aktarılmıştır. Bu maddeler için R programı PerFit (Tendeiro, 2016) paketi yardımıyla birey uyum istatistiklerinden H^t, U3 ve C* indeksleri hesaplanacaktır. Yapılan hesaplamaların ardından anormal yanıt örüntüsü gösteren öğrenciler tespit edilecek ve ülke bazında indeks puan ortalamaları verilecektir. Ayrıca, ülkelerin anormal yanıt örüntüsü gösteren öğrencilerinin oranı varyans analizi ile karşılaştırılacaktır.

Yapılan hesaplamalar sonucunda birey uyum analizi sürecinin üçüncü adımı olan grafiksel keşif adımının gerçekleşmesi için anormal yanıt örüntüsü gösteren bireylerin yine R programı yardımıyla PRF (madde yanıt grafiği/person response functıon plot) grafikleri çizdirilecektir. Son olarak H^t, U3 ve C* indeksleri arasındaki uyum oranları hesaplanacaktır.

Sonuçlar

Elde edilen birey uyum indeksleri sonucunda anormal yanıt örüntüsü gösteren bireylerin incelenen 5 farklı form için ülke bazında benzer sonuçlar üretmesi beklenmektedir. Elde edilen bulgular araştırma soruları altında incelendiğinde "PISA 2018 testini alan ve anormal yanıt örüntüleri gösteren öğrencilerin miktarı farklı ülkelere göre değişmekte midir?" sorusuna yönelik her bir ülkede her bir formu alan öğrenci sayısı farklı olduğundan anormal yanıt örüntüsü sergileyen öğrenci sayısının farklılaşması beklenmektedir. Anormal yanıt örüntüsü gösteren öğrencilerin madde yanıt grafikleri R studio programı yardımıyla çizdirilerek değişen madde zorluklarına göre verilen yanıtlar arasındaki tutarsızlıklar görselleştirilecektir. "Hesaplanan birey uyum indeks değerlerinin ortalamaları ülkelere göre farklılık göstermekte midir?" araştırma sorusuna yönelik yapılacak olan ANOVA testi sonuçlarında ülkelerin eğitim sistemleri ve kültürel normlarından dolayı farklılaşmaların meydana gelmesi beklenmektedir. "Hesaplanan indeks değerlerinin sonuçları birbiri ile uyumlu mudur?" araştırma sorusuna yönelik ise Guttman ölçeğine dayalı indekslerin tümü cevap örüntülerini farklı şekillerde ağırlıklandırdıklarından anormal yanıt örüntüsüne sahip öğrenci sayılarının indekslere göre değişkenlik göstermekle birlikte birbirlerine uyumlu sonuçlar üretmeleri beklenmektedir.

Kaynaklar

Bademci, V. (2019). Geçerlik: Nedir? Ne değildir? *Eğitim ve Toplum Araştırmaları Dergisi*, *6*(2), 373-385. https://dergipark.org.tr/tr/pub/etad/issue/51092/646773

Ferrara, S. & DeMauro, G. E. (2006). Standardized assessment of individual achievement in K-12. *Educational measurement*, *5*, 579-621.

Karabatsos, G. (2003) Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement In Education*, *16*(4), 277–298.

MEB (2019). *PISA 2018 ulusal ön raporu.* http://pisa.meb.gov.tr/wp-content/uploads/2020/01/PISA_2018_Turkiye_On_Raporu.pdf

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, *18*(4), 311–314.

Meijer, R. R., Niessen, A. S. M., & Tendeiro, J. N. (2016). A practical guide to check the consistency of item response patterns in clinical research through person-fit statistics: Examples and a computer program. *Assessment*, *23*(1), 52-62.

Meijer, R. R. & Sijtsma, K. (2001) Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*(2), 107–135.

OECD (2019). *PISA 2018 results (volume I): What students know and can do*? OECD Publishing,. https://doi.org/10.1787/5f07c754-en.

Olson, J., & Fremer, J. (2013). *TILSA test security guidebook: Preventing, detecting, and investigating test security irregularities.* Council of Chief State School Officers.

Tendeiro, J. N., & Meijer, R. R. (2014). Detection of invalid test scores: The usefulness of simple nonparametric statistics. *Journal of Educational Measurement*, *51*(3), 239-259.

Tendeiro J. N., Meijer R. R., and Niessen A. S. M. (2016). PerFit: An r package for person-fit analysis in IRT. *Journal of Statistical Software, 74*(5), 1–27. https://doi.org/10.18637/jss.v074.i05

Veri tipinin makine öğrenmesi yöntemleri tahminleme performansına etkisi

İlhan Koyuncu ve Abdullah Faruk Kılıç

Anahtar kelimeler: Makine öğrenmesi, kategorik veri, sürekli veri, simülasyon

Giriş

Eğitimde yapılan sınıflandırma ve tahminleme çalışmalarında çok çeşitli makine öğrenmesi teknikleri kullanılmaktadır. Bu sayede, öğrencilerinin akademik başarılarının yanında geçme kalma durumları ile ilgili de önemli çıkarımlarda bulunulmaktadır. Yapılan çalışmalarda yapay sinir ağları, regresyon teknikleri, Bayes teknikleri, destek vektör makineleri, karar ağaçları ve kurallı temelli birçok tekniğin performansı sınanmaktadır (örn. Hamalainen ve Vinni, 2011; Koyuncu, 2018; Minaei-Bidgoli ve diğ., 2003; Nghe ve diğ., 2007; Romero ve diğ., 2008; Tepehan, 2011; Tezbaşaran, 2016). Yapılan bu çalışmalarda farklı veri setleri için farklı tekniklerin yüksek performans gösterdiği bulunmuştur. Benzer şekilde, Romero ve diğ., (2013) yaptıkları çalışmada farklı veri tipleri için farklı tekniklerin daha etkili olabileceğini belirtmişlerdir. Bu nedenle, makine öğrenmesi tekniklerinin farklı veri tipleri için nasıl performans gösterdiği önemli bir araştırma konusudur. Nitekim, bu konuda yapılan çalışmalarda öğrenme/test verisi oranı (Brain ve Webb, 1999; Foody ve diğ., 2006; Koyuncuve Gelbal, 2020), Örneklem büyüklüğü (Hamalainen ve Vinni, 2006; 2011), farklı veri yapıları (Romero ve diğ., 2008; Romero ve diğ., 2013), kayıp veri (Hamalainen ve Vinni, 2011) gibi durumlarda sınıflandırma ve tahminleme performansları incelenmiştir. Bu çalışmada ise birinci aşamada PISA 2018 veri setinden hareketle yöntemlerin performansları kategorik, sürekli ve karışık yapıdaki bağımsız değişkenler açısından incelenmiştir. İkinci aşamada ise gerçek veri setinin özellikleri göz önünde bulundurularak yöntemlerin performansları bağımsız değişkenin ölçek düzeyine ek olarak örneklem büyüklüğü ve bağımsız değişken sayısı açısından incelenmiştir. Buna göre, gerçek veri seti için elde edilen sonuçların geçerliği farklı koşullarda üretilen simülasyon verisinde test edilmiştir.

Bu doğrultuda, bu araştırmanın amacı, farklı ölçek düzeylerinde ölçülmüş değişkenlerin makine öğrenmesi tekniklerinin tahminleme performansı üzerindeki etkisini incelemektir. Bu amaç doğrultusunda aşağıdaki araştırma problemlerine yanıt aranmıştır:

Doğrusal regresyon, yapay sinir ağları, en yakın komşuluk, kurallı temelli öğrenme ve karar ağaçları tekniklerinin tahminleme performansı;

1. Gerçek veri setinde kategorik, sürekli ve karışık (%50 kategorik + %50 sürekli) veri tipleri için nasıldır?
2. Gerçek veri setine benzer yapıdaki simülasyon verisinde kategorik, sürekli ve karışık (%50 kategorik + %50 sürekli) veri tipleri için değişen miktarlardaki bağımsız değişken sayısı (10, 20 ve 30) ve örneklem büyüklüğünde (250, 1000 ve 5000) nasıldır?

Yöntem

Farklı veri tipleri ve bunların kombinasyonlarında makine öğrenmesi yöntemlerinin tahmin performansının incelenmesi amacıyla gerçekleştirilen bu araştırma iki çalışmadan oluşmaktadır. Birinci çalışmada PISA verisi ile makine öğrenmesi yöntemlerinin performansları incelenmiştir. İkinci çalışmada ise Monte Carlo simülasyon yöntemi kullanılarak farklı veri tiplerinde makine öğrenmesi yöntemlerinin performansları incelenmiştir. Monte Carlo simülasyonları belirlenen dağılım özelliklerine uygun olarak örneklem verisinin oluşturulduğu ve analiz edildiği çalışmalardır (Sigal ve Chalmers, 2016).

PISA verisi ile yapılan birinci çalışmada 6890 Türk öğrencinin okuma başarısını anlamlı yordayan 26 sürekli değişken tespit edilmiştir. Bu değişkenlerin dağılımlarının normal ya da normale yakın olduğu belirlenmiş ve daha sonra yarısı 5 kategorili hale getirilmiştir. Sonrasında ise tüm değişkenler 5 kategorili hale getirilerek tümü sürekli, tümü kategorik ve karışık (%50 kategorik + %50 sürekli) olmak üzere 3 dosya elde edilmiştir. Bu dosyalar kullanılarak öğrenme verisinin tüm veriden rastgele %67 oranında seçildiği 100 replikasyondan oluşan analizler gerçekleştirilmiştir. Bu analizlerde doğrusal regresyon (LinearRegression), yapay sinir ağları (MLPRegressor, RBFNetwork ve RBFRegressor), en yakın komşuluk (IBkLG ve KStar), kural temeli (DecisionTable ve M5Rules) ve karar ağaçları (AlternatingModelTree, DecisionStump ve REPTree)

Simülasyon çalışmasında ise bağımsız değişken sayısı (10, 20 ve 30), örneklem büyüklüğü (250, 1000 ve 5000) ve veri kategorileri (tüm bağımsız değişkenler sürekli, bağımsız değişkenlerin yarısı sürekli yarısı 5 kategorili ve tüm bağımsız değişkenler kategorili) manipüle edilmiştir. Çalışmada 3x3x3 = 27 simülasyon koşulunda çalışılmış olup her bir koşul için 100 replikasyon yapılmıştır.

Veri üretiminde R yazılımında (R Core Team, 2020) araştırmacılar tarafından yazılan kodlar kullanılmıştır. Bağımsız değişkenler ve bağımlı değişken öncelikle sürekli değişken olarak üretilmiştir. Bağımsız değişkenler ile bağımlı değişken arasındaki korelasyonlar; bağımsız değişkenin 10 olduğu koşulda 4 değişken (-0.15, -0.05), 1 değişken (-0.06, 0.05), 4 değişken (0.06, 0.15) ve 1 değişken (0.16, 0.25) aralıkta üretilmiştir. Bağımsız değişkenin 20 olduğu koşulda 8 değişken (-0.15, -0.05), 2 değişken (-0.06, 0.05), 8 değişken (0.06, 0.15) ve 2 değişken (0.16, 0.25) aralıkta üretilmiştir. Bağımsız değişkenin 30 olduğu koşulda ise 12 değişken (-0.15, -0.05), 3 değişken (-0.06, 0.05), 12 değişken (0.06, 0.15) ve 3 değişken (0.16, 0.25) aralıkta üretilmiştir. Belirlenen korelasyon aralıkları gerçek veri seti olan PISA'dan seçilen 26 bağımsız değişkenin bağımlı değişkeni arasındaki korelasyonlar temel alınmıştır. Sürekli olarak üretilen veri setleri; bağımsız değişkenlerin kategorik olma durumuna belirlenen eşik

noktaları yardımıyla 5 kategorili hale getirilmiştir. Veri analizi ise Weka (Hall ve diğ., 2009) yazılımında gerçekleştirilmiştir.

Yöntemlerin performansları; hata karelerin karekökü (root mean squared error [RMSE]), ortalama mutlak hatalar (mean absolute error [MAE]) ve modelden kestirilen bağımlı değişken değerleri ile bağımlı değişkenin gerçek değerleri arasındaki korelasyon kullanılarak incelenmiştir. Ayrıca, 10 katmanlı çapraz geçerleme yöntemi ile analiz tekniklerinin performansı iyileştirilmiştir.

Sonuçlar

Gerçek veri ile yapılan çalışmada, en az hatalı sonuçlar tüm veri tipleri için karar ağacı (AlternatingModelTree), doğrusal regresyon, yapay sinir ağları (MLPRegressor ve RBFRegressor) ile kural temelli yaklaşım olan M5Rules ile elde edilmiştir. Kategorik veri için en etkili yöntemlerin doğrusal regresyon ve yapay sinir ağları (MLPRegressor ve RBFRegressor) olduğu bulunmuştur. Yöntemlerin genel olarak sürekli veride daha az hatalı sonuçlar verdiği görülmüştür. Korelasyon değerleri de bu bulguları destekler niteliktedir. Ancak gerçek veri ile yapılan çalışmanın 26 bağımsız değişken ve 6890 öğrenci ile yapılmıştır. Bu nedenle, bu sonuçların benzer koşullar için genellenebilirliğine yönelik yapılan simülasyon çalışması sonucunda hata karelerinin karekökü incelendiğinde, örneklem büyüklüğünün 250 olduğu koşulda MLPRegressor yöntemi kategorik veri oranı artıkça daha hatalı kestirimler yapmıştır. Bağımsız değişken sayısının artması da bu örneklem büyüklüğünde MLPRegressor yönteminin performansını düşürmüştür. Bu örneklem büyüklüğünde Kstar yöntemi de diğer yöntemlere göre daha düşük performans göstermiştir. Örneklem büyüklüğünün artması verinin kategori sayısının önemini azaltmıştır. Diğer bir deyişle örneklem büyüklüğü arttıkça yöntemlerin farklı veri tiplerindeki performansı birbirine yaklaşmıştır.

Ortalama mutlak hatalar açısından incelendiğinde de hata karelerinin karekköüne benzer sonuçlara ulaşılmıştır. Korelasyonlar incelendiğinde ise bağımsız değişken sayısındaki artışın bazı yöntemlerde korelasyonu da artırdığı gözlenmiştir. Örneklem büyüklüğünün artması da korelasyon üzerinde kısmi etkiye sahiptir. DecisionStump, KStar ve RBFNetwork yöntemlerinin korelasyon değerleri 0 civarındadır. Ancak RBFNetwork yönteminde bağımsız değişkenlerin sürekli olduğu koşulda korelasyon 0.20 civarındadır. Buna göre RBFNetwork yönteminin kategorik veriye duyarlı olduğu söylenebilir. Diğer taraftan RBFRegresson yöntemi de kategorik-sürekli karma veri tipinde sürekli veri tipine benzer sonuç verirken tüm bağımsız değişkenlerin kategorik olduğu durumda korelasyonun düştüğü gözlenmiştir.

Kaynaklar

Brain, D., & Webb, G. (1999, December, 5-6). On the effect of data set size on bias and variance in classification learning. In *Proceedings of the Fourth Australian Knowledge Acquisition Workshop*, University of New South Wales (pp. 117-128), Sydney, Australia.

Foody, G. M., Mathur, A., Sanchez-Hernandez, C., & Boyd, D. S. (2006). Training set size requirements for the classification of a specific class. *Remote Sensing of Environment, 104*(1), 1-14.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Peter, R., & Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations, 11*(1), 10-18.

Hamalainen, W., & Vinni, M. (2006). Comparison of machine learning methods for intelligent tutoring systems. In M. Ikeda, K. D. Ashley, and T. W. Chan (Eds.), *Proceedings of International Conference on Intelligent Tutoring Systems* (pp. 525-534). Springer. https://doi.org/10.1007/11774303_52

Hamalainen, W. & Vinni, M. (2011). Classifiers for educational technology. In C. Romero, S. Ventura, M. Pechenizkiy, and R. Baker (Eds.), *Handbook of educational data mining* (pp. 54-74). CRC Press.

Koyuncu, İ. (2018). *Öğrencilerin PISA matematik başarılarının yordanmasında veri madenciliği yöntemlerinin karşılaştırılması* (Tez No. 494325) [Doktora tezi, Hacettepe Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Koyuncu, İ., & Gelbal, S. (2020). comparison of data mining classification algorithms on educational data under different conditions. *Journal of Measurement and Evaluation in Education and Psychology, 11*(4), 325-345.

Minaei-Bidgoli, B., D.A. Kashy, G. Kortemeyer, & W. Punch. Predicting student performance: An application of data mining methods with an educational web-based system. In Proceedings of *33rd Frontiers in Education Conference*, (pp. 13-18). Westminster, CO.

Nghe, N. T., Janecek, P., and P. Haddawy, P. (2007). *A comparative analysis of techniques for predicting academic performance.* 37th Annual Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports, (pp. T2G-7-T2G-12). Milwaukee, USA. https://doi.org/10.1109/FIE.2007.4417993

Romero, C., Espejo, P. G., Zafra, A., Romero, J. R., & Ventura, S. (2013). Web usage mining for predicting final marks of students that use Moodle courses. *Computer Applications in Engineering Education, 21*(1), 135-146.

Romero, C., Ventura, S., Espejo, P. G. & Hervás, C. (2008). Data mining algorithms to classify students. In *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 8-17). Montréal, Québec, Canada. https://www.researchgate.net/profile/Michel-Desmarais-2/publication/221570516_Adaptive_Test_Design_with_a_Naive_Bayes_Framework/links/09e41512380f4e4bc1000000/Adaptive-Test-Design-with-a-Naive-Bayes-Framework.pdf#page=8

R Core Team. (2020). R: A language and environment for statistical computing. https://www.r-project.org/

Sigal, M. J., & Chalmers, R. P. (2016). Play it again: Teaching statistics with monte carlo simulation. *Journal of Statistics Education*, *24*(3), 136–156. https://doi.org/10.1080/10691898.2016.1246953

Tepehan, T. (2011). *Türk öğrencilerinin PISA başarılarının yordanmasında yapay sinir ağı ve lojistik regresyon modeli performanslarının karşılaştırılması* (Tez No. 308559) [Doktora tezi, Hacettepe Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Tezbaşaran, E. (2016). *Temel bileşenler analizi ve yapay sinir ağı modellerinin ölçek geliştirme sürecinde kullanılabilirliğinin incelenmesi* (Tez No. 421572) [Doktora tezi, Mersin Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Sayısal yetenek farklılaşmasının cinsiyete göre incelenmesi: Bir değişen madde fonksiyonu (DMF) çalışması

Yasemin Yardım ve Tuncay Öğretmen

Anahtar kelimeler: Rasch model, değişen madde fonksiyonu (DMF)

Giriş

Sayısal düşünme yeteneği, demokratik bir vatandaş olmak için çok önemlidir, çünkü evde, işyerinde ve yerel topluluklarını etkileyen karmaşık ulusal ve uluslararası konularda bilinçli kararlar almalarını sağlar.

Sayısal yetenek testlerinde cinsiyet temelli bir farklılaşmanın olup olmadığını araştırmak bu aşamada önemlidir. Meslek seçimi ve kurumlara yerleşme süreçlerinde kız ve erkek bireylerin girdikleri sınavlardaki testlerin kızların veya erkeklerin lehine ya da aleyhine olmaması gerekir.

Hayatlarımız boyunca, ama özellikle de okul yıllarında herkes çok sayıda teste tabi tutulur. Bu testlerden bazıları yüksek öneme sahiptir, yani yaşamları etkileyen kararlar için temel oluştururlar. Yüksek öneme sahip sınavların çoğunda, cinsiyet, ırk veya etnik grup, sosyo-ekonomik durum ve coğrafi bölgenin bir fonksiyonu olarak farklılık gösteren grup farklılıkları veya ortalama farklılıklar vardır (Willingham ve Cole, 1997). Grup farklılıklarının olması testlerin yanlı olduğu anlamına mı geliyor? İstatistikçiler, psikometri uzmanları (ölçme bilimlerinde uzmanlaşmış, genellikle zekâ veya kişilik gibi psikolojik yapıları ölçen kişiler) ve diğer test uzmanları için bir madde, bir değişkendeki başarıyı farklı gruplardaki bireyler için eşit derecede iyi yordarsa, yansızdır.

Madde yanlılığı, sınava giren farklı grupların, aynı test maddesine farklı yanıt vermesidir. Bu farklılıklar, hem test maddesinde hem de sınava giren farklı grupların deneyimlerine ve geçmişlerine ışık tutabileceği için keşfedilmeyi gerektirir (Holland ve Thayer, 1986a). Bir testin öğrencinin bilişsel yeteneğini yanlı ya da yansız bir ölçüt haline getirip getirmediğine bakılmaksızın, testin yanlı tahminlerde bulunması olasılığı yüksektir. Örneğin, akademik bir yetenek testi, erkeklerin gelecekteki akademik performansını rutin olarak fazla tahmin edebilir ve kadınlarınkini düşük tahminleyebilir. Bu durumda akademik seçim veya yerleştirme kararlarını almak için bu test kullanılıyorsa, kadınlar dezavantajlı olacaktır.

Testlerdeki farklılaşmalar madde ve test bazında incelenmektedir. Bu çalışmalar çoğunlukla Değişen madde fonksiyonu (DMF) analizi ile yapılmaktadır. DMF analizlerinin Madde Tepki Kuramı (MTK) yani IRT (Item Response Theory) temelli yöntemlerle yapılması yaygındır (Murphy ve Davidshofer, 2005).

Bu çalışmanın amacı, DMF belirlemek için Rasch modelini kullanarak MTK çerçevesinde, lise 11. sınıflara yapılacak olan sayısal yetenek testindeki (SYT) maddelerin cinsiyete göre DMF içerip içermediğini araştırmaktır.

Bu konu ile ilgili yapılan araştırmalara bakıldığında pek çok araştırmada sayısal yetenek testlerinde erkeklerin kızlardan daha başarılı olduğu sonucuna varılmaktadır. Bu çalışmada, daha önce yapılan birçok araştırmanın tersine, PISA, TIMMS, OKS, SBS dataları gibi hazır data kullanmak yerine bir sayısal yetenek testi oluşturulmuştur. Bu teste, geçerlilik, güvenirlilik çalışmaları yapılıp madde parametreleri hesaplanarak asıl şekli verildikten sonra, test lise 11. sınıf öğrencilerine uygulanmıştır. Bu çalışma kapsamında ortaya konan problem cümlesi şu şekildedir:

Sayısal yetenekte cinsiyet gruplarına göre farklılaşma var mıdır? Problem cümlesi çerçevesinde incelenen alt problemler şu şekildedir:

1. Bu araştırma kapsamında geliştirilen sayısal yetenek testi psikometrik özellikleri nelerdir?
2. SYT Rasch modelinin varsayımlarını karşılamakta mıdır?
3. SYT maddeleri cinsiyet gruplarına göre DMF içermekte midir?

Yöntem

Araştırma evreni İzmir ilinde okuyan ve 2018-2019 eğitim öğretim yılı itibarıyla 11. sınıfa devam etmekte olan öğrencilerden oluşmaktadır. Bu çalışma için seçilen örneklem ise 1036 kişilik devlet Anadolu Liselerinde okuyan 11. sınıf öğrencisinden oluşmaktadır. Örneklem seçilirken Anadolu Liselerine yerleşme yüzdelik dilimi en az %8 olan Anadolu Liselerindeki öğrenciler tercih edilmiştir. Çalışmaya katılan öğrenciler Aritmetik Akıl Yürütme, Eşitlik Kurma, Geometri ve Görsel-Uzamsal Yetenek alt testlerinin her birinden 5 er madde olmak üzere Sayısal yetenek testindeki 20 maddeyi cevaplamışlardır. Öğrencilere cevaplama süresi olarak 40 dakika verilmiştir. Yanıtlar doğru yanlış (1-0) olarak değerlendirmeye alınmıştır.

Elde edilen veriler madde tepki kuramı çerçevesinde Rasch modeline göre Winsteps programı kullanılarak analiz edilmiştir. Araştırmanın ikinci alt problem cümlesini incelemek için veri setinin Rasch modelinin varsayımlarını karşılayıp karşılamadığı araştırılmıştır. Araştırmanın ikinci ve üçüncü alt problem cümlesindeki, SYT'nin Rasch modelinin varsayımlarını karşılayıp karşılamadığının ve SYT maddelerinin cinsiyete göre DMF içerip içermediğinin araştırılması doğrultusunda aşağıdaki çalışmalar yapılmıştır:

i) Model seçimi: Bu çalışmada maddeler iki yanıt seçeneğine sahip olduğundan iki boyutlu Rasch modeli seçilmiştir.

ii) Modele uyum testleri: Gözlemlenen yanıt ile modelin beklediği yanıt arasındaki farka vurgu yapıldığında, uygun istatistiklerin çoğu ki-kare temellidir. Winsteps'de bunlara INFIT ve OUTFIT istatistikleri denir. Winsteps'in standartlaştırılmış uyum istatistikleri de vardır (ZSTD). Bu istatistikler gözlemlenen ve beklenen değerler arasındaki tüm farkların tüm bireyler üzerinde toplandığı standartlaştırılmış bir toplamdır (Conaghan, 2007). Bu testlerden elde edilen sonuçlar test ve kişi uyum iyiliği kapsamında raporlanmıştır.

iii) DMF testleri: Bunun için Rasch modeli çerçevesinde Winsteps programındaki analizler sonucunda öğrenci DMF grafikleri incelenerek cinsiyete göre DMF içeren maddeler tespit edilmiş ve bu DMF'lerin büyüklükleri ile manidarlık dereceleri rapor edilmiştir.

Sonuçlar

Bu çalışmada, Sayısal Yetenekteki farklılaşma cinsiyete göre incelenmiştir. 20 maddeden oluşan SYT, 2018-2019 eğitim öğretim yılında 11. Sınıfında okuyan 1382 öğrenciye uygulanmıştır. İlk analizler yapıldıktan sonra modele uymayan öğrenciler datadan temizlenmiştir. Eleme sonrası 1036 öğrenci üzerinde analizler gerçekleştirilerek rapor edilmiştir. Rasch analizlerindeki kişi-model uyumuna bakıldığında outfit mnsq değerinin 0.99 ve infit mnsq değerinin de 1.00 olduğu görülmüş olup, bu da kişi-model uyumunun mükemmel olduğunu göstermektedir. Standart sapma infit mnsq ve outfit mnsq değerlerinin sırasıyla .11 ve .20 olması test-data uyumun iyi olduğunu göstermektedir. Testin tek boyutluluk ve yerel bağımsızlık koşulları da analiz edilmiş ve karşılanmıştır.

Oluşturulan testin ön uygulmasında 355 öğrenci üzerinden yapılan KR-20 güvenirlik hesaplamalarında güvenirlik katsayısı .61, madde güçlük ortalaması .411 ve madde ayırdedicilik ortalaması 0.286 olarak bulunmuştur. Güvenirlik katsayısı düşük görünse de daha sonra yapılan Rasch analizinde 1036 öğrenci üzerinden yapılan test güvenirlik hesaplamalarında test güvenirliği .99 olarak bulunmuştur. Örneklem büyüklüğü arttığında testin genel güvenirliğinin arttığı görülmüştür.

Bu çalışmada SYT'nin alt testlerindeki, eşitlik kurma-sayısal akıl yürütme alanlarındaki 2 maddenin erkekler lehine DMF içerdiği ve geometri-görsel uzamsal alanlarındaki 3 maddenin ise kızlar lehine DMF içerdiği tespit edilmiştir. Literatürde maddelerin DMF göstermesinin sebebinin kız ve erkek öğrencilerin farklı ilgi alanları ve yeteneklere sahip olmasından kaynaklandığı veya genetik olarak farklı düşünme yapılarına sahip olmalarından kaynaklandığı gibi birçok araştırma vardır. Bu çalışmada DMF içeren maddeler tespit edilmiş fakat nedenleri üzerinde durulmamıştır.

Erkekler lehine DMF içeren madde örneği

7.

	Arda	Mert	Zeynep	Polen	Orhan
Ağırlık	27	45	51	63	69

Yukarıdaki tabloda Arda, Mert, Zeynep, Polen ve Orhan' ın ağırlıkları kg cinsinden verilmiştir. Bu beş kişi bir miktar elmayı aşağıdaki kurallara göre paylaşıyorlar.

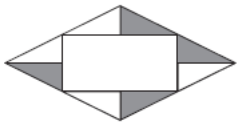
- İlk olarak elmaları bu beş kişi aralarında eşit paylaşıyorlar.
- Daha sonra bu kişiler ağırlıkça kendinden ağır olanlara her 1 kg farkı için 3 elma veriyor.
- Paylaşım sonucunda Arda' nın 20 elması oluyor.

Paylaşım sonucunda kimin elmalarının sayısı ilk paylaşımdaki elmaların sayısıyla aynıdır?

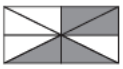
A) Arda B) Mert C) Zeynep D) Polen E) Orhan

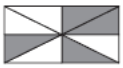
Kızlar lehine DMF içeren madde örneği

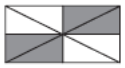
15.

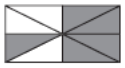


Yukarıda açılmış biçimi verilen şeffaf zarf kapatıldığında aşağıdakilerden hangisi elde edilir?

A) 

B) 

C) 

D) 

E) 

Kaynaklar

Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a rasch paper? *Arthritis & Rheumatism, 57*(8), 1358–1362. https://doi.org/10.1002/art.23108.

Halpern, D. F. (2012). *Sex differences in cognitive abilities* (4th ed.). Psychology Press.

Holland, P. W., & Thayer, D. T. (1986a). Differential item functioning and the Mantel-haenszel procedure. *ETS Research Report Series* (Tecnical Report No. 86-69). chrome-extension://dagcmkpagjlhakfdhnbomgmjdpkdklff/enhanced-reader.html?pdf=https%3A%2F%2Fbrxt.mendeley.com%2Fdocument%2Fcontent%2F6d96fb3b-1615-3f5e-a3f2-8b247f2be7d1

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing: Principles and applications.* Pearson/Prentice Hall.

Yükseköğrenime devam eden lisans öğrencilerinin hipotez kurma becerilerinin değerlendirilmesi

Uğur Hassamancıoğlu ve Fatma Betül Kurnaz Adıbatmaz

Anahtar kelimeler: Hipotez kurma becerisi, bilimsel düşünme becerisi, yükseköğrenim

Giriş

Bilimsel düşünme becerileri uygun ortam -bilimsel düşünme becerilerinin kullanılabileceği ortamlar- yaratıldığında desteklenebilir. Dolayısıyla öğretmenlerin çocuklara aktarabileceği konularda "neyin, nasıl olduğu" bilgisinden ziyade "neyin, neden olduğunu ya da araştırma ve analizin neden gerekli olduğunu" düşünmelerini sağlamaları gerekmektedir (Kuhn, 2011). Desteğin içeriğiyle ilgili alanyazında farklı çalışmalar bulunmaktadır. Örneğin yapılan bir araştırmada (Haverlikova, 2017) tıp fakültesi öğrencilerinin öğretim sürecinde sürekli yeni hipotezler kurmaları, hipotez kurma becerilerinde bir değişikliğe neden olmuştur. Eckel ve diğ., (2019) çalışmalarında birinci sınıf tıp fakültesi öğrencilerine fizyoloji dersinde konuyla ilgili bir model verilmiştir (sağlam fare kası) ve yeni bir durum yaratılarak incelemeler gerçekleştirmeleri istenmiştir. Daha sonra tartışma sorusu belirlenmiş ve öğrencilere bilimsel bir çalışma gerçekleştirecekleri ve bu süreçte konuyu tanımlamaları, bilinmeyenleri belirlemeleri, hipotez geliştirmeleri, hipotezi test etmeleri ve verileri yorumlamaları gerektiği belirtilmiştir. Öğretmen, süreci yönlendirme ve bilgileri özetleme sorumluluğunu üstlenmiştir. Bu çalışmaların sonunda öğrencilerin bilimsel düşünme becerileri puanının arttığı sonucuna ulaşılmıştır. Ayrıca bir rehberin yönlendirmesiyle sorgulamaya dayalı öğrenme materyallerinin, öğretimde dikkat etme, problemleri anlama ve açıklama, hipotez kurma, deney yapma, gözlemleri kaydetme, tartışma, sonuçları açıklama etkinlikleriyle birlikte kullanımının (Widia, ve diğ., 2021); uygun bir hipotez seçme, hipotez oluşturma, deney planı hazırlama ve sonuçları açıklama, hipotezi test etme, verileri çözümleme ve sonuçları yorumlama gibi becerileri içeren müdahale programın ders programlarıyla bütünleştirilerek kullanılmasının bilimsel düşünme becerilerini desteklediği (Zimbardi ve diğ., 2013) belitilmiştir.

Morris ve diğ. tarafından yapılan bir araştırmada video oyunlarının güdüleyici özelliğinin olması, oyun sırasında gerçekleşen her davranışla ilgili geri bildirimde bulunması, akıl yürütme stratejilerini içermesi, üst bilişsel yapıyı desteklemesi gibi özelliklere sahip olması nedeniyle bilimsel düşünme becerilerini etkileyebileceğinden söz edilmiştir. Örneğin "River City" oyununda, oyuncu küçük gruplarla salgına karşı bir şehri korumak üzere bilimsel araştırmalar yürütmektedir. "Mad City Mystery" oyununda

ise oyuncular gizemli bir ölümü açığa çıkarmak için araştırma yapar, sorgular ve tartışmalarda bulunurlar. Bilimsel düşünme becerileriyle fen bilimleri alanında eğitici olan "Supercharged!, Quest Atlantis, Environmental Detectives ve Biohazard" gibi oyunlar da önerilmektedir (Morris ve diğ., 2013).

Bu çalışmada yükseköğrenim öğrencilerinin hipotez kurma becerileri incelenmiştir. Sosyal bilimler alanında bir yükseköğrenim programına devam eden ve psikolojiye giriş ve gelişim psikolojisi dersi almış lisans öğrencilerine gelişim psikolojisi konularıyla ilişkili belirli durumlar/araştırma ve deney soruları verilerek öğrencilerin hipotez kurma becerileri değerlendirilmiştir. Bu süreçte biçimlendirici değerlendirme kullanılmıştır. Biçimlendirici değerlendirme öğretim sürecinde öğrencilerin bilgilerinin ortaya çıkarıldığı, öğretimin bu doğrultuda şekillendiği (Bulunuz ve Bulunuz, 2017) temel amacı öğrenmeyi arttırmak ve istenen davranışların ortaya çıkmasını sağlayan değerlendirme türüdür. Biçimlendirici değerlendirme öğrenciye geri bildirimler sağlar ve öğrenci de aldığı bu geri bildirimlerle eksik yönlerini tamamlayıp bir sonraki adımı buna göre yönlendirir (Metin ve Özmen, 2010). Biçimlendirici değerlendirme doğası gereği sürekli, dinamik ve ilerleyicidir, öğrenciler kendi öğrenme sorumluluklarını üstlenip kendini yönlendirirken öğretmenler de öğrencilerin öğrenimlerini topladıkları bilgilerle değerlendirir ve süreci bu bilgilerin ışığında ilerletebilir (Bell ve Cowie, 2001). Verilen etkili geri bildirimlerle öğrencinin motivasyonu ve öz düzenleme becerileri desteklenir, öğrenci kendine yeni hedefler belirleyebilir ve bu hedeflere ulaşmada stratejiler geliştirebilir (Clark, 2012). Öğrencilerin verilen durumlar/araştırma ve deney sorularına yönelik yanıtları araştırmacılar tarafından analiz edilmiştir. Yapılan analizler sonucunda öğrencilere verilen her durum için geri bildirimler verilmiştir. Bilimsel düşünme süreçlerinin ders müfredatıyla bütünleştirilmiş şekilde, bir rehber yönlendirmesiyle ve uygun düşünme ortamları yaratıldığında desteklenebileceği yapılan araştırmalar sonucunda görülmektedir. Hipotez kurma becerisinin varolan bilgilerden yararlanma, problem durumunu anlama ve açıklama, kanıta dayalı olarak akıl yürütme, verilen durumu eleştirme ve tüm olasılıkları gözden geçirme gibi becerileri içerdiği bilinmektedir. Bu araştırmada sosyal bilimler alanında bir yükseköğretim programına devam eden lisans öğrencilerinin hipotez kurma becerilerinin incelenmesi amaçlanmıştır.

Yöntem

Bu araştırma sosyal bilimler alanında bir yükseköğretim programına devam eden lisans öğrencilerinin hipotez kurma becerilerinin incelenmesini amaçladığından, var olan durumu olduğu gibi açıklamaya çalışan tarama modelli bir araştırmadır.

Araştırmada hipotez kurma becerilerinin değerlendirilmesi açık uçlu maddelerle gerçekleştirildiğinden ve veri toplama sürecinin güçlükleirnden dolayı evren ve örneklem belirleme yoluna gidilmemiş, veriler çalışma grubundan toplanmıştır. Çalışma grubu bir devlet üniversitesinde sosyal bilimler alanında öğrenim gören, psikolojiye giriş ve gelişim psikolojisi dersleri almış birinci (n=74) ve üçüncü (n=104) sınıfa devam eden 178 öğrenciden oluşmaktadır. Öğrencilerin 9,5'i erkek, 90,5'i kadındır. Araştırmada hipotez kurma becerileri araştırmacılar tarafından oluşturulan açık uçlu maddelerle ölçülmüş ve bu maddelere verilen yanıtların değerlendirilmesinde dereceli puanlama

anahtarları kullanılmıştır. Aşağıda bu maddelerin oluşturulma süreci ve dereceli puanlama anahtarlarının geliştirilmesiyle ilgili bilgiler verilmiştir.

Araştırmada hipotez kurmayla ilgili beceriler araştırmacılar tarafından oluşturulmuş açık uçlu maddelerle ölçülmüştür. Açık uçlu maddeler Kurnaz-Adıbatmaz ve Kutlu (2020) tarafından oluşturulan bilimsel düşünme becerileri sınıflandırmasında yer alan hipotez kurma becerilerine ilişkin açıklamalar dikkate alınarak oluşturulmuştur. Bu sınıflandırmada hipotez kurma becerisi aşağıdaki alt becerilerden oluşmaktadır.

- Açık uçlu verilmiş bir problemle ilgili gözlem sonuçlarından yola çıkarak problemi tanımlar.
- Araştırma problemini açıklar.
- Problemin nedenlerini açıklar.
- Problem durumuna dayalı test edilebilir bir hipotez kurar.
- Hipotezin test edilmesinde çözüme götürebilecek birden fazla uygun kanıt olabileceğini açıklar.
- Gözlenen etkinin nasıl kontrol edileceği konusunda açıklamalar yapar.
- Hipoteze yöneltilebilecek eleştirileri giderebilmek için tüm olasılıkları kontrol eder.

Hipotez kurmayla ilgili bu beceriler dikkate alınarak yedi madde oluşturulmuştur. Maddeler, tüm öğrencilerin piskolojiye giriş ve gelişim psikolojisi dersi almış olmaları nedeniyle gelişim psikolojisi konu alanıyla ilişkilendirilmiştir. Maddelerin oluşturulmasında gerçek yaşama dayalı sorunlar kullanılmış ve bu sorunlarla ilgili hipotez kurma becerileri ölçülmüştür. Maddeler iki ölçme ve değerlendirme uzmanı, bir çocuk gelişimi uzmanı tarafından ölçülen özelliğe uygunluk, açıklık, anlaşılırlık, öğrenci düzeyine uygunluk vb. bakımlardan değerlendirilmiş ve uzman görüşlerine dayalı olarak maddeler gözden geçirilmiştir ve maddelerin geçerliliğine ilişkin kanıtlar toplanmıştır.

Sonuçlar

Bu araştırmada bir yükseköğrenim kurumuna devam eden lisans öğrencileirnin hipotez kurma becerileri incelenmiştir. Araştırmanın bulguları incelendiğinde elde edilen sonuçlar aşağıdaki gibi özetlenebilir. Öğrenciler aşağıda liste olarak verilen becerileri gerçekleştirmekte önemli güçlükler yaşamışlardır.

- Maddede verilen problem durumla ilgili çıkarımlar yapma; bu durumu bilimsel ilkelerle, kuramlarla, olgularla, vb. ilişkilendirme
- Maddede verilen problemin nedenleri açıklama; problemin çözümüne ilişkin tahminler sunma
- Maddede verilen durumla ilgili değişkenleri belirleme; hipotezin araştırılmasına yönelik bilimsel araştırma yöntemleriyle tutarlı bir yöntem önerme

- Maddede verilen durumla ilgili hipotez oluşturma; hipotezin sonuçlarına ilişkin tahminde bulunma
- Maddede verilen durumla ilgili oluşturduğu hipotezin kontrol edilmesinde hangi kanıtların toplanması gerektiğini ve bu kanıtların ne tür bilgiler sağlayabileceğini açıklama
- Maddede verilen hipotezi test etmek için bilimsel araştırma yöntemleriyle tutarlı bir deney planlama
- Bir hipotezin test edilmesinde, araştırma sonucuna getirilebilecek eleştirilerin neler olduğunu ve bu eleştirileri gidermek için ne tür önlemlerin alınabileceğini açıklama

Kaynaklar

Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education*, 85, 536-553.

Bulunuz, N. ve Bulunuz, M. (2017). Biçimlendirici değerlendirme uygulamalarının lise öğrencilerinin denge ve tork kavramlarını anlamalarına etkisinin değerlendirilmesi. *Araştırma Temelli Etkinlik Dergisi, 7*(1), 21-33.

Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review*, 24, 205–249. https://doi.org/10.1007/s10648-011-9191-6

Eckel, J., Zavaritskaya, O., Schüttpelz-Brauns, K., & Schubert, R. (2019). An explorative vs. traditional practical course: how to inspire scientific thinking in medical students. *Advances in Physiology Education, 43*(3), 350-354.

Haverlikova, V. (2017). *Development of hypothesisıng skills in biophysical context among medical students.* 11th International Technology, Education and Development Conference (pp. 6242-6245). Valencia: INTED2017 Proceedings.

Kuhn, D. (2011). What is scientific thinking and how does it develop? In U. Goswami (Ed.), *The Wiley-Blackwell Handbook of Childhood Cognitive Development* (pp. 497-523). A John Wiley & Sons, Ltd., Publication.

Kurnaz-Adıbatmaz, F. B. ve Kutlu, Ö. (2020). *Bilimsel düşünme becerilerinin ölçülmesi.* Pegem Akademi.

Metin, M. ve Özmen, H. (2010). Biçimlendirici değerlendirmeye yönelik öğretmen adaylarının düşünceleri. *Milli Eğitim*, 187, 293-309.

Morris, B. J., Croker, S., Zimmerman, C., Grill, D., & Roming, C. (2013). Gaming science: The "Gamification" of scientific thinking. *Frontiers in Psychology, 4*(607), 1-16. https://doi.org/10.3389/fpsyg.2013.00607

Widia, F., Sartina, A., Irawan, Syafrudin, Armansyah, Nurdiana, Hunaepi, Sapnowandi, Prayogi, and Asy'ari, M. (2021). The effectiveness of guided inquiry learning tools in increasing students' activities and creative thinking skills. *Journal of Physics: Conference Series, 1816*(1). http://dx.doi.org/10.1088/1742-6596/1816/1/012102

Zimbardi, K., Bugarcic, A., Colthorpe, K., Good, J. P., & Lluka, L. (2013). A set of vertically integrated inquiry-based practical curricula that develop scientific thinking skills for large cohorts of undergraduate students. *Advances in Physiology Education*, *37*(4), 303-315. https://doi.org/10.1152/advan.00082.2012

Matematikte akademik yılmazlığı yordayan değişkenlerin incelenmesi: Bir açıklayıcı madde tepki modellemesi uygulaması

Sevilay Kilmen ve Naime Şahin Baloğlu

Giriş

Matematik başarısını yordayan değişkenlerin incelenmesiyle ilgili birçok araştırma mevcuttur. Matematik başarısının bilişsel yodayıcılarının yanında duyuşsal yordayıcıları da son zamanların araştırma konusu olmuştur. PISA, TIMSS gibi uluslararası sınavlarda da akademik beceriyi ölçen testlerin yanında öğrenci özellikleri ile ilgili daha fazla bilgi toplamayı amaçlayan anketler de uygulanmaktadır. Özellikle literatür, akademik başarı ve yılmazlığın birbiriyle ilişkili olduğunu göstermektedir (Arastaman ve Balcı, 2013). Bu nedenle matematikte akademik başarıyı yükseltmek için matematikte akademik yılmazlığın incelenmesi önemlidir.

Matematiksel yılmazlık, öğrencilerin matematiğe güvenle yaklaşma, zorluk karşısında çabalamaya devam etme, öğrenmede ısrarlı ve başarılı olma istekliliği olarak tanımlanmaktadır (Borman ve Overman, 2004; Hutauruk ve Priatna, 2017; Johnston-wilder ve Lee, 2010; Kooken diğ., 2013, 2016; Ricketts ve diğ., 2015; Atahan, 2019; Layco, 2020). Yılmazlık olgusuna sahip bireyler matematik öğrenme sürecindeki zorluklar karşısında güven ve ısrarla çalışma, tartışma, düşünme ve araştırma istekliliğiyle çalışmaya ve çabalamaya devam ederler.

Literatürde matematikte akademik yılmazlığı ölçen iki ölçek yer almaktadır. Bunlar; Kooken ve arkadaşları tarafından 2016 yılında geliştirilen ve matematiksel yılmazlığı üç boyutuyla ölçen Matematiksel Yılmazlık Ölçeği ve Ricketts ve arkadaşları tarafından 2015 yılında geliştirilen Matematikte Akademik Yılmazlık Ölçeği'dir. Bu araştırmada matematikte akademik yılmazlığı yordayan değişkenler açıklayıcı madde tepki modellemesi kapsamında incelendiği için matematik dersine yönelik yılmazlığı tek boyutta ölçen Pekdemir ve diğ. (2019) tarafından Türk kültürüne uyarlanan "Matematikte Akademik Yılmazlık Ölçeği" kullanılmıştır.

Araştırma kapsamında ele alınan değişkenler ilgili literatür kapsamında cinsiyet, matematik başarısı, matematik kursuna katılım durumu, anne ve baba eğitim durumu olarak belirlenmiştir. Araştırma değişkenlerinin matematikte akademik yılmazlık üzerindeki etkisini daha iyi anlayabilmek için üç basamaklı bir yol izlenmiştir. Araştırmanın ilk basamağında geleneksel olmayan MTK yöntemlerinden açıklayıcı madde tepki modellemesi yardımıyla tüm değişkenlerin matematikte akademik yılmazlık ile

ilişkisine bakılmıştır. İkinci basamakta cinsiyet, anne ve baba eğitim durumu değişkenleri için ölçeğin madde tepki kuramına dayalı ölçme değişmezliği incelenmiştir. Bu aşamada ölçeğin eşik parametrelerinin eşitliğini incelemek için her bir değişken için ayrı ayrı iki ölçme değişmezliği modeli kurulmuş ve bu modeller karşılaştırılmıştır. Üçüncü basamakta ise ölçekte cinsiyet, anne ve baba eğitim durumu değişkenleri açısından madde düzeyinde değişen madde fonksiyonu (DMF) olup olupmadığı incelenmiştir. Tüm bu aşamalar matematikte akademik yılmazlığı yordayan değişkenleri tespit etmenin yanı sıra, araştırma aynı zamanda söz konusu ölçeğin bir geçerleme çalışması niteliğindedir. Araştırma ölçeğin genel kalitesiyle beraber farklı gruplarda nasıl işlediği ile ilgili bilgi sunmaktadır. Analiz sonuçları ölçeğin yapısı ve ölçek maddeleri ile ilgili daha detaylı bilgiler sunma ve değişikliğe gidilmesi gereken noktaları aydınlatma konusunda araştırmacılara ışık tutacaktır.

Yöntem

Çalışma grubunu 2020-2021 Eğitim Öğretim Yılında, Bolu ili Merkez ilçe Milli Eğitim kurumuna bağlı resmi liselerde öğrenim gören toplam 734 lise öğrencisi oluşturmaktadır. Uç değerlerin çıkarılmasından sonra araştırmanın analizleri 553 veri üzerinden yapılmıştır. Araştırmada öğrenci kişisel bilgiler formu ve matematikte akademik yılmazlık ölçeği kullanılmıştır. Kişisel bilgiler formu ile öğrencilerin cinsiyet, matematikte akademik başarı durumu, matematik kursuna katılım durumu, anne ve baba eğitim durumu bilgileri elde edilmiştir.

Ölçme aracının faktör yapısı için AFA ve DFA analizleri yapılmış olup yapı doğrulanmıştır. Ölçeğin güvenirliğini sorgulamak için ise Cronbach alfa katsayısı hesaplanmıştır. Tüm analizler ölçeğin bu araştırmanın çalışma grubu için ölçmek istediği yapıyı güvenilir bir şekilde ölçtüğünü göstermiştir.

Araştırma kapsamında kullanılan ölçeğin madde eşiği ve kişi parametreleri kısmi kredi modeli (KKM) kullanılarak kestirilmiştir. Açımlayıcı madde tepki modellemesi için R (R Core Team,2020) "eirm" paketi (Bulut, 2021) kullanılmıştır. Tek boyuttan oluşan ölçme aracı için toplam altı adet model tahmin edilmiştir. Bu modeller sırasıyla; (0) KKM; (1) cinsiyet ile KKM; (2) cinsiyet ve matematikte akademik başarı ile KKM; (3) cinsiyet, matematikte akademik başarı ve kurs ile KKM; (4) cinsiyet, matematikte akademik başarı, kurs ve anne eğitim durumu ile KKM; (5) cinsiyet, matematikte akademik başarı, kurs, anne eğitim durumu ve baba eğitim durumu ile KKM şeklindedir. Bu yöntem eklenen her açıklayıcı değişkenin gözlemlendiğinde ve gözlemlenmediğinde oluşan değişimin belirlenmesine olanak sağlamaktadır.

Madde tepki kuramına dayalı olarak üç değişmezlik modelinden söz edilebilir. (1) Şekil değişmezliği tüm madde parametrelerinin gruplar arasında serbest olduğu değişmezlik modelidir. (2) Metrik değişmezlik modeli sadece ayırt edicilik parametrelerinin gruplar arasında eşit olacak şekilde sınırlandırıldığı modeldir. (3) Skalar değişmezlik modeli ise tüm madde parametrelerinin (ayırt edicilik ve eşik parametresi) gruplar arasında eşit olarak sınırlandırıldığı modeldir. Ki-kare analizine dayalı olarak şekil ve metrik değişmezlik modelleri arasında anlamlı bir fark gözlenirse bu durum madde ayırt edicilik parametresinin eşitliği varsayımının ihlal edildiğine işaret etmektedir. Metrik ve skalar değişmezlik

modelleri arasındaki ki-kare değerlerindeki anlamlı farklılık ise madde eşik parametrelerinde ölçüm değişmezliği varsayımının ihlal edildiğini göstermektedir (Spencer ve diğ., 2019). Bu araştırmada kısmi kredi modeline dayalı olarak analizler gerçekleştirildiği için ayırt edicilik parametresi sabit olduğundan sadece şekil ve skalar ölçme değişmezliği modelleri kurulmuştur. Bu iki model arasında ki-kare karşılaştırması yapılarak araştırmaya katılan tüm değişkenlere göre madde eşiklerinin ölçme değişmezliği incelenmiştir. Ölçme değişmezliği analizlerinin ardından DMF analizleri MTK kapsamında olabilirlik oranı yöntemi ile yapılmıştır. Bu yöntemde madde parametreleri için kısıtlı ve genişletilmiş modeller üretilmiştir. Ölçme değişmezliği ve DMF analizleri için R (R Core Team, 2020) "mirt" paketi (Chalmers, 2012) kullanılmıştır.

Sonuçlar

Bu çalışmanın amacı öğrencilerin matematikte akademik yılmazlık özelliklerini yordayan değişkenlerin incelenmesidir. Bu amaçla AMTM, ölçme değişmezliği ve DMF analizleri birlikte kullanılmış ve öğrencilerin yılmazlık özelliklerini yordayan değişkenlerle ilgili daha kapsamlı bilgiler elde edilmeye çalışılmıştır. Genel olarak sonuçlar öğrencilerin yılmazlık özelliklerinin matematikte akademik başarıya göre farklılaştığını göstermektedir. Ancak araştırmanın diğer değişkenleri olan cinsiyet, matematik kursuna katılım durumu, anne ve baba eğitim durumun göre yılmazlık ölçeğindeki maddelere verdikleri tepkilerdeki değişim anlamlı bir etkiye sahip değildir. AMTM sonuçlarına göre matematikte akademik başarısı ile matematikte akademik yılmazlık özelliklerinin arasındaki ilişkiyi tekrar ortaya koymuştur (Alva,1991; Atahan, 2019). Ölçme değişmezliği analizi sonuçlarına göre cinsiyet alt gruplarına göre ölçme değişmezliğinin kurulmadığı görülmüştür. Baba eğitim durumuna göre bir maddede DMF tespit edilmiştir.

Kaynaklar

Alva, S. A. (1991). Academic invulnerability among Mexican-American students: The importance of protective resources and appraisals. *Hispanic Journal of Behavioral Sciences*, *13*(1), 18-34. https://doi.org/10.1177/07399863910131002

Arastaman, G. ve Balci, A. (2013). Investigation of high school students' resiliency perception in terms of some variables. *Educational Sciences: Theory and Practice*, *13*(2), 922-928. https://files.eric.ed.gov/fulltext/EJ1017335.pdf

Atahan, Ş. (2019). *Matematiksel modellemeye dayalı öğretimin matematiksel yılmazlık algısı ve modelleme becerisine etkisi* (Tez No. 561476) [Yüksek lisans tezi, Balıkesir Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Borman, G. D. ve Overman, L. T. (2004). Academic resilience in mathematics among poor and minority students. *The Elementary School Journal*, *104*(3), 177-195. https://doi.org/10.1086/499748

Bulut, O. (2021). *eirm: Explanatory item response modeling for dichotomous and polytomous item responses* (version 0.3) [Computer software]. https://cran.r-project.org/package=eirm.

Revelle, W. (2020). *psych: Procedures for psychological, psychometric, and personality research* (version 2.0.12) [Computer software]. https://cran.r-project.org/package=psych

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software, 48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Hutauruk, A. J. B. ve Priatna, N. (2017, September). Mathematical resilience of mathematics education students. *Journal of Physics: Conference Series, 895*, IOP Publishing. https://doi.org/10.1088/1742-6596/895/1/012067

Johnston-Wilder, S., and Lee, C. (2010, September, 1-4). *Developing mathematical resilience* [Paper presentation]. Warwick University BERA annual conference, Coventery, England.

Kooken, J., Welsh, M. E., and Mccoach, D. B. (2013, October). *Running head: Mathematical resilience as a higher order factor* [Paper presentation]. New England Educational Research Association, Rocky Hill, CT.

Kooken, J., Welsh, M. E., McCoach, D. B., Johnston-Wilder, S., and Lee, C. (2016). Development and validation of the mathematical resilience scale. *Measurement and Evaluation in Counseling and Development*, *49*(3), 217-242. https://doi.org/10.1177/0748175615596782

Layco, E. P. (2020). Discerning the intervening roles of students mathematical resilience and academic emotions between the relationship of home-school ecological structures and achievement. *International Journal of Innovation, Creativity and Change. 14*(8), 439-469. https://www.ijicc.net/images/Vol_14/Iss_8/14806_Layco_2020_E1_R.pdf

Pekdemir, Ü., Yazıcı, H., Altun, F. ve Tosun, C. (2019). Matematikte akademik yılmazlık ölçeği'nin Türk kültürüne uyarlanması. *Türk Bilgisayar ve Matematik Eğitimi Dergisi*, *10*(1), 217-231. https://doi.org/10.16949/turkbilmat.446722

Ricketts, S. N., Engelhard Jr, G., and Chang, M. L. (2015). Development and validation of a scale to measure academic resilience in mathematics. *European Journal of Psychological Assessment*, *33*(2), 79–86. https://doi.org/10.1027/1015-5759/a000274

Spencer, M., Cho, S. J. ve Cutting, L. E. (2019). Item response theory analyses of the Delis-Kaplan executive function system card sorting subtest. *Child Neuropsychology, 25(*2), 198-216. https://doi.org/ 10.1080/09297049.2018.1433156

IrtGUI: Tek boyutlu madde tepki kuramı analizlerini bir kullanıcı arayüzü ile gerçekleştiren bir R paketi

Hüseyin Yıldız

Anahtar kelimeler: Madde tepki kuramı, R paketi, kullanıcı arayüzü, yerel bağımsızlık

Giriş

Madde Tepki Kuramı (MTK) ve bu kurama bağlı uygulamaların ölçme değerlendirme alanında sıklıkla çalışılan konulardan biri olduğu söylenebilir. MTK çerçevesindeki modeller ve uygulamalar sürekli olarak artmakta ve yenilenmektedir. Klasik Test Kuramının aksine MTK kestirimlerini yapmak, uygun modelin seçilmesi ve varsayımlarının sınanması oldukça karmaşık işlemler gerektirmektedir. MTK analizlerinin yapılabilmesi için birçok farklı bilgisayar programı geliştirilmiştir. Bu programlardan bazıları IRTPRO, BILOG, MULTILOG olarak sıralanabilir. Söz konusu programların kullanımı lisansa ya da ücrete tabi olabilmektedir. Ya da ücretsiz versiyonlarında belirli kısıtlamalarla karşılaşılmaktadır. Ayrıca son yıllarda R programlama dilinde MTK analizlerini yapmaya imkan veren açık kaynak kodlu ve ücretsiz bir çok R paketi yazılmış ve yayınlanmıştır. Mirt, ltm, irtoys, irtplay gibi R paketleri örnek olarak gösterilebilir. Söz konusu R paketleri kullanılarak hemen hemen tüm MTK analizleri ücretsiz olarak gerçekleştirilebilmektedir.

R paketleri ve barındırdıkları fonksiyonlar kullanışlı araçlar olsa da, temel düzeyde kod yazma becerisi gerektirmektedir. Kod okur-yazarlığı konusunda yeterli deneyimi olmayan araştırmacı ve uygulamacılar bu işlevsel analiz araçlarından sağlıklı bir şekilde faydalanamamaktadır. R dilinde yer alan "shiny", "shinydashboard" gibi kütüphaneler, yazılmış kod veya fonksiyonların grafik kullanıcı arayüzüne (Graphical User Interface - GUI) dönüştürülmesine imkan tanımaktadır. Bu sayede herhangi bir kodlama yapmadan söz konusu fonksiyonların kullanımı, çıktıların görüntülenmesi mümkün olmaktadır.

Bu çalışmanın amacı farklı MTK modelleri altında; madde ve yetenek parametre kestirimleri, model ve madde uyum indekslerinin elde edilmesi, boyutluluk ve yerel bağımsızlık varsayımlarının sınanması, madde karakteristik ve bilgi fonksiyonlarının çizimi işlemlerinin kullanıcı dostu bir arayüzle gerçekleştirilebildiği "irtGUI" (Yildiz, 2021) isimli R paketini geliştirmek ve tanıtmaktır.

Yöntem

Bu paketin kullanımı diğer R paketleri ile karşılaştırıldığında oldukça basittir. Paket içeriğinde yalnızca bir fonksiyon yer almaktadır. Bu fonksiyonun paketle aynı ismi taşımaktadır. Kullanıcılar "irtGUI" paketini yükleyip, hiç bir argüman kullanmadan irtGUI () fonksiyonunu çalıştırdıklarında bir kullanıcı arayüzü otomatik olarak açılmaktadır. İhtiyaç duyulan kodlar aşağıda paylaşılmıştır.

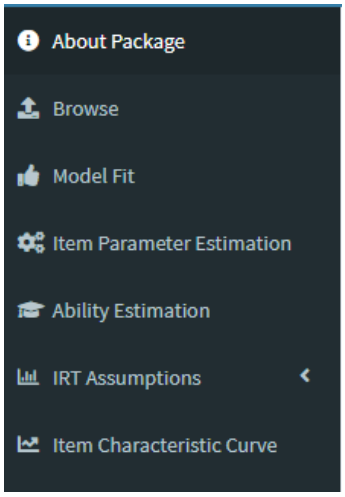
```
install.packages("shinyIRT")

library(shinyIRT)

shinyIRT()
```

Yukarıda sunulan 3 satırlık kodun ilk satırı irtGUI paketini yüklenmesini, ikinci satırı ise paketi aktifleştirilmesini sağlar. Son satır çalıştırıldığında ise irtGUI arayüzü başlatılacaktır. Bu çalışmada "irtGUI" paketine ait kodlar paylaşılmamıştır. Pakete ait tüm kaynak kodlara https://github.com/huseyinyildiz35/irtGUI adresinden ulaşılabilir.

shinyIRT Arayüzü

Uygulama arayüzünün sol tarafında bir sidebar menü yer almaktadır. İlgili menü sekmesiyle ile ilgili girdi ya da çıktılar ise sağ tarafta yer alan main panelde yer almaktadır. Sidebar menüye ait görsel Şekil 1'de paylaşılmıştır.

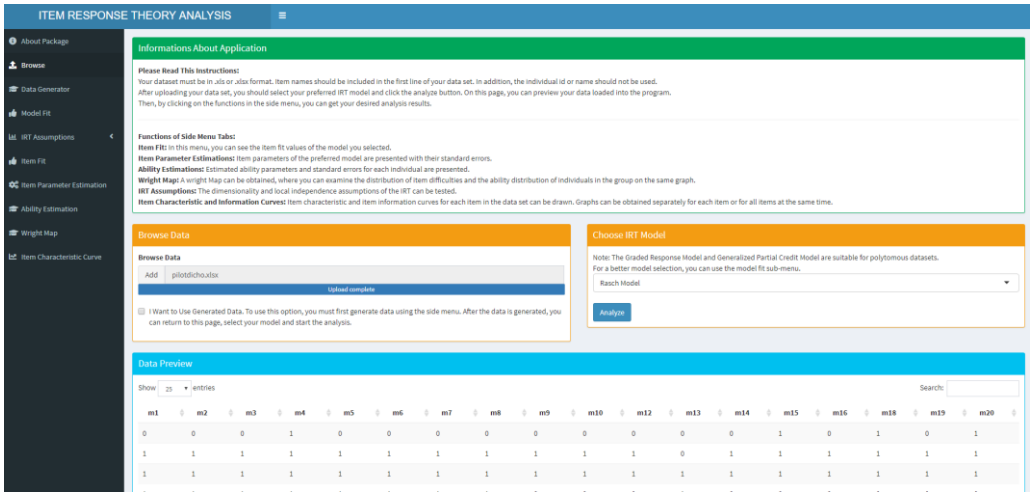
Şekil 1. Sidebar Menu



Şekil 1'de görülebileceği gibi uygulama menüsünde 7 farklı sekme yer almaktadır. "About Package" sekmesinde paket ile ilgili temel bilgiler sunan bir metin bulunmaktadır. "Browse" sekmesinde kullanıcılar analizde kullanmak istedikleri veri setini yükleyebilir ve yüklenen veriyi görebilirler.

Şekil 2

Browse Tab



Şekil 2'de görülen "Browse" sekmesinde kullanıcılar "Browse Data" kutusunun içerisindeki "Add" butonunu kullanarak bilgisayarında bulunan veri dosyasını uygulamaya yükleyebilir. Yükleme tamamlandığında veri seti otomatik olarak "Data Preview" kutusunda görülecektir. Yüklenen veri setinin .xls ya da .xlsx uzantılı bir dosya olması gerekmektedir. Veri seti her satır bireyleri, her sütun ise maddeleri temsil edecek şekilde birey tepkilerinden oluşmalıdır. Dosyanın ilk satırında madde isimleri yer almalıdır. Analizlerin arka planında "mirt" (Chalmers, 2012), "irtoys" (Partchev ve Maris, 2017), "psych" (Revelle, 2020), "WrightMap" (Irribarra ve Freund, 2014) paketlerinden bazı fonksiyonlar kullanılmıştır.

Sonuçlar

Araştırma kapsamında "irtGUI" paket programının ve arayüzünün tanıtılması amacıyla bildiri sunumunda örnek analizler gerçekleştirilecektir. Bu analizlerde 1000 öğrenciden elde edilmiş ve çoktan seçmeli 20 maddeden oluşan gerçek bir veri seti kullanılacaktır. İlgili veri setine ait model veri uyumu ve madde uyumu değerleri, tek boyutluluk ve yerel bağımsızlık varsayımlarının sınama sonuçları, madde ve yetenek parametreleri, Wright haritası, madde karakteristik ve bilgi eğrisi grafikleri katılımcılarla paylaşılacaktır. Söz konusu analizler "irtGUI" paketi yardımıyla Rasch Model, 2 PL Model, 3 PL Model, Aşamalı Tepki Modeli, Genelleştirilmiş Kısmi Puan Modelleri altında sorunsuz olarak gerçekleştirilebilmektedir. Ayrıca kullanıcılar analiz sonuçlarını "download" butonları yardımıyla bilgisayarlarına excel formatında indirebilmektedir. "irtGUI" paketinin manuel dosyasına https://cran.r-project.org/web/packages/irtGUI/irtGUI.pdf adresinden ulaşılabilir.

Kaynaklar

Chalmers, R. (2012). mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Sofware, 48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Irribarra D. T., and Freund, R. (2014). *Wright Map: IRT item-person map with ConQuest integration*. http://github.com/david-ti/wrightmap

Partchev, I. and Maris, G. (2017*). Irtoys: A collection of functions related to item response theory (IRT)* (version 2.0.12) [Computer software]. https://cran.r-project.org/package=irtoys

Revelle, W. (2020). *psych: Procedures for psychological, psychometric, and personality research* (version 2.0.12) [Computer software]. https://cran.r-project.org/package=psych

Yıldız, H. (2021). *irtGUI: Item response theory analysis with a graphic user interface* (version 0.2) [Computer software]. https://cran.r-project.org/package=irtGUI

Hiyerarşik yapıların modellenmesine alternatif bir bakış: Bileşke puanlar

Abdullah Faruk Kılıç, Metin Buluş ve İbrahim Uysal

Anahtar kelimeler: Hiyerarşik yapılar, yapısal eşitlik modelleri, yol analizi, faktör puanı, bileşke puan

Giriş

Hiyerarşik yapılar, birincil düzeydeki faktörlerin üst düzey gizil değişkenleri tanımladığı yapılardır (Mehta & Neale, 2005). Yapısal eşitlik modellerinde (YEM) hiyerarşik yapılar günümüzde daha sık kullanılmaktadır. Bu modellerde parametre kestirimlerinin yapılabilmesi ise örneklemin yeterli büyüklükte olmasına bağlıdır (Kline, 2016). YEM modellerinde hiyerarşik yapılar tamamen gizil değişkenler şeklinde modellenmekte ve modeller karmaşıklaştıkça daha fazla parametre kestirimi gerektirdiğinden ihtiyaç duyulan örneklem büyüklüğü artmaktadır (Devlieger ve Rosseel, 2017). Nitekim veri çok değişkenli normalliğe sahip olduğu durumlarda YEM analizinde en çok olabilirlik (EÇO) yöntemi kullanılarak doğru kestirimler yapılabilmesi için en az 500 veriye ihtiyaç duyulmaktadır. Çok değişkenli normallik varsayımı ihlal edildiğinde ise EÇO ile kestirim yapılabilmesi için 2500'den fazla sayıda veri gerekmektedir (Ullman, 2019). Ancak bu öneriler her durum için uygun olmayabilir. Göstergelerin sürekli olmaması ve normal dağılmaması bunun yanı sıra gösterge değişkenler arasındaki ilişkilerin doğrusal olmaması, puanların güvenirliğinin düşük olması, kayıp veri sayısının fazla olması, modelin karmaşıklaşması gibi durumlar ihtiyaç duyulan gözlem sayısını arttırabilmektedir (Kline, 2016). Örneğin modeller karmaşıklaştıkça örneklem yetersizliğinden dolayı ya yakınsama sağlanmamakta ya da parametre kestirimleri yanlı olmaktadır (Gagne ve Hancock, 2006). Bunun yanında uygunsuz çözümlerle (improper solutions, örn. Heywood vakası) karşılaşılabilmektedir (Loncke ve diğ., 2018).

Yol analizi ve gizil değişkenlerin alt boyutlarına ait puanlarla ilişkilendirildiği bileşke puanlarda (composite scores) (McNeish ve Wolf, 2020), tamamen gizil değişken olarak ele alınan modele göre kestirilecek parametre sayısı daha az olduğundan daha küçük örneklemler yeterli olabilmektedir. Yapılan araştırmalar son yıllarda hiyerarşik yapıda bulunan alt boyut toplam puanları ya da yol analizinde kullanılan ölçek toplam puanları yerine faktör puanlarının da kullanılabileceğine işaret etmektedir. Buna gerekçe olarak ise faktör analizine dayalı puanlarla yapılan geçerlik çalışmasının ardından toplam puanların kullanılmasının geçerliği belirlenen modelle çalışılmadığı anlamına geldiğini belirtmektedirler (McNeish ve

Wolf, 2020). Ayrıca faktör skorlarına dair fonksiyonların kullanımının modelin hatalı tanımlanmasına karşı daha dirençli olduğuna, karmaşık modellerde daha etkili olduğuna, küçük örneklemlerde daha doğru sonuçlar verdiğine, daha az yakınsama sorununa sahip olduğuna ve hata terimleri ilişkili olduğunda kullanılabileceğine ilişkin görüşler bulunmaktadır (Devlieger ve Rosseel, 2017; Hayes ve Usami, 2020a; Loncke ve diğ., 2018). Ayrıca faktör skorlarının kullanımı yukarıda belirtilen ve YEM'de karmaşık modellerden kaynaklı sorunları engelleyebilmektedir (Loncke ve diğ., 2018).

Alanyazında faktör puanları ve toplam puanlar kullanılarak gerçekleştirilen bazı araştırmalar bulunmaktadır (Devlieger ve Rosseel, 2017; Devlieger ve diğ., 2016; Devlieger ve diğ., 2019; Hayes ve Usami, 2020a; Hayes ve Usami, 2020b; Loncke ve diğ., 2018; Lu ve diğ., 2011). Örneğin Devlieger ve diğ. (2016) araştırmalarında regresyon katsayılarını, determinasyon katsayılarını, örneklem büyüklüğünü ve madde sayısını simülasyon koşulları olarak belirleyerek yol analizi ve yapısal eşitlik modelleri üzerinde faktör puanlarını ve toplam puanları karşılaştırmıştır. Devlieger ve diğ. (2016) araştırma sonucunda yanlılık düzeltmesi uygulanan faktör skor yöntemi ile yapısal eşitlik modelinin yanlılık, efficiency, MSE, power ve 1. tip hata oranı açısından en iyi sonuçları sergilediğini, belirsizliğin az (göstergeler arasındaki ilişkinin ve göstergelerin gizil değişkenle ilişkisinin yüksek olması ya da gösterge sayısının fazla olması) olduğunda ise faktör skor yönteminin kullanılabileceğini belirtmiştir. Loncke ve diğ. (2018) ise kayıp verinin bulunduğu ve bulunmadığı koşullarda yapısal eşitlik modelleri ile iki faktör puanı yöntemini (factor score regression and Bartlett factor score regression) karşılaştırdığı araştırmada bağımsız değişkenlerde faktör skorlarının kullanımının yansız kestirimlerde bulunduğunu belirlemiştir. Literatür incelendiğinde örneklem büyüklüğü, ortalama faktör yükü, etki büyüklüğü ve çapraz yüklerin varlığı durumlarıyla tamamen gizil, bileşke puanlar ve yol analizi tekniklerini faktör puanları ve toplam puanlarla aynı araştırmada ele alan bir çalışmaya rastlanmamıştır. Bu yönde araştırmada ortalama faktör yükü, örneklem büyüklüğü, etki büyüklüğü, model tipi ve çapraz yüklerin varlığı koşulları altında modellerin performansları incelenmiştir.

Yöntem

Hiyerarşik ölçme modellerinden elde edilen sonuçların yapısal modeldeki yol katsayısına etkisinin incelendiği bu çalışmada Monte Carlo simülasyonu kullanılmıştır. Monte Carlo simülasyonları verinin belli bir dağılıma göre üretildiği, üretilen verinin farklı istatistiksel metotlarla analiz edildiği ve sonuçlarının karşılaştırıldığı çalışmalardır (Sigal ve Chalmers, 2016). Bu nedenle bu çalışmada da Monte Carlo simülasyonu kullanılmıştır.

Çalışmada ortalama faktör yükü (.40 ve .70), örneklem büyüklüğü (200, 500 ve 1000), etki büyüklüğü (.00, .30, .50 ve .70) ve çapraz yük olma durumu (var ve yok) simülasyon koşulu olarak belirlenmiştir. Her bir simülasyon için değişkenler; tamamen gizil değişken, yarı-gizil değişken (bileşke puan) ve yol analizi modelleriyle analiz edilmiştir. Modellerden yarı-gizil ve yol analizlerinde hem toplam puan hem de faktör puanları kullanılmıştır. Çalışmada 2x3x4x2 = 48 simülasyon koşulunda çalışılmış olup her bir koşul için 1000 replikasyon yapılmıştır.

Veri üretiminde R yazılımında (R Core Team, 2020) bulunan lavaan paketi (Rosseel, 2012) kullanılmıştır. Sürekli ve çok değişkenli normal dağılım gösteren veri setleri 5'li Likert tipinde olacak şekilde kategorize edilmiştir. Faktör puanları ve analizler Mplus (Muthén ve Muthén, 2012) yazılımında gerçekleştirilmiştir. Sürekli veri setlerinin analizinde güçlü en çok olabilirlik (robust maximum likelihood [MLR]) yöntemi kullanılırken kategorik veri setlerinin analizinde varyans ve ortalamaların düzeltildiği ağırlıklandırılmamış en küçük kareler (weighted least squares mean and variance adjusted [WLSMV]) kestirim yöntemi kullanılmıştır.

Yöntemlerin performansları; göreli yanlılık (Moshagen ve Musch, 2014; Rhemtulla ve diğ., 2012), kapsam oranı (coverage rate) (Flora ve Curran, 2004), 1. tip hata, istatiksel güç ve uyum indeksleri (CFI, TLI, RMSEA, χ^2/sd) açısından değerlendirilmiştir.

Sonuçlar

Çalışma sonucunda faktör puanları kullanılarak gerçekleştirilen yol analizi ve bileşke puanların 1. tip hatasının yüksek olduğu (.09-.67) gözlenmiştir. Tamamen gizil modelde ortalama faktör yükü düşük olduğu koşullarda 1. tip hata yükselmiştir. Toplam puanla gerçekleştirilen yol analizinde ise faktör puanlarına benzer şekilde 1. tip hata yüksektir. Ancak bileşke puanların toplam puan ile elde edildiği koşulda birinci tip hata .05 civarındadır.

Güç açısından inceleme yapıldığında sonuçların etki büyüklüğü, örneklem büyüklüğü ve ortalama faktör yüküne göre farklılıklar gösterdiği anlaşılmıştır. Küçük örneklemlerde etki büyüklüğü düşük olsa bile faktör puanlarının kullanıldığı yol analizinin gücü yüksek bulunmuştur. Çapraz yükün yöntemlerin gücü üzerinde etkili olmadığı belirlenmiştir.

Göreli yanlılık açısından toplam puanlar kullanılarak elde edilen bileşke puanın, tamamen gizil modelin ve faktör puanlarıyla elde edilen bileşke puan modelinin kabul edilebilir aralıkta yanlı olduğu gözlenmiştir. Yöntemlerin kapsam oranları incelendiğinde ise toplam puanların bileşke puan olarak kullanıldığı modelin kabul edilebilir aralığa en yakın sonuçları gösterdiği söylenebilir. Uyum indekslerinde tamamen gizil model ve toplam puanlar kullanılarak elde edilen bileşke puan modeli kabul edilebilir aralıktadır. Sonuç olarak hiyerarşik yapılarda yakınsama probleminden kaçınmak için toplam puanın bileşke puan olarak kullanıldığı modelin tercih edilmesi önerilebilir.

Kaynaklar

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement, 76*(5), 741–770. https://doi.org/10.1177/0013164415607618

Devlieger, I., & Rosseel, Y. (2017). Factor score path analysis: An alternative for SEM? *Methodology, 13*, 31–38. https://doi.org/10.1027/1614-2241/a000130

Devlieger, I., Talloen, W., & Rosseel, Y. (2019). New developments in factor score regression: Fit indices and a model comparison test. *Educational and Psychological Measurement, 79*(6), 1017-1037. https://doi.org/10.1177%2F0013164419844552

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491. https://doi.org/10.1037/1082-989X.9.4.466

Gagne, P., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research, 41*(1), 65-83. https://doi.org/10.1207/s15327906mbr4101_5

Hayes, T., & Usami, S. (2020a). Factor score regression in the presence of correlated unique factors. *Educational and Psychological Measurement, 80*(1), 5–40. https://doi.org/10.1177/0013164419854492

Hayes, T., & Usami, S. (2020b). Factor score regression in connected measurement models containing cross-loadings, structural equation modeling. *A Multidisciplinary Journal, 27*(6), 942-951. https://doi.org/10.1080/10705511.2020.1729160

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). The Guilford.

Loncke, J., Eichelsheim, V. I., Branje, S. J. T., Buysse, A., Meeus, W. H. J., & Loeys, T. (2018). Factor score regression with social relations model components: A case study exploring antecedents and consequences of perceived support in families. *Frontiers in Psychology, 9*, 1699. https://doi.org/10.3389/fpsyg.2018.01699

Lu, I. R. R., Kwan, E., Thomas, D. R., & Cedzynski, M. (2011). Two new methods for estimating structural equation models: An illustration and a comparison with two established methods. *International Journal of Research in Marketing, 28*(3), 258–268. https://doi.org/10.1016/j.ijresmar.2011.03.006

McNeish, D., Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods, 52*, 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological methods, 10*(3), 259–284. https://doi.org/10.1037/1082-989X.10.3.259

Moshagen, M., & Musch, J. (2014). Sample size requirements of the robust weighted least squares estimator. *Methodology*, *10*(2), 60–70. https://doi.org/10.1027/1614-2241/a000068

Muthén, L. K., & Muthén, B. O. (2012). *Mplus statistical modeling software: Release 7.0*. Muthén & Muthén.

R Core Team. (2020). *R: A language and environment for statistical computing* [Computer software]. https://www.r-project.org/

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. https://doi.org/10.1037/a0029315

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Sigal, M. J., & Chalmers, R. P. (2016). Play it again: Teaching statistics with monte carlo simulation. *Journal of Statistics Education*, *24*(3), 136–156. https://doi.org/10.1080/10691898.2016.1246953

Ullman, J. B. (2019). Structural equation modeling. In B. G. Tabachnick & L. S. Fidell (Eds.), *Using multivariate statistics* (7th ed., pp. 528-612). Pearson.

Nokta-çift serili korelasyon katsayısı, sınıflamalı tepki modeli ve önerilen yeni bir yöntemden elde edilen çeldirici analizi sonuçlarının karşılaştırılması

Hüseyin Yıldız, Erdem Boduroğlu ve Mahmut Sami Yiğiter

Anahtar kelimeler: Çeldirici analizi, nokta-çift serili korelasyon, sınıflamalı tepki modeli, klasik test kuramı, madde tepki kuramı.

Giriş

Uygun bir ölçme aracının geliştirilebilmesi, ölçme aracında yer alacak maddelerin de uygun niteliklerde hazırlanmasına bağlıdır. Bu sebeple herhangi bir başarı testinin geliştirilmesi sürecindeki en temel adım test maddelerinin hazırlanması aşamasıdır (Crocker ve Algina,1986; Baykul, 2000). Çoktan seçmeli maddeler sağladığı çeşitli avantajlardan dolayı başarı testlerinde yaygın olarak kullanılmaktadır. Göreceli olarak kapsamın daha iyi temsil edilebilmesi, puanlamanın objektif olması, ölçme kuramlarının ikili (1-0) yanıtları, kolayca analiz edebilmesi bu avantajlardan bazılarıdır (Haladyna ve Downing, 1989).

Hazırlanan geniş madde havuzundan nihai test formuna alınacak maddelerin seçilmesi amacıyla madde analizleri yapılır. Guilford (1954) madde analizinde kullanılan 60'tan fazla sayıda yöntemin bulunduğunu belirtmiştir. Madde analizi sürecinin en önemli çıktıları madde güçlüğü, madde ayırt ediciliği ve çeldirici analizleridir. Birçok test geliştiricisine göre çeldiricilerin hazırlanması, çoktan seçmeli madde oluşturma sürecinin en zor aşaması olarak görülmektedir. Bu yüzden çeldiriciler hazırlanırken çok fazla özen gösterilmelidir (Haladyna & Downing, 1989). Bir çeldiriciyi seçmek için gereken yeterlilik düzeyi, doğru cevabı seçmek için gereken yeterlik düzeyinden daha düşük ancak yanlış bir cevap seçmek için gereken yeterlilik düzeyinden daha yüksek olmalıdır. Çeldirici analizi özellikle düşük yeterlik grupları için yanlış öğrenmeleri ve kavram yanılgılarını ortaya çıkarmada önemlidir (Cavanagh ve Waugh, 2011).

Bireylerin çoktan seçmeli maddelere verdiği yanıt performansı çeldiricilerin niteliğine ve kalitesine göre değişmektedir. Bir çeldiriciyi, alt gruptan seçen bireylerin sayısının üst gruptan seçenlere göre daha fazla olması istenilen durumdur. Aksi durumlarda ise çeldiricinin etkili çalışmadığı ifade edilir. Bu sebeple her bir çeldiricinin de ayırt edicilik değerinin hesaplanması gerekmektedir. Klasik test kuramına göre alt-üst %27'lik grubun incelendiği basit yöntem ve madde test korelasyonuna dayanan yöntemler (nokta-çift serili veya çift serili korelasyon) yaygın olarak kullanılmaktadır. Bu yöntemler madde sayısının

az olduğu durumlarda ve test puanlarının normal dağılmadığı durumlarda hatalı sonuçlar verebilmektedir. Ayrıca klasik test kuramının sahip olduğu bazı sınırlılıklar da madde analizlerinde alternatif yöntemlere ihtiyaç olduğunu göstermektedir (Henrysson, 1971; Crocker ve Algina,1986).

Madde Tepki Kuramı (MTK) sağladığı bazı avantajlar sebebiyle son dönemde yaygın kullanım alanlarına sahiptir. Özellikle madde parametrelerinin evrende değişmediği varsayımı ile bireylerin yetenek düzeylerinden bağımsız olarak madde parametrelerinin kestirilmesi söz konusudur. MTK'da a parametresi madde ayırt ediciliğinin bir ölçüsü olarak ifade edilir ve madde karakteristik eğrisinin eğimi ile ilişkilidir. Eğimi az olan madde karakteristik eğrileri maddenin göreceli olarak düşük ayırt edicilikte olduğuna işaret eder. Bu durum düşük ve yüksek yetenek düzeylerinde maddenin doğru yanıtlanma olasılığının yeteri kadar farklılaşmadığı anlamına gelir. Eğrinin dikleşmesi ise maddenin yüksek ayırt ediciliğe sahip olduğu şeklinde yorumlanır (Baker, 2001; De Mars, 2010).

MTK, çeldiricilerin etkili olarak çalışıp çalışmadığını değerlendirdiği gibi aynı zamanda araştırmacılara çeldiricileri kullanarak bireylerin yetenekleri hakkında tahminde bulunma olanağı sunar. MTK'da, Bock (1972) tarafından önerilen Sınıflamalı Tepki Modeli (Nominal Response Model) çoktan seçmeli maddelerde çeldiricilerin analiz edilmesinde yaygın olarak kullanılmaktadır. Sınıflamalı Tepki Modeli (STM) seçenekler arasında herhangi bir sıralama olmaksızın, bireylerin yetenek düzeyleri ve her bir seçeneği işaretleme olasılıkları arasındaki ilişkiyi inceleme imkânı sunar. Bu modele göre bireyin yetenek düzeyi azaldıkça belirli bir çeldiriciyi seçme olasılığının artması beklenir (De Ayala, 2009; Gierl ve diğ., 2017). STM'nin matematiksel yapısı gereği bir maddenin tüm seçeneklerinin ayırt edicilikleri toplamı 0 olacak şekilde ölçeklenmektedir (Desjardins ve Bulut, 2018). Bu sebeple bir çeldiriciye ait a parametresi değeri diğer seçeneklerin ayırıcılık düzeyinden etkilenmektedir. Bu durum bir çeldiricinin ayırıcılık gücünü temsil eden a parametresi değerinin tek başına yorumlanmasını güçleştirmektedir. STM'nin bu sınırlılığından dolayı ayırıcılık gücüne ilişkin kesme puanı belirlenememektedir. Bu sınırlılıklardan dolayı araştırmada MTK'nın 2 parametreli lojistik modeline dayalı olarak yeni bir çeldirici analizi yaklaşımı önerilmiştir.

Bu araştırmada, KTK'ya dayalı çeldirici analizi yöntemlerinden olan nokta-çift serili korelasyon katsayısı (rpbis) yöntemi, MTK'ya dayalı STM yöntemi ve MTK'nın 2PL modeline dayalı olarak önerilen yeni yöntemden elde edilen sonuçlar karşılaştırılmıştır.

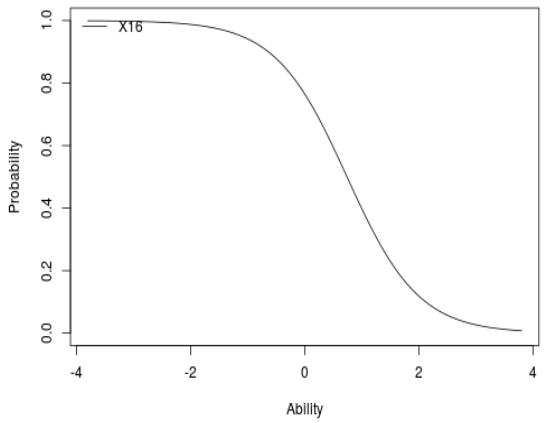
Yöntem

Bu araştırmada MTK'nın 2PL modeline dayalı olarak çoktan seçmeli maddelerin çeldirici performanslarını değerlendirmek adına yeni bir yöntem önerilmektedir. Bu sebeple araştırmanın türü temel araştırma olarak değerlendirilebilir (Fraenkel, Wallen & Hyun, 2012).

2PL modelde a parametresi madde karakteristik eğrisinde (MKE) %50 doğru yanıt olasılığının bulunduğu noktadaki eğimi ile ilişkilidir (Hambleton ve diğ., 1991; Yang, 2014). Dolayısıyla söz konusu noktada eğim değerinin yüksek olması, doğru yanıt olasılığının daha dar yetenek ranjlarında daha hızlı değişim gösterdiği ve maddenin daha ayırıcı olduğu şeklinde yorumlanır.

MTK'nın lojistik modellerinde model denklemini kullanarak bireyin maddeye doğru yanıt verme olasılığını hesapladığımızı söyleriz. Aslında istatistiksel olarak, bireyin doğru kabul edilen seçeneği işaretleme olasılığını hesapladığımız söylenebilir. Hangi seçeneği doğru kabul edersek (Cevap Anahtarında) o seçeneği işaretleyen bireylere "1" değeri atanacak bu seçeneğin tercih edilme olasılığının yeteneğe bağlı fonksiyonu elde edilecektir. Bu tip bir analizde a parametresi de çeldiricinin ayırt ediciliği olarak yorumlanacaktır. Çeldiricinin düşük yetenek düzeylerinde fazla, yüksek yetenek düzeylerinde ise daha az tercih edilme olasılığı beklendiği için bu durumda negatif ve düşük a parametreleri daha başarılı çeldirici performanslarına işaret edecektir. Şekil 1.'de örnek bir çeldirici karakteristik eğrisi paylaşılmıştır.

Şekil 1.Çeldirici Karakteristik Eğrisi Örneği (a= -1.59; b= 0.74)



Her bir çeldiriciye ait parametre değerlerinin incelenmesi için cevap anahtarında her analizde sadece bir cevap manipüle edilerek analizler yapılması gerekmektedir. Yani n madde m cevap seçeneği olan bir test için nxm adet 1-0 veri matrisi oluşturulup aynı sayıda MTK analizi gerçekleştirmek gerekmektedir. Bu işlemin manuel olarak yapılması oldukça zahmetli olacaktır. Bu sebeple tüm bu analizlerin tek seferde yapılmasını ve sonuçların toplu olarak elde edilmesini sağlayacak R kodları yazılmıştır.

Bu araştırmada 20 maddelik, 4 seçenekli 6. Sınıf matematik testinden elde edilen 2000 kişilik gerçek veri seti ile örnek analizler gerçekleştirilerek 3 farklı yöntem ile elde edilen çeldirici performans katsayıları elde edilmiştir. Söz konusu yöntemlere ait katsayılar arasındaki ilişkinin düzeyi Pearson momentler çarpımı korelasyon katsayısı ile incelenmiştir.

Sonuçlar

Araştırma kapsamında KTK'ya dayalı rpis, MTK'ya dayalı STM ile çeldirici analizleri gerçekleştirilmiş ve sonuçlar raporlaştırılmıştır. Tablo 1'de 20 madde ve 80 cevap seçeneğine ait rpbis değerleri sunulmuştur.

Tablo 1

20 Madde 80 Cevap Seçeneğine Ait rpbis Değerleri

Madde No	A	B	C	D
1	-0.360	0.530	-0.480	-0.190
2	-0.320	0.530	-0.420	-0.330
3	-0.290	-0.370	0.530	-0.400
4	0.330	-0.250	-0.280	-0.290
5	-0.200	0.370	-0.330	-0.320
6	-0.330	0.540	-0.420	-0.300
7	0.380	-0.250	-0.310	-0.290
8	-0.370	-0.400	-0.240	0.510
9	-0.370	-0.330	0.520	-0.350
10	-0.160	-0.570	0.640	-0.330
11	-0.370	0.380	-0.340	-0.140
12	-0.320	-0.300	-0.250	0.380
13	-0.330	-0.280	-0.240	0.390
14	0.350	-0.300	-0.230	-0.300
15	-0.310	-0.340	0.430	-0.290
16	0.530	-0.210	-0.400	-0.420
17	-0.450	-0.230	0.530	-0.350
18	-0.380	0.470	-0.310	-0.290
19	-0.390	-0.220	-0.240	0.450
20	-0.280	-0.340	0.450	-0.330

Tablo 2'de STM ile her cevap seçeneği için hesaplanan a parametreleri listelenmiştir.

Tablo 2

STM ile Elde Edilen Seçeneklere Ait a Parametreleri

Madde No	A	B	C	D
1	-1.641	1.611	-0.282	0.312
2	-0.446	1.740	-1.124	-0.170
3	-0.585	-0.492	1.448	-0.370
4	1.031	-0.482	0.276	-0.825
5	-0.189	0.698	-0.412	-0.098
6	-0.456	1.264	-0.532	-0.276
7	0.816	0.049	-0.506	-0.358
8	-0.628	-0.513	-0.086	1.227
9	-0.489	-0.583	1.372	-0.300
10	-0.543	-1.096	1.898	-0.260
11	-0.621	0.838	-0.582	0.364
12	-0.563	0.156	-0.553	0.961
13	-0.241	-0.388	-0.166	0.795
14	0.995	0.185	-0.146	-1.035
15	-0.671	-0.362	1.178	-0.145
16	1.462	-0.419	-0.392	-0.651
17	-0.759	0.067	1.264	-0.571
18	-0.534	1.208	-0.511	-0.162
19	-0.239	-0.265	-0.539	1.042
20	-0.356	-0.443	1.004	-0.205

KTK ve MTK'ya dayalı yöntemlerde çeldirici gücü ile ilgili bilgi veren parametrelerin yanı sıra, MTK'nın 2PL modeline dayalı olarak önerilen yeni çeldirici analizi yaklaşımına ait çeldirici gücü parametreleri de Tablo 3'te paylaşılmıştır.

Tablo 3

Önerilen Yeni Yönteme Ait a Parametreleri

Madde No	A	B	C	D
1	-2.213	1.886	-1.315	-0.637
2	-1.175	2.214	-1.976	-1.323
3	-1.027	-1.253	1.918	-1.393
4	1.031	-1.173	-0.563	-1.475
5	-0.265	0.948	-0.784	-0.899
6	-0.849	1.747	-1.169	-1.200
7	0.970	-0.323	-0.827	-0.688
8	-1.149	-1.081	-0.622	1.642
9	-1.031	-1.249	1.848	-1.499
10	-0.549	-1.754	2.550	-0.773
11	-1.031	0.960	-0.967	-0.047
12	-1.166	-0.550	-1.089	1.093
13	-0.663	-0.730	-0.518	1.049
14	1.046	-0.598	-0.847	-1.677
15	-1.365	-1.177	1.512	-0.921
16	1.957	-1.016	-1.217	-1.443
17	-1.131	-0.425	1.681	-1.543
18	-1.276	1.583	-1.142	-0.843
19	-0.676	-0.469	-0.688	1.328
20	-0.677	-0.979	1.350	-0.967

Tablo 3'teki değerlerin negatif olması yetenek düzeyi düştükçe seçeneğin tercih edilme olasılığının arttığını ifade etmektedir. Bu sebeple daha düşük değerler daha başarılı çeldirici performansına işaret etmektedir. Tablo 3'teki pozitif değerler ise doğru cevap seçeneğine ait olduğu için doğrudan söz konusu maddenin a parametresidir.

Son olarak kullanılan 3 farklı yöntemle elde edilen ve çeldirici performansına işaret eden katsayılar arasındaki ilişki Pearson momentler çarpımı korelasyon katsayısı ile incelenmiştir. Elde edilen sonuçlara göre rpbis ve STM ile elde edilen katsayılar arasında .933'lük bir korelasyon katsayısı elde edilmiştir. Rpbis ve yeni önerilen yöntem ile elde edilen katsayılar arasında ise .969'luk bir ilişki düzeyi tespit edilmiştir. Son olarak STM ve yeni yöntem ile elde edilen katsayılar arasında 0.971'lik çok güçlü bir ilişki tespit edilmiştir. Elde edilen bu yüksek korelasyon katsayısı çeldirici performansına ilişkin farklı yöntemlerle bilgi veren bu değerlerin birbirlerinin yerine kullanılabilir olduğu söylenebilir. STM'nin, problem durumu bölümünde bahsedilen ve çeldirici performansını tek başına yorumlamayı güçleştiren ve kesme puanına göre karar vermeyi engelleyen özellikleri, pbis yönteminin çeldiricinin farklı yetenek

düzeylerinde nasıl çalıştığı ile ilgili bilgi vermemesi göz önünde bulundurulduğunda, önerilen yeni yöntemin araştırmacılara çeldirici analizi konusunda katkı sağlayabileceği düşünülmektedir.

Kaynaklar

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). http://ericae. net/irt/baker.

Baykul, Y. (2000). *Egitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması.* ÖSYM yayınları.

Cavanagh, R. F., & Waugh, R. F. (Eds.). (2011). *Applications of rasch measurement in learning environments research.* Sense Publishers.

Bock, D. R. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29–51. https://doi.org/10.1007/BF02291411

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Cengage Learning.

Çelen, Ü. (2008). Klasik test kuramı ve madde tepki kuramı yöntemleriyle geliştirilen iki testin geçerlilik ve güvenilirliğinin karşılaştırılması. *İlkogretim Online*, *7*(3), 758-768.

De Ayala, R. J. (2009). *The theory and practice of item response theory.* The Guilford Press.

DeMars, C. (2010). *Item response theory.* Oxford University Press.

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R.* Boca Raton: CRC Press.

Fraenkel, J.R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.). McGraw-Hill, Inc.

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: a comprehensive review. *Review of Educational Research*, *87*(6), 1082-1116. https://doi.org/10.3102/0034654317726529

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). McGraw-Hill.

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied measurement in education*, *2*(1), 37-50.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

Henryson, S. (1971). Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.), (pp. 124-153). Council on Education.

Samejima, F. (1968), Estimation of latent ability usıng a response pattern of graded scores. ETS Research Bulletin Series, i-169. https://doi.org/10.1002/j.2333-8504.1968.tb00153.x

Stage, C. (1998). *A comparison between item analysis based on item response theory and classical test theory.* A study of the SweSAT subtest ERC. (Educational Measurement). Umeå University, Department of Educational Measurement. www.edusci.umu.se/digitalAssets/60/60608_enr3098sec.pdf

Tarrant. M., Ware, J., and Mohammed A. M. (2009). An assessment of functioning and non-functioning distracters in multiple-choice questions: A descriptive analysis. *BMC Med Educ, 9*(40), 1-8. https://doi.org/10.1186/1472-6920-9-40

Wiberg, M. (2004). *Classical test theory vs. item response theory: An evaluation of the theory test in the Swedish driving license-test.* Umea: Umea Universiteit. http://www.jus.umu.se/digitalAssets/59/59529_em-no-50.pdf.

Yang, F. M. (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry, 26*(3), 171-177. https://doi.org/10.3969/j.issn.1002-0829.2014.03.010

Değişen madde fonksiyonu gösteren ortak maddelerin test eşitlemeye etkisinin incelenmesi

Feyzi Güneş ve Hülya Kelecioğlu

Anahtar kelimeler: Madde yanlılığı, değişen madde fonksiyonu, test eşitleme, ortak test, eşitleme hatası

Öz

Bu araştırmada değişen madde fonksiyonu (DMF) gösteren ortak maddelerin test eşitlemeye olan etkisinin incelenmesi amaçlanmıştır. Testlerin eşdeğer olmayan gruplar ortak test deseni yöntemiyle eşitlenmesinde, DMF gösteren ortak test maddelerinin testten çıkarılmasının eşitlenmiş puanlara ve eşitleme hatasına olan etkisi doğrusal ve doğrusal olmayan eşitleme yöntemleriyle incelenmiştir. DMF değişkenleri İngilizce, İspanyolca dilleri ve bu dillerde testi alan bireylerin cinsiyetleri olarak belirlenmiştir. DMF gösteren maddelerin belirlenmesinde Mantel-Haenszel ve SIBTEST teknikleri kullanılmıştır. Eşitleme çalışması doğrusal eşitleme yöntemlerinden Tucker, Levine, zincir eşitleme yöntemi, doğrusal olmayan yöntemlerde ise frekans kestirim, zincir eşit yüzdelikli ve Daire-Yay Levine yöntemleri ile yürütülmüştür. Çalışma verilerini, PISA 2018 uygulaması 18 ve 24. kitapçıkları İngilizce, İspanyolca dillerinde bilgisayar ortamında almış fen okuryazarlığı birey yanıtları oluşturmaktadır. Tekniklere göre B ve C düzeyinde DMF gösteren maddelerin ortak testten çıkarılmasının eşitleme hatasına etkisi ağırlıklandırılmış hata kareleri ortalaması ile değerlendirilmiştir. Araştırma sonucunda DMF'li maddelerin ortak testten çıkarılması ile eşitleme hatalarında tutarlı değişimler gözlenmemiştir. Eşitleme hatalarındaki tutarsızlığa; kitapçıkların ortalamaları arasındaki istatistiksel olarak anlamlı fark, DMF değişkenlerine ait alt gruplar arasındaki yetenek farklılaşması ve eşitleme yöntemlerine ait varsayımların karşılanma derecesinin kaynak olabileceği düşünülmüştür.

Bayesian hiyerarşik modelle ölçme değişmezliğinin incelenmesi

Merve Ayvallı ve Hülya Kelecioğlu

Anahtar kelimeler: Ölçme değişmezliği, rastgele etkiler modeli, Bayesian modelleme

Giriş

Eğitimde ve psikolojide yapılan araştırmalarda ölçülen gizil yapılarda gruplar arası anlamlı karşılaştırmaların yapılabilmesi için ölçülen yapının tüm alt gruplarda aynı olması gerekir. Ölçme değişmezliği, birçok grup arasında ve zaman içinde faktör ortalamalarını karşılaştırmak için önemli bir ön koşuldur. Ölçme değişmezliğinin incelenmesi için farklı yöntemler kullanılmaktadır (Millsap, 2011). Bunlar doğrulayıcı faktör analizi temelli yöntemler ve madde tepki kuramı temelli yöntemler olmak üzere iki farklı grupta incelenebilir.

Çoklu-grup doğrulayıcı faktör analizi temelli yöntemlerde söz konusu yapının oluşturulan ölçme modeline ait parametreler tüm alt gruplarda kestirilir. Bu parametrelerin tüm alt gruplarda anlamlı bir şekilde farklılaşıp farklılaşmadığı test edilir (Brown, 2006). Madde tepki kuramı temelli modellerde ise ölçme değişmezliği bir maddeyi doğru cevaplama koşullu olasılığının grup özelliklerine bağlı olmadığında sağlanır (Thissen ve diğ., 1986). Bu modellerden en çok kullanılan ise parametrik bir yöntem olan olabilirlik oranı testidir. Bu yöntem MTK modellerini, değişmezlikle sınırlandırılmış bütün maddeler ile değişmezliği sağlanmış bazı ortak maddeler ve değişmezliği sağlamayan tüm maddelere ait parametrelere göre karşılaştırır (Thissen ve diğ., 1993). Bununla birlikte çok düzeyli madde tepki modellerinden rastgele madde etkisi modeli (random item effects model) ortak maddeye ihtiyaç duymadan ölçme değişmezliğini test etmeye imkân verir. Bu modelde ölçme değişmezliği rastgele madde etkisinin varyans bileşenlerinin Bayes faktörü ile doğrudan değerlendirmesi ile test edilir. (Fox ve Verhagen, 2010).

Rastgele etkiler modelinde, tüm parametrelerin belli düzeyde varyansa sahip olduğu ve parametre değişmezliğinin diğer tüm parametrelerden bağımsız olduğu varsayılır. Rastgele etkiler modelinin daha genişletilmiş hali olan araştırmada kullanılan Bayesian modelde ise parametreler arası varyansın sıfır olduğu ya da olmadığı durumlar göz önünde bulundurularak ölçme değişmezliği testi yapılır Bu araştırmada çok düzeyli yapı gösteren geniş ölçekli testlerde ölçme değişmezliğinin hiyerarşik rastgele etkiler modeli ile incelenmesi amaçlanmıştır. Araştırma birçok eğitim politikasının belirlenmesinde etkili olan geniş ölçekli testlerin geçerliğine ilişkin kanıt oluşturması açısından oldukça önemlidir.

Yöntem

Araştırma nicel araştırma yöntemlerinden olan ilişkisel tarama modelinde yürütülmüştür. Araştırmada TIMSS 2015 uygulamasındaki fen ve matematik başarı testlerine ait veriler kullanışmıştır. Bu kapsamda tüm ülkelerde ortak olarak uygulanmış bir kitapçık belirlenerek, bu kitapçıkta yer alan çoktan seçmeli fen ve matematik maddeleri üzerinden analizler R yazılımı kullanılarak gerçekleştirilmiştir (R Core Team, 2014). Ölçme değişmezliğine ilişkin kestirimler rastgele madde etkisi modelinde değişmezliğin sağlanmadığı madde parametreleri ve ülkelere ait dağılım özelliklerinin gizil değişken olarak analize dâhil edilmesine imkân veren Markov chain Monte Carlo (MCMC) algoritmaları kullanılarak gerçekleştirileştirilmiştir (Fox, 2010).

Verilerin analizinde değişmezliğin sağlanıp sağlanmadığına ilişkin karar verilirken parametreler arasındaki farka dayalı olarak hesaplanan Bayes faktörleri (BF_{01}) kullanılmıştır. Araştırmada kullanılan Bayesian modelde yer alan gruba özgü her bir parametrenin (λ,σ,ν) standart sapmasının Bayes faktörü incelenmiştir. Değişmezlik kararı alınırken Bayes faktör için kesme noktası oldukça katı düzey olarak kabul edilen $BF_{01}<\frac{1}{6}$, ve $BF_{01}>6$ olarak belirlenmiştir (Martin, Williams ve Rast, 2019). Araştırmada gruba özgü her bir parametrenin (λ,σ,ν) standart sapmasının Bayes faktörü incelenmiştir.

Sonuçlar

Sonuç olarak, TIMSS 2015 Uygulamasında yer alan 8 çoktan seçmeli maddenin 3 tanesinde tüm parametrelerde ölçme değişmezliği sağlanırken, ölçme değişmezliğinin sağlanmıyor olma ihtimali olan 5 madde bulunmaktadır.

TIMSS 2015 Uygulamasında yer alan 8 çoktan seçmeli matematik maddesinin 4 tanesinde tüm parametrelerde ölçme değişmezliği sağlanırken, ölçme değişmezliğinin sağlanmıyor olma ihtimali olan 4 madde bulunmaktadır.

Kaynaklar

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* Guilford Publications.

Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P., & Meuleman, B. (2012). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *Journal of Cross-Cultural Psychology*, *43*(4), 558-575. https://doi.org/10.1177/0022022112438397

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications.* Springer.

Fox, J.-P., and Verhagen, A. J. (2010). Random item effects modeling for cross-national survey data. In E. Davidov, P. Schmidt, and J. Billiet (Eds.), *Cross-cultural Analysis: Methods and Applications* (pp. 467–488). Routledge Academic.

Hoijtink, H. (2012). *Informative hypotheses. Theory and practice for behavioral and social scientists.* Boca Raton: Chapman & Hall/CRC.

Martin, S. R., Williams, D. R., & Rast, P. (2019, June 18). Measurement invariance assessment with bayesian hierarchical inclusion modeling. *PsyArXiv*, 1-16. https://doi.org/10.31234/osf.io/qbdjt

Millsap, R. E. (2011). *Statistical approaches to measurement invariance.* Routledge. https://doi.org/10.4324/9780203821961

Rensvold, R. B., & Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. In C. A. Schriesheim & L. L. Neider (Eds.), *Research in management*, (pp. 25-50). Information Age Publishing.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*(1), 118–128. https://doi.org/10.1037/0033-2909.99.1.118

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Lawrence Erlbaum Associates, Inc.

Yazma becerileri puanlarını etkileyen faktörlerin çok-yüzeyli Rasch modeli ile incelenmesi

Sümeyra Soysal, Nuri Doğan ve Mehmet Ali Aydoğmuş

Giriş

Yazma becerilerinin ölçülmesinde sıklıkla kullanılan performans değerlendirmelerde, öğrencilerin çoktan seçmeli testler kadar objektif puanlanması zordur. Dolaysıyla öğrencilerin objektif olarak puanlanamadığı herhangi bir testten aldığı puanlar, testi puanlayan kişiye göre farklılık gösterebilmektedir (Tekin, 2009). Öğrencilerin yapılandırma gerektiren test maddelerine verdikleri cevaplar yalnızca öğrencinin performans düzeyinden etkilenmez; görevin zorluğu, değerlendiricinin katılığı/cömertliği, puanlayıcıların puanlama deneyimi, puanlayıcıların cinsiyeti, puanlayıcıların eğitim düzeyi gibi çeşitli özelliklerle puanlama rubriklerinin uygun kullanımı gibi puanlamayı etkileyebilecek durumlar söz konusudur. Çünkü ölçme amacımızla ilgili olmayan varyans, performans değerlendirmenin adaletini, güvenirliğini ve geçerliğini tehdit eder (Messick, 1998, Prieto ve Nieto, 2014). Ölçme aracının uygulama aşamalarını doğru ve yeterli seviyede yerine getirmelerine rağmen bazı puanlayıcıların daha katı bazılarının daha cömert puanlama yapabildikleri bilinmektedir. Hatta puanlayıcıların puanlama deneyimlerine bağlı olarak puanlama performanslarında farklılık olduğunu raporlayan araştırmalar mevcuttur (Attali 2016; Davis 2016; Alp ve diğ., 2018; Ahmadi-Shirazi 2019). Puanlayıcı etkisini kontrol altına almak amacıyla birden fazla puanlayıcının görev alması, rubriklerin kullanılması gibi önlemler alınmakta ve farklı değişkenlerin puanlamaya etkisine ilişkin değişkenliği ortaya çıkaracak istatistiksel tekniklere dayalı analizler yapılmaktadır. Bu araştırmada yazma becerilerini ölçmek amacıyla kullanılan bir testteki görevlere ilişkin puanlara karışabilecek başta puanlayıcılar olmak üzere çeşitli hata kaynakları incelenmiştir. Bu amaçla potansiyel olarak test puanlarını etkileme olasılığına sahip tüm değişkenlik kaynaklarının dikkate alınmasını sağlayan (Kim ve diğ., 2012) ve bu değişkenlik kaynakları arasındaki etkileşimleri de belirleyebilen (Abu Kassim, 2007) çok yüzeyli Rasch modeli tercih edilmiştir. Çalışmanın amacı doğrultusunda "Yazma becerileri test puanları üzerinde puanlayıcı, puanlayıcının deneyimi, rubrik türü, görev türü değişkenlerinin ve bu değişkenlerin etkileşimlerinin etkisi nedir?" sorusuna yanıt aranmıştır.

Yöntem

Araştırma kapsamında MEB ÖDSGM tarafından 2017 yılında uygulanan "Yazma Becerileri Testi"ne 4., 7. ve 9. Sınıftan katılan toplam 9241 öğrenci arasından seçkisiz olarak 240 öğrenci belirlenmiştir. Araştırmanın örneklemi bu öğrencilerin cevaplarını puanlayacak olan Milli Eğitim Bakanlığı'na bağlı okullarda görev yapan ve daha önce yazma becerilerinin puanlanması çalışmasına katılan deneyimli puanlayıcılar ile puanlama deneyimi olmayan puanlayıcılar arasından seçkisiz olarak belirlenmiş, her branşta (sınıf öğretmeni, Türkçe öğretmeni ve Türk dili ve edebiyatı öğretmeni) 4 olmak üzere toplam 12 öğretmenden oluşmaktadır. Tablo 1 puanlayıcıların branş ve deneyimlerine göre dağılımlarını göstermektedir.

Tablo 1

Puanlayıcıların Branş ve Deneyimlerine Göre Dağılımı

Branş Türü ve Sınıf Düzeyi	Puanlama deneyimine sahip	Daha önce puanlama deneyimi olmayan	Toplam
Sınıf öğretmeni (4. Sınıf)	2	2	4
Türkçe öğretmeni (7. Sınıf)	2	2	4
Türk Dili ve Edebiyatı öğretmeni (9. Sınıf)	2	2	4
Toplam	6	6	12

Kullanılacak veri toplama aracı MEB Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü tarafından geliştirilen ve 2016-2017 eğitim öğretim yılı Nisan ayında 4, 7 ve 9. sınıflarda öğrenim gören öğrencilere uygulanan "Yazma Becerileri Testi"dir. Puanlayıcılar öğrenci cevaplarını MEB tarafından hazırlanan analitik ve bütüncül dereceli puanlama anahtarlarını kullanarak çevrimiçi olarak puanlamışlardır. Bu araştırma için alınan izin doğrultusunda, analizler yalnızca A kitapçığında bulunan veriler üzerinde yapılmıştır. Yazma beceri testinde cümle, paragraf ve metin olmak üzere üç bölümde görevler tanımlanmıştır. 4. ve 7. sınıf testinde 4 cümle, 3 paragraf ve 8 metin; 9. sınıf testinde 4 cümle, 3 paragraf ve 6 metin sorusu yer almaktadır.

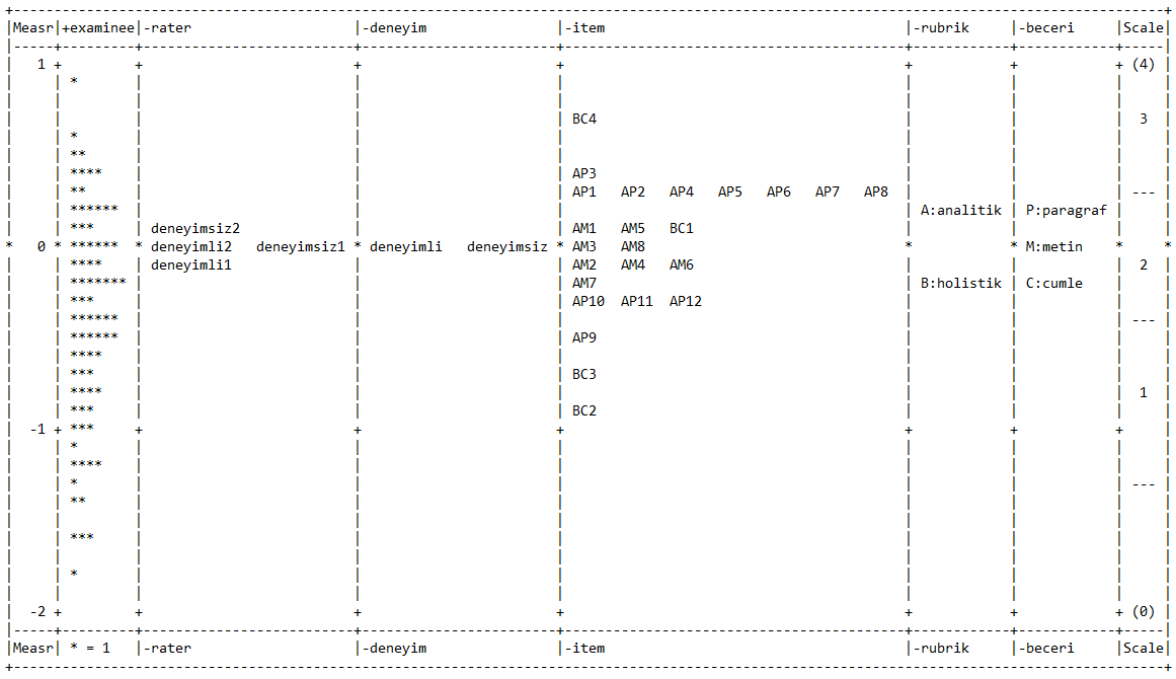
Puanlayıcıların, bütüncül ve analitik rubrikleri kullanarak yazma becerileri testini puanlaması sonucunda elde edilen veriler çok yüzeyli Rasch modeline göre analiz edilmiştir. Analizler, FACET (Linacre, 2014) paket programından yararlanılarak gerçekleştirilmiştir. Araştırmada birey, madde, puanlayıcı, deneyim, rubrik türü ve alt beceri olmak üzere altı facetli bir Rasch modeli test edilmiştir. Analizler her bir sınıf düzeyi için tekrarlanmıştır.

Sonuçlar

4. Sınıf düzeyinde 7680 cevap analiz edilmiştir. 4. Sınıf verileri için elde edilen değişken haritası Şekil 1'de özetlenmektedir.

Şekil 1

4. Sınıf Verileri İçin Elde Edilen Değişken Haritası



Şekil 1 ve Tablo 2 birlikte incelendiğinde puanlayıcılardan deneyimli 1 en cömert puanlamayı yaparken deneyimsiz 2 en katı puanlamayı yapmıştır. Puanlayıcı yüzeyi için ayırma oranı 2.59 ve güvenirlik indeksi .87 olarak belirlenmiştir. Puanlayıcı yüzeyi için hesaplanan güvenirlik indeksi, puanlayıcılar arasındaki güvenilir benzerliği değil; güvenilir farkı göstermektedir (Haiyang, 2010). Bu güvenirliğin yüksek olması puanlayıcıların katılıkları/ cömertlikleri yönüyle farklılık gösterdiği anlamına gelmektedir. Bu araştırmada puanlayıcıların katılık/cömertlik açısından farklılık gösterdiğine işaret etmektedir. Ki-kare sonucu (χ^2= 23.4, p< .001) da bu farkın anlamlı olduğunu göstermektedir. Ayrıca modele göre analitik rubriklerden elde edilen puanların bütüncül rubriklerden daha düşük olduğu, deneyimli puanlayıcıların deneyimsizlere göre daha cömert puanlama yaptığı gözlenirken cümle alt görevlerinin en kolay, paragraf alt görevlerinin ise en zor olduğu belirlenmiştir.

Model-veri uyumu açısından uygunluk içi ve uygunluk dışı istatistiklerinin kabul edilebilir sınırları Wright ve Linacre (1994), 0.6-1.4 olarak belirtmiştir. Analiz sonucunda uygunluk istatistiklerinin puanlayıcıların hiçbirinde kabul edilebilir aralığın dışında kalmadığı saptanmıştır (Min: 0,92 Max:1,25). Ayrıca diğer yüzeylerde de model uyumunu bozan düzey gözlenmemiştir.

Tablo 2

Yüzeylere İlişkin Betimsel Sonuçlar

İstatistik	Birey	Puanlayıcı	Deneyim	Madde	Rubrik	Alt Beceri
Minimum (logit ölçüsü)	-1.78	-0.09	-0.04	-0.9	-0.23	-0.22
Maksimum (logit ölçüsü)	0.89	0.06	0.04	0.72	0.23	0.23
Ortalama (logit ölçüsü)	1.04	0	0	0	0	0
SD (logit ölçüsü)	0.25	0.05	0.04	0.37	0.23	0.19
Standart Hata (model)	0.1	0.02	0.02	0.06	0.02	0.02
RMSE	0.11	0.02	0.02	0.06	0.02	0.02
Ayırma indeksi	5.42	2.59	3.76	6.72	10.84	10.68
Güvenirlik	0.97	0.87	0.93	0.98	1	0.99
Ki-kare	2059**	23.4**	15.1**	1018.3**	237.1**	230.3**

$^{**}p < .01$

Analizler devam etmektedir.

Kaynaklar

Abu Kassim, N.L. (2007, June). Exploring rater judging behaviour using the many-facet Rasch model. Paper Presented in the Second Biennial International Conference on Teaching and Learning of English in Asia: Exploring New Frontiers (TELiA2) Universiti Utara Malaysia.

Ahmadi Shirazi, M. (2019). For a greater good: Bias analysis in writing assessment. *SAGE Open, 9*(1), 1–14. https://doi.org/10.1177/2158244018822377

Alp, P., Epner, A., & Pajupuu, H. (2018). The influence of rater empathy, age and experience on writing performance assessment. *Linguistics Beyond And Within, 3*(3), 7–19.

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing, 33*(1), 99–115.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117–135.

Haiyang, S. (2010). An application of classical test theory and many facet Rasch measurement in analyzing the reliability of an English test for non-English major graduates. *Chinese Journal of Applied Linguistics, 33*(2), 87-102.

Kim, Y., Park, I., & Kang, M. (2012). Examining rater effects of the TGMD-2 on children with ıntellectual disability. *Adapted Physical Activity Quarterly, 29*, 346-365.

Linacre, J.M. (2014). A user's guide to FACETS Rasch-model computer programs. http://www.winsteps.com/a/facets-manual.pdf

Messick, S. (1998). Test validity: A matter of consequences. *Social Indicators Research*, 45, 35-44. https://doi.org/10.1023/A:1006964925094

Prieto, G., & Nieto, E. (2014). Analysis of rater severity on written expression exam using Many Faceted Rasch Measurement. *Psicológica, 35*(2), 385-397.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370-371.

Aykırı değer içeren ve içermeyen meta analiz çalışmalarına dahil edilen araştırma sayısına göre sabit etkiler modeli ve rastgele etkiler modelinin karşılaştırılması

Seda Demir ve Mehmet Fatih Doğuyurt

Anahtar kelimeler: Meta analiz, aykırı değer, sabit etkiler modeli, rastgele etkiler modeli

Giriş

Teknolojinin her geçen gün biraz daha gelişmesiyle farklılaşan ve artan ihtiyaçlar yapılan bilimsel araştırma sayısında da artış sağlamaktadır. Bu artışa bağlı olarak, alanyazında aynı veya benzer araştırma sorusu üzerinde çok sayıda birbirinden bağımsız araştırmayla karşılaşmak mümkündür. Araştırmalarda etki büyüklüklerinin pozitif ve negatif aralıklarda değişmesi, araştırma metodolojilerinin, evren ve örneklemlerinin farklılaşması gibi faktörlerden dolayı genel bir yorum çıkarmak oldukça güçleşmektedir (Demir ve Başol, 2013). Bu noktada, aynı konuda yapılmış olan birden fazla çalışmayı bir araya getirip incelemek ve büyük resmi gösteren ortak bir sonuca ulaşmak için yapılacak meta analiz çalışmalarının gerekliliği ortaya çıkmaktadır. Meta analiz, bir konudaki farklı araştırma sonuçlarını bir ya da birden fazla istatistiksel yöntem yardımıyla birleştirerek ortak bir metrikte standartlaştıran, hesaplanan istatistiksel sonuçları araştırma karakteristikleriyle birlikte özetleyen ve birincil çalışma sonuçlarından daha fazla bilgi veren bir analiz tekniğidir (Glass, 1976; Hedges ve Olkin, 1985).

Birincil çalışmalarda olduğu gibi meta analizde de verilerin dağılımı sonuçların özetlenmesinde önemlidir (Riley ve diğ., 2011). Ayrıca meta analizde kullanılacak modelin seçiminde ise birincil çalışmalardan elde edilen sonuçların varyansları (değişkenliği) dikkate alınmalıdır (Borenstein ve diğerleri 2009). Meta analizle etki büyüklüklerinin birleştirilip ortak (genel) etki büyüklüğünün hesaplanması amacıyla alanyazında sıklıkla kullanılan modeller; Sabit Etkiler Modeli (SEM-Fixed Effects Model) ve Rasgele Etkiler Modelidir (REM-Random Effects Model) (Borenstein ve diğ., 2009; Hedges ve Vevea, 1998). Momentler yöntemi (DerSimonian Laird) REM'de en sık kullanılan yöntem olmasına karşın bunun dışında çeşitli REM yöntemleri de bulunmaktadır (Schwarzer ve diğ., 2015). Kullanılan yönteme göre çalışmaların meta analizdeki ağırlıkları değişeceğinden kullanılan yöntem ortak (genel) etki büyüklüğünün belirlenmesinde doğrudan etkilidir (Borenstein ve diğ., 2009; Hedges ve Vevea, 1998). Araştırmacılar kullanacakları yönteme karar verirken meta analize dahil edilen çalışmaların etki

büyüklüklerini dikkate almaktadırlar. Genel olarak çalışmaların etki büyüklükleri homojen ise SEM, homojen değil ise REM tercih edilmektedir.

Meta analiz çalışmalarında, çalışmalar arası homojenliğin bozulmasının örnekleme hatası, aykırı değer sorunu, farklılaşan çalışma karakteristikleri gibi çeşitli sebeplerinin olduğu söylenebilir (Borenstein ve diğ., 2010; Demir ve Başol, 2014; Schwarzer ve diğ., 2015). Aykırı değer sorunu, bir veya daha çok çalışmadan elde edilecek etki büyüklüklerinin diğerlerinden büyük ya da küçük değerler alarak farklılık göstermesiyle ortaya çıkar. Dolayısıyla en uygun modelin seçimi için meta analize dahil edilen çalışmalar arasından farklılık gösteren etki büyüklüklerinin aykırı değer olup olmadığının istatistiksel olarak belirlenmesinin kritik öneme sahip olduğu söylenebilir.

Bir meta analiz çalışmasında aykırı değer(ler) bulunması durumunda, aykırı değerin ait olduğu çalışma(lar) meta analizden çıkarılabilir (Viechtbauer ve Cheung, 2010). Ancak veriyi silmek yerine tüm verilerin dahil edildiği daha uygun bir modelin kullanımı da tercih edilebilir (Viechtbauer ve Cheung, 2010). Buradan yola çıkarak yapılan bu araştırmada aykırı değer içeren ve içermeyen meta analiz çalışmalarına dahil edilen araştırma sayılarına göre SEM ve REM altındaki yöntemlerin performansının karşılaştırılması amaçlanmıştır. Bu amaç doğrultusunda aşağıdaki araştırma problemine yanıt aranmıştır.

Aykırı değer içeren 5, 10, 20 ve 40 çalışmanın dahil edildiği ve aykırı değer içermeyen 4, 9, 19 ve 39 çalışmanın dahil edildiği meta analiz çalışmalarında, SEM ile REM altındaki momentler kestirimi DerSimonian Laird yöntemi (DL), en çok olabilirlik yöntemi (ML), kısıtlanmış en çok olabilirlik yöntemi (REML), deneysel Bayes yöntemi (EB), Sidik Jonkman yöntemi (SJ), Paule Mandel yöntemi (PM), Hunter Schmidt yöntemi (HS) ve Hedges yönteminin (HE) performansı kestirilen ortak etki büyüklüğü, güven aralığı için kapsama oranı ve heterojenlik ölçüleri (Q istatistiği, Higgins ve Thompson'ın I^2 istatistiği ve çalışmalar arası varyans için τ^2 istatistiği) bakımından nasıldır?

Yapılan bu araştırma, aykırı değer bulunan meta analiz çalışmalarında, SEM veya REM altında bulunan meta analiz yöntemlerinden hangisini kullanmanın diğerlerinden göreli olarak daha avantajlı olacağı hakkında fikir vermesi bakımından alanyazına sağlayacağı düşünülen katkılardan dolayı önemlidir.

Yöntem

Bu araştırma betimsel araştırma türündedir. Betimsel araştırmalar, belirli bir durumu mümkün olan en detaylı şekilde açıklamaya çalışan araştırmalardır (Fraenkel ve diğ., 2012). Bu araştırmada gerçekleştirilen meta analizlerde Doğuyurt'un (2013) çalışmasında yer alan gerçek veri setinin bir bölümü kullanılmıştır. Dolayısıyla dokuz farklı meta analiz yöntemini (bir SEM yöntemi, sekiz REM yöntemi) karşılaştırmak için *Öğretmenlerde duygusal tükenmişliğin cinsiyet değişkenine göre incelendiği* ve meta analize dahil edilme kriterlerine uygun olan çalışmalardan veri setleri oluşturularak toplam 72 meta analiz gerçekleştirilmiştir. Meta analize dahil edilen çalışma sayıları; aykırı değer bulunan meta analizler için 5, 10, 20 ve 40 olacak şekilde, aykırı değer veri setlerinden çıkartılarak gerçekleştirilen meta analizler içinse 4, 9, 19 ve 39 olacak şekilde ayarlanmıştır. Aykırı değer içeren veri setleri, her bir veri setinde bir adet

aykırı değer bulunacak şekilde oluşturulmuştur. Bu aykırı değerin ait olduğu birincil çalışma aykırı değer içeren tüm veri setleri için ortaktır. Aykırı değer içermeyen veri setleri ise aykırı ve etkili gözlemlerin belirlenmesine yönelik grafiksel (funnel plot-huni grafikleri ve radyal grafikler) ve istatistiki (DL yöntemi ile elde edilen student türü artık) değerlerin test edilip elde edilen sonuçlara dayanarak varlığı belirlenen aykırı değerin veri setinden silinmesiyle oluşturulmuştur. Ardından aykırı değer içeren ve aykırı değerin silindiği, farklı sayıda çalışmadan oluşan veri setleri kullanılarak SEM'e göre ve REM altında yer alan DL, ML, REML, EB, SJ, PM, HS ve HE yöntemlerine göre meta analizler gerçekleştirilmiştir. Tüm veri setleri için elde edilen meta analiz sonuçları; kestirilen ortak etki büyüklüğü, güven aralığı için kapsama oranı ve heterojenlik ölçülerindeki (*Q* istatistiği, Higgins ve Thompson'ın I^2 istatistiği ve çalışmalar arası varyans için τ^2 istatistiği) değişim bakımından karşılaştırılmış ve ardından kullanılan yöntemlerin performansları hakkında çıkarımda bulunulmuştur. Araştırma verilerinin analizi R programı ile gerçekleştirilmiştir. SEM ve REM ile gerçekleştirilen meta analizlerde Viechtbauer (2010) tarafından yazılan "metafor" paketinden yararlanılmıştır.

Sonuçlar

Meta analize dahil edilen çalışmaların aykırı değer içermesi durumunda; veri setindeki çalışma sayısı fark etmeksizin ortak (genel) etki büyüklüğü kestirimlerinde en fazla SEM'in etkilendiği, dolayısıyla aykırı değer içeren meta analizler için REM'in SEM'e kıyasla daha avantajlı olduğu söylenebilir. Ayrıca aykırı değer içeren veri setleri ile gerçekleştirilen meta analizlerde beklentiye uygun şekilde Q, I^2 ve çalışmalar arası varyans (τ^2) gibi heterojenlik ölçülerinin yüksek değerler aldığı sonucuna ulaşılmıştır.

Meta analize dahil edilen çalışmaların aykırı değer içermediği durumda ise ortak (genel) etki büyüklüğünün tüm veri setleri ve tüm yöntemler için benzer olarak kestirildiği sonucuna ulaşılmıştır. Bunların yanı sıra meta analize dahil edilen çalışma sayısındaki artışın aykırı değerin etki büyüklüğü kestirimi üzerindeki etkisini azalttığı ve heterojenliği düşürdüğü görülmüştür. Kullanılan meta analiz yöntemlerinin güven aralığı kapsama oranları incelendiğinde ise güven aralıklarının tüm veri setlerinde ve tüm yöntemlerde kestirilen etki büyüklüklerini içerdiği belirlenmiştir.

Ayrıca aykırı değer veri setinden çıkartılıp meta analizin yinelendiği tüm veri setleri için sadece HS yönteminin kullanıldığı durumlarda heterojenliğin tamamen ortadan kalktığı görülmüştür. HS yönteminin bu anlamda avantaj sağladığı söylenebilir. Araştırmanın en önemli sonuçlarından biri olarak, özellikle 20 ve daha fazla çalışmanın dahil edildiği meta analiz çalışmalarında kullanılan yöntemlerin, genel (ortak) etki büyüklüğü kestiriminde aykırı değerden daha az etkilendiği belirlenmiştir.

Kaynaklar

Borenstein, M., Hedges, L.V., Higgins, J. P. T., and Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed effect and random effect models for meta-analysis. *Research Synthesis Methods, 1*(1), 97-111. https://doi.org/10.1002/jrsm.12

Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach.* Sage.

Demir, S., & Başol, G. (2014). Bilgisayar destekli matematik öğretiminin (BDMÖ) akademik başarıya etkisi: Bir meta analiz çalışması. *Kuram ve Uygulamada Eğitim Bilimleri, 14*(5), 2013-2035. https://doi.org/10.12738/estp.2014.5.2311

Doğuyurt, M. F. (2013). *Öğretmenlerde tükenmişliğin çeşitli değişkenlere göre incelenmesi: Bir meta analiz çalışması* (Tez No: 350210) [Yüksek lisans tezi, Gaziosmanpaşa Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Fraenkel, J.R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.). McGraw-Hill, Inc.

Glass, G.V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3-8. https://doi.org/10.3102/0013189X005010003

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*(4), 486–504. https://doi.org/10.1037/1082-989X.3.4.486

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. Measuring inconsistency in meta-analyses. *BMJ, 327*(7414), 557-60. https://doi.org/10.1136/bmj.327.7414.557

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Sage.

Normand, S-L.T. (1999). Meta-analysis formulating, evaluating, combining, and reporting. *Statistics in Medicine, 18*, 321-59. https://doi.org/10.1002/(SICI)1097-0258(19990215)18:3<321::AID-SIM28>3.0.CO;2-P

Riley, R. D., Higgins, J. P., & Deeks, J. J. (2011). Interpretation of random effects meta-analyses. *BMJ (Clinical research ed.)*, *342*(2), 964-967. https://doi.org/10.1136/bmj.d549

Rudy, A. C. (2001). *A meta-analysis of the treatment of anorexia nervosa: A proposal.* Ithaca College.

Schmid, E. J., Koch, G. G, & LaVange L. M. (1991). An overview of statistical issues and methods of meta-analysis. *Journal of Biopharmaceutical Statistics, 1*(1), 103-20. https://doi.org/10.1080/10543409108835008

Schwarzer, G., Carpenter, J. R, & Rücker, G. (2015). *Meta-Analysis with R.* Springer international Publishing.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1-48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta-analysis. *Res Synth Methods, 1*(2), 112-125. https://doi.org/10.1002/jrsm.11

Akdeniz üniversitesi uluslararası öğrenci kabul sınavına (AKDENİZ YÖS-2019) ilişkin ölçme değişmezliğinin ve değişen madde fonksiyonlarının incelenmesi

Güçlü Şekercioğlu ve Ahmet Kütük

Anahtar kelimeler: Geniş ölçekli testler, ölçme değişmezliği, değişen madde fonksiyonları

Öz

Bu çalışmanın amacı, Akdeniz Üniversitesi tarafından 2019 yılında uygulanmış olan Uluslararası/Yabancı Uyruklu Öğrenci Kabul Sınavı'na (Akdeniz YÖS-2019) katılmış olan 5383 adaydan elde edilen puanlarda, testin orijinal dili olan Türkçe ile diğer diller arasında ölçme değişmezliğinin sağlanıp sağlanamadığının ve değişen madde fonksiyonlarının (DMF) bulunup bulunmadığını belirlenmesidir. Araştırmada öncelikle her bir grup için doğrulayıcı faktör analizleri (DFA) yapılmış daha sonrasında ölçme değişmezliği dört model bağlamında test edilmiştir. DMF belirleme yöntemlerinden Mantel-Haenszel ile Raju'nun Alan Ölçümleri seçilmiştir. Sonuç olarak, testin Genel Yetenek bölümünde dil değişkenine bağlı olarak ölçme değişmezliğinin sağlanamadığı, Matematik bölümünde ise ölçme değişmezliğinin sağlandığı görülmüştür. DMF incelemesi sonucunda da Genel Yetenek Testi'nde bir maddede B düzeyinde ve bir madde de C düzeyinde DMF olduğu görülmüştür. Testin Matematik bölümünde ise B ve C düzeylerinde DMF gösteren maddeye rastlanmamıştır.

Ensemble yöntemlerin eğitim alanında karşılaştırmalı olarak incelenmesi: Bagging ve Boosting algoritmaları

Hikmet Şevgin

Anahtar kelimeler: Ensemble learning, bagging, boosting, treenet, random forest

Öz

Bu çalışmanın amacı, Ensemble yöntemlerden Bagging ve Boosting algoritmalarını karşılaştırmalı olarak incelemek ve bu algoritmaları kullanan TreeNet ve Random Forest yöntemleri ile eğitim alanında elde edilen veriler üzerinde her iki yöntemin sınıflama performansını incelemektir. Ensemble yöntemler; sınıflamaya ve tahmine dayalı tekli yöntemlerden elde edilen sonuçların aksine, birden fazla yöntemin bir araya gelerek geçerliği ve güvenilirliği daha yüksek ve hata varyansı daha düşük olan topluluk yöntemlerini ifade etmektedir. Analizler için TreeNet ve Random Forest yöntemlerinin seçilmesindeki temel etken, her iki yöntemin tek tek karar ağaçlarını Bagging ve Boosting algoritmalarıyla bir araya getirip, her birinden elde edilen sonuçları birleştirerek tek bir sonuç oluşturan Ensemble yöntemlerden olmalarıdır. Eğitim alanında elde edilen verilerin arka planda farklı algoritmaları kullanan çeşitli istatistiksel yöntemler ile sınıflamaya ve tahmine dayalı sonuçlar üretmesi, öğrenciler adına verilen kararlar ve bu alanda verilen hizmetler için önemli görülmektedir. Çalışmaya ait veri seti ABİDE 2016 uygulamasına ait matematik puanları ile öğrencilere ait çeşitli demografik değişkenlerden oluşmaktadır. Çalışma grubu ABİDE örnekleminden rastgele çekilen 5000 öğrenciden oluşmaktadır. Kayıp veri silme ve atama işlemleri sonrası bu sayı 4568 düşmüştür. Sürekli bir özellik gösteren bağımlı değişken (öğrencilerin matematik başarısı) %25'lik birinci çeyrek (düşük matematik başarısı) ve %25'lik dördüncü çeyrek (yüksek matematik başarısı) dikkate alınarak 2'li kategorik hale getirilmiştir. Birinci çeyrek 1142 ve dördüncü çeyrek 1142 olmak üzere 2284 (1034 kız - 1250 erkek) öğrenci bu araştırmanın çalışma grubunu oluşturmaktadır. Yapılan analizler sonrası örneklem büyüklüğü, yüzdelik değerler ve performans kriterleri olmak üzere tüm koşullar dikkate alınarak her iki analiz yöntemi kendi içinde kıyaslanacak olursa, TreeNet yönteminin Random Forest yöntemine göre daha yüksek sınıflama performansı gösterdiği söylenebilir. Tek bir sınıflayıcı ya da tahminleyici yöntemin aksine, Bagging ve Boosting algoritmalarını kullanarak topluluk oluşturan birden fazla yöntemin sınıflamada ya da tahminde bulunması eğitim alanında elde edilen sonuçlar adına önemli olduğu düşünülmektedir. Bu doğrultuda arka planda Bagging ve Boosting algoritmalarını kullanan Ensemble yöntemlerin Eğitim Bilimleri çalışmalarında kullanılması önerilmektedir.

Yabancı dillerde görevli öğretim elemanlarının madde yazma eğitimine ilişkin ihtiyaçları ve görüşleri

Levent Yakar, Elif Kantarcıoğlu, Erkan Hasan Atalmış, Tuba Arabacı Atlamaz, Reyhan Ağçam ve Nuri Doğan

Anahtar kelimeler: Yabancı diller yüksekokulu, madde yazma eğitimi, çoktan seçmeli soru, madde istatistikleri

Giriş

Yabancı Diller Yüksekokulları (YADYO) uyguladıkları dereceli kur sistemi nedeniyle üniversitelerde en yoğun sınav takvimine sahip birimler arasındadır. Sık aralıklarla gerçekleştirilen sınavların hazırlanması, uygulanması, değerlendirilmesi ve sonuçların duyurulması ciddi bir iş yükü oluşturmaktadır. Bir sınava ilişkin aşamaların tamamlanmasının hemen ardından bir sonraki sınava hazırlıklar başlamaktadır. Bu yoğunluk nedeniyle soruların niteliği üzerinde yeterince durulmamaktadır. Pilot çalışma olmaksızın, özellikleri tam olarak bilinmeyen maddelerden oluşturulan testlere dayalı yapılan değerlendirmelerin kimi sorunları da olası beraberinde getirmesi kuvvetle muhtemeldir.

YADYO sınavlarında güvenirlik ve geçerlik soruları sıkça gözlenebilmektedir. Aydın ve diğ. (2016)'nin büyük ölçekteki 12 devlet üniversitesi YADYO'su gerçekleştirdiği çalışmaya katılan kurumlardan çoğu (f=9) hazırladıkları yeterlik sınavlarının ön denemelerini bir pilot çalışma olarak yapmadıklarını belirtmiştir. Bu kurumların genellikle sınav sorularının güçlüğünü ve maddelerin kalitesini sınavı hazırlayan ya da kurumda çalışan diğer öğretim elemanlarının deneyimleri doğrultusunda belirledikleri ve kurumlardan yedisinde sınav sonrası istatistiki analizlerin yapılmadığı görülmüştür. Çalışmaya katılan kurumların sınav birimlerinden çok azı (f=3) her yeterlik sınavı için yeni sorular hazırlamaktadır. Diğer kurumlarda ise (f=9) gerekli değişiklikler yapıldıktan sonra geçmişte kullanılan soruların bir kısmı yeni sınavlarda kullanılmakta ya da tekrar kullanılan sorular farklı hedef öğrencilere uygulanmaktadır. Buradan büyük üniversiteler de dahi gerekli ön çalışmaların yapılmadığı, sınav sonrası soru niteliklerinin araştırılmadığı ve yine de bu sorunlu soruların defalarca kullanıldığı sonucu ortaya çıkmaktadır. Bu sonuçları doğuran bir diğer değerli sonuç ise bu kurumların sınav birimlerinin sadece birinde ölçme ve değerlendirme uzmanı bulundurması olarak görülmektedir.

YADYO sınav birimlerinde görev alanların ölçme konusunda yeterince niteliğe sahip olmamaları sorunlara yol açabilmektedir. Yapılan bir çalışmada İngilizce seviye belirleme ve yeterlik sınavları test puanlarına ait güvenirliğin beklenen düzeyin altında olduğu ve testlerin çoğunluğunda ayırt edici madde sayısının az olduğuna ulaşılmıştır. Ayrıca test puanlarının sadece A1 seviyesinde elde edilen puanları düşük yordayıcılık gücü ile yordadığı diğer seviyeler için yordayıcı olmadığı sonucuna varılmıştır. Bu durum ise test puanlarının geçerliği ve güvenirliğine şüpheyle yaklaşmaya sebep olmaktadır (Gökçe-Taşdelen ve Kezer, 2020). Bu tür sonuçlar aynı kurum tarafından yapılan sınavların dahi standarda sahip olmadığını göstermektedir.

Türkiye'de, özellikle yabancı dilde eğitim yapan yükseköğretim kurumlarında uygulanan sınavların zorluğu, kolaylığı, yeterlik düzeyi gibi konular sıklıkla tartışılmaktadır (Üstünlüoğlu, 2011). Bu tartışmaların standart ölçme işlemlerinin yoksunluğundan kaynaklandığı söylenebilir. Bu sorun Aydın ve diğ. (2016)'nin çalışmasında şu şekilde dile getirilmektedir: "Türkiye'de içerisinde farklı becerileri barındıran yabancı dil sınavlarının uygulanmamakta ve farklı kurumların bahsedilen seviyeleri farklı algıladığı, yabancı dil sınavlarına dair bir standardın olmadığı bir gerçektir".

Çoktan seçmeli maddeler uygulama ve değerlendirme kolaylığı ve objektifliği sayesinde eğitimde en sık kullanılan ölçme araçlarındandır. Ancak hazırlama aşamasındaki zorluk göz önünde bulundurulduğunda ise çoktan seçmeli maddelerin herkes için kullanışlı olduğu söylenemez (Doğan, 2020). Avantajları sayesinde YADYO'larda sıklıkla kullanılan çoktan seçmeli maddelerin buralarda çalışan öğretim elemanlarınca ne anlam ifade ettiği, nasıl kullanıldığı tespit edilerek çoktan seçmeli maddelerin özelliklerinin öğretim elemanlarına tam olarak tanıtılması YADYO'larda testlere bağlı alınan kararların daha isabetli olması adına önem arz etmektedir.

Bu çalışmada YADYO'larda görevli öğretim elemanlarına çoktan seçmeli madde yazımı eğitimine yönelik ihtiyaçları ve bu eğitim sonrasında eğitime ilişkin görüşlerinin elde edilmesi amaçlanmıştır, Bu amaçla öğretim elemanlarına göre,

a) eğitim öncesinde madde yazmaya ilişkin yeterlik algıları ne düzeydedir?
b) madde yazmada yaşadıkları zorluklar nelerdir?
c) madde yazma eğitimi hangi düzeyinde verilmelidir?

araştırma soruları yanıtlanmaya çalışılmıştır.

Yöntem

Bu araştırmada katılımcıların öznel görüşlerinin toplanıp değerlendirildiği bir çalışma olması nitel araştırma desenine sahiptir (Yıldırım ve Şimşek, 2019). Araştırma, 2021 Mart-Mayıs ayları arasında İngilizce Hazırlık Sınıflarında Bilgisayarda Bireyselleştirilmiş Test Kullanımı isimli Tübitak projesi kapsamında düzenlenen Avrupa Dilleri için Ortak Çerçeve (ADOÇ) ve Çoktan Seçmeli Madde Yazma Eğitimine başvuran ve katılan YADYO İngilizce öğretim elemanlarına yönelik yapılmıştır. Eğitime katılmak için başvuru yapan 247 öğretim elemanı ilk araştırma sorusu için, başvuranlar arasından eğitime

seçilen 36 öğretim elemanı ise araştırmanın 2. ve 3. araştırma sorusu için çalışma grubunu oluşturmaktadır.

Araştırmanın verileri çevrimiçi veri toplama araçlarından Google Forms aracılığıyla elde edilmiştir. Eğitim için başvuru alma esasında öğretim elemanlarının madde yazmaya ilişkin yeterlik algılarını 1-Çok Yetersiz, 2-Yetersiz, 3-Orta Yeterlik, 4-Yeterli ve 5- Çok yetersiz düzeylerinden birini seçerek ifade etmeleri istenmiştir. Eğitim sürecindeki madde yazma oturumları öncesinde ise öğretim elamanlarından "Madde yazmada en çok zorlandığım nokta …." İfadesini tamamlamaları istenmiştir. Eğitim sonrasında ise öğretim elemanlarından madde yazma eğitiminin hangi düzeyinde verilmesi gerektiğine yönelik düşünceleri alınmıştır.

Google Formda elde edilen ilk ve üçüncü alt probleme ilişkin veriler Office Excel ile analiz edilmiş ve elde edilen sonuçlara ilişkin frekanslara ulaşılmıştır. İkinci alt problemde ise öğretim elemanlarının görüşleri nitel veri analiz yöntemlerinden betim analize tabi tutulacaktır.

Sonuçlar

Araştırmanın ilk sorusu için elde edilen yanıtlar incelendiğinde 247 öğretim elemanından 73 tanesi soru yazma eğitimi aldığını ifade etmiştir. Öğretim elemanlarının tamamının "Kendinizi soru yazmada ne derece yetkin hissediyorsunuz?" sorusuna verdiği yanıtlara bakıldığında ise öğretim elamanlarının ikisi 1-Çok Yetersiz, 20'si 2-Yetersiz, 107'si 3-Orta Yeterli, 93'ü 4-Yeterli ve 18'i 5- Çok yetersiz düzeyinde yeterlik algısına sahip olduğu görülmüştür. Grup geneli yeterlik algılarının orta-yeterli düzeyinde olduğu söylenebilir. Eğitime kabul edilen öğretim elamanlarının da kısmen yüksek de olsa benzer yeterlik algı düzeyine sahip olduğu söylenebilir.

Araştırmanın ikici sorusu için eğitim katılan öğretim elemanlarının 16'sı görüş belirtmiştir. Ön analiz sonucunda öğretim elemanlarının mantıklı ve güçlü çeldirici yazımında zorlandıkları geçici sonucu elde edilmiştir.

Araştırmanın üçüncü sorusu için ise eğitim katılan öğretim elemanlarının 29'u görüş bildirmiştir. Öğretim elemanlarından ikisi eğitimin lisans veya yüksek lisans düzeyinde verilmesi gerektiğini belirtirken, beşi ise madde yazma eğitiminin hizmetçi eğitimde verilmesi gerektiğini görüşünü bildirmişlerdir. Grubun %76'sını oluşturan (n=22) büyük bir çoğunluğu ise her iki düzeyde de madde yazma eğitiminin verilmesi gerektiğini bildirerek bu eğitime olan ihtiyacı ve gerekliliği vurgulamışlardır.

Kaynaklar

Aydın, B., Akay, O. E., Polat, O. M., ve Geridönmez, O. S. (2016). Türkiye'deki hazırlık okullarının yeterlik sınavı uygulamaları ve bilgisayarlı dil ölçme fikrine yaklaşımlar. *Anadolu Üniversitesi Sosyal Bilimler Dergisi, 16*(2), 1-20.

Doğan, N. (2020). Geleneksel ölçme ve değerlendirme teknikleri I: Yanıtı seçmeyi gerektiren ölçme araçları. N. Doğan (Ed.) *Eğitimde ölçme ve değerlendirme* içinde (ss. 114-139). Pegem Akademi.

Gökçe-Taşdelen, Z. ve Kezer, F. (2020). Bir İngilizce yeterlik ve seviye belirleme test puanlarının yordama geçerliği. *Kocaeli Üniversitesi Eğitim Dergisi. 3*(1), 1-25.

Üstünlüoğlu, E. (2011) Yeni bir sınav kültürü geliştirmek: Olasılık sanatı. *Eğitim Bilimleri Araştırmaları Dergisi, 1(*1), 19-31.

Yıldırım, A. ve Şimşek, H. (2019). *Sosyal bilimlerde nitel araştıma yöntemleri.* Seçkin Yayıncılık

Artistik buz pateni yarışması sonuçlarının genellenebilirlik kuramı ve çok yüzeyli Rasch modeli ile incelenmesi

İsmail Karakaya, Umur Öç ve Nazira Tursynbayeva

Anahtar kelimeler: Performans değerlendirilmesi, farklılaşan puanlayıcı davranışları, genellenebilirlik kuramı, çok yüzeyli Rasch ölçme modeli

Giriş

Performans değerlendirme sürecinde en çok karşılaşılan problemlerden birisi puanlayıcı yanlılığı (katılık/cömertlik)dır. Bu tarz puanlayıcı davranışına sahip puanlayıcılar birey performansını değerlendirirken bazı öğrencilere daha düşük puan verme eğiliminde iken (örneğin; erkek öğrencilere daha düşük not verme); bazı öğrencilere performansından daha yüksek puan verme (örneğin; kız öğrencilere daha yüksek not verme) eğiliminde olabilmektedir (Myford ve Wolfe, 2004).

Performansların ölçülmesi ve değerlendirilmesi, eğitim ve psikolojinin yanı sıra spor bilimlerinde de önemli bir yere sahiptir (Arsan, 2012). Performansı ölçmek için kullanılan yöntemler; davranışı gözlemlemek ve ölçmek, sonuca bakarak bireyin yeteneği hakkında bir değer yargısı yapmak, bireyi çeşitli düzeylerde gözlemleyerek, ara ara puan verme işidir (Turgut, 1993). Performansa dayalı değerlendirmede birey performans gösterirken hakem tarafından gözlenir ve bu ortamda değerlendirilir. Böylece hakem, bireyin hareketinden yorumlanabilir bir sonuç çıkarır ve bunu puanlar (McNamara, 1996).

Spor alanlarında yapılan tüm yarışmalarda sporcuların, müsabaka esnasında performanslarını en üst düzeyde sergilemeleri beklenmektedir ve bu performanslar farklı hakemler tarafından değerlendirilmektedir. Spor müsabakalarında sporcuların hakemler tarafından öznel olarak değerlendirilmesi, sporcuların kafasında soru işareti bırakmaktadır. Hakemlerin veya puanlayıcıların öznel değerlendirmeleri sonucunda ne kadar doğru, adil, geçerli ve güvenilir kararlar aldığı spor performansının ölçülmesinde önemli bir sorun olarak karşımıza çıkmaktadır. Birden fazla hakemin görüşlerinden oluşan bu değerlendirmeler, öznel kararlara karşı bir ölçü olmakla birlikte, güvenilirlik çalışmaları için birden fazla değişkenlik kaynağının bir arada ele alınması ihtiyacını vurgulamaktadır (Arsan, 2012).

Benzer şekilde artistik buz pateninde sporcu ile ilgili verilecek kararların, sürecin başlangıcında ve süreç içerisinde puanlayıcıların ölçütleri doğru ve objektif puanlama işlemlerinin gerçekleştirmesinde,

sporculardaki becerilerin geçerli ve güvenilir ölçme araçları ve yaklaşımları kullanılarak ölçülmesi önemlidir. Çalışmanın esas konusu olan; buz pateni, diğer spor türleriyle kıyaslandığında oldukça zarif ve estetik spor türü olarak bilinmektedir. Hızlı hareketlerle beraber göze çarpmayan, ince değişimleri barındırmaktadır. Artistik buz pateninde ise; yarışmacıların buz üzerinde bireysel, çiftler veya gruplar halinde, dönüş, zıplama, kaldırma ve atlama gibi bir dizi adım attıkları bir olimpik spordur. Müzik eşliğinde, dansın veya gösterinin konusuna uygun kostümlerle yapılır (ISU [International Skating Union], 2021). Bu nedenle artistik buz pateni sanatsal alanlarda teknik yetenekleri başta olmak üzere hem bir sanat hem de bir spor dalı olarak değerlendirilmektedir.

Alanyazın incelendiğinde, performans değerlendirmede puanlayıcılardan kaynaklı çok fazla hata türleri bulunmaktadır. Bunlardan en yaygın olanları: cömertlik ya da katılık, merkezi eğilim, ranj daralması (ranjın kısıtlanması) ve halo etkisidir.

Bu çalışmada, performans ölçümünde birden fazla puanlayıcı tarafından puanlama sonucunda elde edilen puanlara genellenebilirlik kuramı ve çok yüzeyli Rasch ölçüm modeli uygulanarak hangisinin daha kullanışlı sonuçlar verdiğini ve problemli puanlayıcı davranışların belirlemek amaçlanmıştır.

Yöntem

Yapılan bu çalışma Genellenebilirlik Kuramı'nın 2018-2021 yılları arasında yapılan Buz Pateni Yarışması verilerine uygulanmasını temel aldığı ve yarışmanın özelliklerini ortaya çıkarmaya çalıştığı için betimsel araştırma kapsamındadır. Bu çalışmanın çalışma grubu, 2018-2021 yılları arasında yapılan Buz Pateni Yarışması'na katılan toplam 8 sporcu ve yarışmada görev alan 6 puanlayıcıdan oluşmaktadır.

Bu çalışmada puanlayıcılar tarafından puanlanan beş program bileşen puanları üzerinden analizler yapılmıştır. Bu programlar; kayma becerileri (skating skills-SS), geçişler (transitions- TR), performans (performance-PE), yapı (composition) ve müziği yorumlamadır (interpretation of the music) (ISU Rules). Yarışmalarda tüm hakemler tüm bireyleri puanladığı için çaprazlanmış desen olarak değerlendirilmiştir. Çaprazlanmış desenlerde (tüm sporcuların tüm hakemler tarafından puanlandığı desenler) norm temelli karşılaştırmalar kullanıldığında hakemlerin cömertliği ya da katılığı sporcuların yarış sıralamasında farklılığa sebep olmayacaktır. Bu problemli puanlayıcı davranışı tüm sporcuları etkileyeceğinden durum dengelenecektir. Puan ortalamasının alındığı düşünüldüğünde puan dağılımındaki kayma sabit olacaktır. Bu sayede cömertlik ya da katılık dışında başka bir problemli puanlayıcı davranışı olmazsa norm temelli karşılaştırmaların tercih edildiği çaprazlanmış desenler oldukça güvenilir olacaktır (Wolfe, 2004).

Çalışmada kullanılan veriler de Genellenebilirlik Kuramı'nda tümüyle çaprazlanmış desen olarak tasarlanmıştır. Verilerden desene ait genellenebilirlik ve güvenirlik katsayıları hesaplanmıştır. Ayrıca problemli puanlayıcı davranışlarından olan katılık ve cömertlik de incelenmiştir. Yapılan çalışmada öncelikle puanlayıcılardan kaynaklanan varyansı belirlemek için EduG (Swiss Society for Research in Education Working Group, 2010) programından faydalanılmıştır. Sonrasında puanlayıcılardan

kaynaklanan varyansın yüksek olduğu yarışmalarda problemli puanlayıcı davranışlarını belirlemek amacıyla FACETS (Linacre, 2017) bilgisayar programı kullanılmıştır.

Sonuçlar

Yapılan çalışmada 2018-2021 yılları arasında yapılan Buz Pateni Yarışma sonuçları incelenmiştir. Sonuçlar değerlendirilirken tümüyle çaprazlanmış desen bxgxp (b: birey, g: görev, p: puanlayıcı) uygulanmıştır. Verilerde program puanları kullanılmış ve 6 puanlayıcı, 5 ölçüt ve 8 birey ile veriler analiz edilmiştir. Değerlendirme sonucunda 240 etkileşim meydana gelmiştir. Bu araştırmada birey, ölçütler ve puanlayıcılar olmak üzere üç yüzey bulunduğundan her bir yüzey için ölçüm değerleri ve uygunluk istatistikleri incelenmiştir.

Mevcut yarışmada performansın değerlendirilirken 3 anlamlı etkileşimin yanlılık değerlerinin işaretleri incelendiğinde yanlılık değerleri pozitif işaretli olan 2 etkileşimde farklılaşan puanlayıcı cömertliği, yanlılık değerleri negatif işaretli olan 1 etkileşimde ise farklılaşan puanlayıcı katılığı olduğu görülmektedir. Yarışmada bulunan puanlayıcıların istatistiksel olarak benzer puanlayıcı davranışları gösterdikleri belirlenmiştir. Tespit edilen farklılaşan puanlayıcı davranışlarını olabildiğince en aza indirebilmek geçerli ve güvenilir sonucu elde edebilmek için puanlayıcılara puanlayıcı eğitimi verilmesi planlanmaktadır.

Kaynaklar

Arsan, N. (2012). *Buz pateninde hakem değerlendirmelerinin genellenebilirlik kuramı ve rasch modeli ile incelenmesi* (Tez No: 330403) [Doktora Tezi, Hacettepe Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

ISU (2021). *Special regulations and technical rules: Synchronized skating.* https://www.isu.org/synchronized-skating/rules/sys-regulations-rules/file

Linacre, J. M. (2017). *A user's guide to FACETS: Rasch-model computer programs.* MESA Press.

McNamara, T. F. (1996). *Measuring second language performance.* Longman

Myford, C.M., and Wolfe, E.W. (2004) Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189-227.

Turgut, M. F. ve Baykul, Y. (1992). *Ölçekleme teknikleri.* ÖSYM.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46*, 35-51. http://psycnet.apa.org/record/2004-19990-003

Bayesçi yaklaşık ölçme değişmezliği: Önsellerin ve örneklem büyüklüğünün kestirimlere etkisi

Gizem Uyumaz, Gözde Sırgancı ve Akihito Kamata

Anahtar kelimeler: Geçerlik, Bayeşci kestirim, esneklik payı, Monte Carlo simülasyonu, çoklu grup karşılaştırması

Giriş

Farklı alt gruplarda yer alan bireylerden toplanan verilerde gruplar arası karşılaştırmaların yapıldığı araştırmalarda yöntemsel bir hataya düşmemek adına ölçme değişmezliğinin incelenmesi gerekir. Ölçme değişmezliği, gözlenen ve örtük değişkenler arasındaki ilişkinin, incelenen alt gruplar arasında aynı olmasıdır (Drasgow ve Kanfer, 1985; Widaman ve Reise, 1997). Ölçme değişmezliği bir geçerlik sorunudur ve çalışmada ele alınan alt gruplar arasında ölçme değişmezliği sağlanmıyorsa gruplar arası karşılaştırmalar yapılamaz. Ölçme değişmezliğinin belirlenmesinde kullanılabilecek farklı yöntemler bulunmaktadır (Muthén ve Asparouhov, 2013; Asparouhov ve Muthén, 2014; Kim ve diğ., 2017). Bu çalışmada ölçme değişmezliğinin belirlenmesinde kullanılan Bayesçi Yaklaşık Ölçme Değişmezliği (BYÖD) yöntemine ilişkin incelemeler yapılmıştır.

BYÖD, Muthén ve Asparouhov (2012) tarafından Bayesçi yapısal eşitlik modellemesine dayalı olarak geliştirilmiştir. Tam ölçme değişmezliğinde faktör yükleri ve kesenler gruplar arasında eşit olacak şekilde kısıtlanırken, BYÖD'de parametreler arası bu eşitlik varsayımının daha esnek olmasına izin verilir. Bu esneklik Van De Schoot, Kluytmans, Tummers, Lugtig, Hox ve Muthén (2013) tarafından esneklik payı (wiggle room) olarak tanımlanır. Faktör yükleri ve kesenlerin gruplar arasındaki farklılaşması sıfıra oldukça yakın olarak varsayılır ancak tam olarak sıfır değildir. Yaklaşık olarak karşılaştırılabilirliği sağlamak için parametre farkları minimumda tutulmaktadır. Esneklik payı önsel dağılım ile belirlenir ve bu pay genellikle ortalaması sıfır varyansı küçük bir değer olan normal dağılımdan gelir, $N(0, \upsilon)$. Önsel dağılımda varyans, farklılıkların sıfır (yakın) olduğu varsayımıyla farklı güven düzeylerini temsil edecek şekilde değiştirilebilir. Geniş varyans değeri, önselin daha az bilgi verici olduğu anlamına gelir. Varyans 1'den küçük ise dar, yani önsel daha fazla bilgi vericidir (Asparouhov ve Muthén, 2017). BYÖD ile ölçme değişmezliğinin test edilme aşamaları geleneksel yöntemle benzer olup kestirim yöntemi bayes ile yapılmaktadır. BYÖD yöntemi ile yaklaşık ölçme değişmezliği incelemesinde ilk adımda gruplar arası faktör yükleri ve kesenlerinin eşitliğine dair herhangi bir kısıt konmadan bayes kestirimi kullanılarak ÇG-DFA

yapılır (yapısal model). Bu aşamada faktör yükü ve kesenlerinin ortalaması sıfır varyansı geniş olan bir önsel dağılıma göre kestirim yapılır. Yapısal modeli tanımlamak için ise tüm gruplarda faktör ortalaması 0 ve varyansı 1 olarak kısıtlanır ve tüm faktör yükleri ve kesenleri serbest kestirilir. İkinci adımda yaklaşık ölçme değişmezliğini sağlayacağı düşünülen faktör yükleri ve kesenlerine ilişkin farkları yansıtan önsel varyans hesaplanır. Üçüncü adımda önceden belirlenmiş önsel varyans dahil olmak üzere farklı önsel varyanslara sahip bir dizi yaklaşık metrik ölçme değişmezliği modeli oluşturulur ve en uygun model, model karşılaştırmalarıyla seçilir. Seçilen modelin önsel varyansı önceden belirlenen değerden küçükse veya ona eşitse, yaklaşık metrik değişmezlik sağlanır; aksi takdirde, yaklaşık metrik değişmezlik reddedilir. Yaklaşık metrik değişmezlik modelleri için kesenler serbest bırakılarak kestirim yapılır. Model tanımlaması için, faktör varyansları serbest kestirilirken tüm faktör ortalamaları sıfıra kısıtlanır. Yaklaşık metrik değişmezlik modeli iyi bir uyum sağlıyorsa, hem faktör yüklerinin hem de kesenlerin gruplara arası farkları için bir dizi önsel aracılığı ile bir önceki adımdaki işlemler tekrarlanarak skaler yaklaşık ölçme değişmezliği test edilir. Kestirilen modelin önsel varyansı gruplar arası kesenlerin farkları için önceden hesaplanan önsele eşit veya ondan küçükse yaklaşık ölçme değişmezliği (skaler) sağlanır. Bu da gruplar arasında önemli bir ölçme değişmezliğinin olmadığını gösterir. Modelin tanımlanması için de bir grubun faktör ortalaması sıfıra sabitlenir ve diğer faktör ortalamaları ve faktör varyansları serbest olarak kestirilir. Bayesçi yaklaşık ölçme değişmezliğini test etmek için farklı yaklaşımlar bulunmaktadır (van de Schoot ve diğ., 2013). Bu çalışmada hem faktör yükleri hem de kesenlerin gruplar arasında eşit tutulduğu yaklaşım ile yaklaşık ölçme değişmezliği test edilmiştir. Bu çalışmada ölçme değişmezliğinin belirlenmesinde kullanılan yöntemlerden biri olan bayesçi yaklaşık ölçme değişmezliği yönteminde kullanılan farklı önsellerin ve örneklem büyüklüğünün kestirimlere etkilerinin belirlenmesi amaçlanmıştır.

Yöntem

Araştırmada simülatif verilerin üretilmesinde gerçek veri seti temel alınmıştır. Gerçek veri seti olarak, Avrupa Sosyal Araştırması 2018 uygulamasında kullanılan psikolojik ölçme araçlarından *Kendini Geliştirme Anketi* kullanılmıştır. Anket Schwartz (2003) tarafından iki temel değeri (başarı ve güç) ölçmek için geliştirilmiştir. Analizlerde kullanılmak üzere bu dört maddeyi 11 ülkeden yanıtlayan 17580 katılımcının verisi kullanılmıştır. Verilerin analizine başlanmadan önce varsayımlar incelenmiştir. Ülkeler arasındaki ölçme değişmezliğinin belirlenmesinde ilgili yöntemin adımları için gerekli işlemler yapılmıştır. Çalışmada analizler Mplus 8.5 programında yapılmıştır. Çalışma kapsamında gerçek verilere dayalı olarak yapılan simülasyonlarda üç farklı örneklem büyüklüğü (50, 100 ve 500) ile altı farklı önsele (.001, .005, .010, .025, .050, .100) ilişkin incelemeler yapılmıştır. Farklı örneklem büyüklüklerinde hangi önselin daha uygun olduğunun belirlenmesi hedeflenmiştir. 18 farklı koşul için 400 replikasyon yapılmıştır.

BYÖD, Bayes analizi için mevcut olan bir model değerlendirme stratejisi ile değerlendirilebilir. BYÖD analizinde, tekrarlanan verilerin gözlemlenen verilere dayalı olanlardan daha büyük olmasına dayanan test istatistiklerinin olasılığını veren sonsal tahmin p değeri (posterior predictive p, PPP), sapma bilgi kriteri olan DIC (Spiegelhalter ve diğ., 2002) ve BIC kullanılmıştır. Model veri uyumu sağlandığında, gözlenen veriden elde edilen test istatistikleri ile tekrarlanan veriden elde edilen istatistiklerin benzer olması ve bir değerin

diğerinden daha büyük olma oranının (PPP) neredeyse yarı yarıya (.50) olması beklenir. Bununla birlikte, bir model yanlış tanımlandığında (yani, doğru şekilde belirlenmiş modelin dağılımına ait olmadığında), PPP'nin aşırı olması beklenir. Bu nedenle, aşırı PPP değerleri (örneğin,.05 veya .01'den küçük) modelin yanlış tanımlanmasının bir göstergesi olarak kabul edilir (Muthén ve Asparouhov, 2012; Sinharay ve diğ., 2006). Yaklaşık ölçme değişmezliği testinde, değişmezliğin miktarı tanımlanan önsel varyansla kıyaslandığında büyük olursa PPP'nin düşük olması beklenir ve bu da modelin yanlış tanımlandığını gösterir. Daha küçük bir BIC veya DIC değeri, veriye daha iyi uyan bir modeli gösterir.

Bulgular

Gerçek veri setine ilişkin kestirimler Tablo 1'de sunulmuştur.

Tablo 1

Uyum Değerleri

Ülke	PPP (güven aralığı)	DIC
1	0.604 (-16.303, 12.236)	17197.747
2	0.400 (-12.799, 15.970)	7171.879
3	0.269 (-10.311, 18.545)	22942.104
4	0.533 (-15.293, 13.598)	25309.884
5	0.111 (-5.833, 22.960)	26628.432
6	0.590 (-16.204, 12.618)	15759.056
7	0.304 (-11.097, 18.057)	15925.241
8	0.492 (-14.314, 14.382)	23433.616
9	0.561 (-15.819, 13.402)	14160.740
10	0.130 (-6.589, 22.262)	17315.179
11	0.444 (-13.740, 15.263)	26460.305

Simülasyon analizleri bulguları Tablo 2'de sunulmuştur.

Tablo 2

Uyum İyiliği Değerleri

Örneklem Büyüklüğü		PPP	DIC	BIC	RMSEA	CFI	TLI
50	0.001	0.094	6638.665	7929.842	0.042	0.872	0.981
	0.005	0.123	6628.014	7924.777	0.039	0.885	0.983
	0.010	0.158	6630.029	7917.400	0.036	0.900	0.985
	0.025	*0.251*	*6629.742*	*7904.169*	*0.031*	*0.929*	*0.989*
	0.050	0.347	6650.991	7891.791	0.026	0.952	0.992
	0.100	0.422	6664.081	7881.038	0.021	0.967	0.995
100	0.001	0.012	13491.55	15090.990	0.037	0.894	0.983
	0.005	0.036	13467.34	15072.710	0.032	0.917	0.988
	0.010	0.084	13460.857	15057.498	0.028	0.937	0.990
	0.025	*0.247*	*13491.100*	*15031.410*	*0.200*	*0.968*	*0.995*
	0.050	0.395	13484.140	15015.130	0.012	0.985	0.998
	0.100	0.472	13472.460	15005.230	0.008	0.992	0.999

(devam ediyor)

Tablo 2 (devam)

Örneklem Büyüklüğü		PPP	DIC	BIC	RMSEA	CFI	TLI
500	0.001	0.000	70207.355	71758.039	0.060	0.916	0.953
	0.005	0.000	69970.683	71595.511	0.034	0.967	0.984
	0.010	0.042	69885.369	71535.216	0.021	0.985	0.994
	0.025	*0.336*	*69716.253*	*71491.244*	*0.008*	*0.997*	*0.999*
	0.050	0.467	69524.444	71482.097	0.003	0.999	1.000
Gerçek veri	*0.025*	*0.135*	*211380.694*	*%95 [-20.947, 74.825]*			

Bayes analizlerine çok küçük bir önselle başlanmış ve gözlemlenen ve tekrarlanan ki-kare değeri arasındaki fark için DIC, PPP ve %95 güvenilirlik aralıkları izlenerek art arda artırılarak gerçek veri seti için en iyi uyum sağlayan önsel seçilmiştir. Ayrıca, PPP ve = 0.025 için gözlemlenen ve tekrarlanan ki-kare değerleri arasındaki fark için %95 güvenilirlik aralığı sınırları, tam metrik ölçme değişmezliği modelinden önemli ölçüde farklı olmadığından, .025 yeterli kabul edilmiştir. Tablo 2 incelendiğinde, kestirimlerin örneklem büyüklüğüne göre farklılaşmadığı görülmektedir. Uygun önsel olan .025'e ilişkin kestirimlerle, hatalı belirlenen önsellere ilişkin kestirimler karşılaştırıldığında, değişimler açısından DIC indeksinin, doğru önseli belirlemede daha doğru yönlendirdiği bulunmuştur.

Sonuçlar

Alanyazın incelendiğinde, genel Bayes için önsellerin seçimi ayrıntılı bir şekilde çalışılmış olsa da (van de Schoot ve diğ., 2013), yaklaşık ölçme değişmezliğinde önsellerin seçimine ilişkin kılavuz ilkelerin büyük ölçüde eksik olduğu belirtilmiştir (Pokropek ve diğ., 2020). Van Erp ve diğ., (2019) tarafından, modelin bir düzenlenme biçimi olarak önsellerin kullanılması önerilmiştir. Pokropek ve diğ. (2020) tarafından yapılan çalışmada da buradan yola çıkılmıştır ve iyi belirlenmiş önsellerin daha kesin güvenirlik aralıkları ve sonsal standart sapmalarla sonuçlandığı BIC, DIC ve PPP indekleri üzerinden karşılaştırmalar yapılarak bulunmuştur. İyi belirlenmiş önseller, daha doğru tahminleri sağlamaktadır. Doğru önseli belirlemede genel olarak en etkili indeksin DIC olduğu kanıtlanmıştır. Bu çalışmada da düşük önsellilerle başlanarak ve bunları BIC, DIC ve PPP ile RMSEA, CFI ve TLI uyum indekslerini kullanarak daha yüksek önselli modellerle karşılaştırılmıştır. Bu çalışmada ele alınan koşullarda da, özellikle DIC ve bir dereceye kadar da PPP indeksinin, doğru önseli belirlemede potansiyel bir yardımcı olarak kullanılabileceği bulunmuştur. Ancak farklı veri setleri üzerinde, aynı ve farklı önseller ile örneklem büyüklükleri ve uyum indeksleri ele alınarak ek çalışmalar yapılmalıdır.

Kaynaklar

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(4), 495–508. https://doi.org/10.1080/10705511.2014.919210

Asparouhov, T., & Muthén, B. (2017). Prior-posterior predictive p-values. *Mplus Web Notes*, *22*.

Byrne, B.M., Shavelson, R.J., & Muthén, B.O. (1989). Testing for equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70*(4), 662-80.

Fan, X., & Sivo, S. A. (2009). Using Δ-goodness-of-fit indexes in assessing mean structure invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(1), 54-69. https://doi.org/10.1080/10705510802561311

Jöreskog, K.G., Sörbom, D., Du Toit, S.H.C., & Du Toit, M. (2001). *LISREL 8: New statistical features* (3rd ed.). Lincolnwood, IL: Scientific Software International.

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement invariance testing with many groups: a comparison of five approaches, *Structural Equation Modeling: A Multidisciplinary Journal, 24*(4), 524-544. https://doi.org/10.1080/10705511.2017.1304822

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological methods*, *17*(3), 313.

Muthén, B., & Asparouhov, T. (2013). BSEM measurement invariance analysis. *Mplus web notes*, *17*, 1-48.

Pokropek, A., Schmidt, P., & Davidov, E. (2020). Choosing priors in Bayesian measurement invariance modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(5), 750-764.

Schwartz, S. H. (2003). A proposal for measuring value orientations across nations. *Questionnaire package of the european social survey*, *259*(290), 261.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30,* 298–321. https://doi.org/10.1177/0146621605285517

Sokolov, B. (2019). *Sensitivity of goodness of fit indices to lack of measurement invariance with categorical indicators and many groups.* Higher School of Economics Research Paper No. WP BRP 86/SOC/2019. https://wp.hse.ru/data/2019/07/09/1480015921/86SOC2019.pdf

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series B (statistical methodology)*, *64*(4), 583-639. https://doi.org/10.1111/1467-9868.00353

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the MI literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-70. https://doi.org/10.1177/109442810031002

Van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, *4*(770). 1-15.

Van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, 89, 31–50. https://doi.org/10.1016/j.jmp.2018.12.004

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). American Psychological Association. https://doi.org/10.1037/10222-009

ÖSYM sınavlarında engelli salon görevlisi olarak bulunan akademisyenlerin engelli adaylara yönelik sınav uygulamaları hakkındaki görüşleri

Mustafa İlhan, Melek Gülşah Şahin ve Bayram Çetin

Anahtar kelimeler: Geniş ölçekli testler, engelli adaylara yönelik sınav uygulamaları, engelli salon görevlisi, okuyucu, işaretleyici

Giriş

Ölçme seçme ve yerleştirme merkezi (ÖSYM), uyguladığı sınavlara katılan engelli adaylar için sınav koşullarında diğer adaylarınkinden farklı birtakım düzenlemeler yapmaktadır. Adaya mevcut engel durumu ile ilgili ilaç, araç-gereç, cihaz ve materyalleri salona getirme imkânı sunulması, okuyucu ve işaretleyici yardımı sağlanması, sınavda ek süre verilmesi, kitapçığın yazı puntosunun büyütülmesi, adayın bazı maddelerden muaf tutulması ve tek kişilik salonlarda sınava alınması ÖSYM'nin engelli adaylara yönelik uygulamaları arasındadır. ÖSYM'nin engelli adaylar için hayata geçirdiği bu ve benzeri uygulamaların olumlu yanları ile işlemeyen yönlerinin tespit edilmesi, engelli adaylara yönelik sınav uygulamalarının iyileştirilmesi açısından oldukça önemlidir. Bu kapsamda araştırmada; ÖSYM tarafından engelli adaylar için uygulanan sınavlarda, engelli salon görevlisi olarak bulunan akademisyenlerin bu sınavların uygulanma koşullarına ilişkin görüşlerinin incelenmesi amaçlanmıştır. Alanyazına bakıldığında ulusal düzeydeki sınavlarda engelli adaylara yönelik uygulamaları konu edinen araştırmalar bulunduğu görülmektedir. Şenel (2015) tarafından yapılan çalışmada görme engelli öğrencilerin üniversiteye giriş sınavı deneyimleri incelenmiştir. Karabay (2016) canlı okuyucu ve bilgisayar destekli okumanın görme engelli öğrencilerin test başarıları üzerindeki etkisini araştırmıştır. Özarkan ve diğ. (2017) temel eğitimden ortaöğretime geçiş kapsamında uygulanan 2015–2016 Eğitim Öğretim yılı birinci dönem merkezi ortak sınavı matematik alt testindeki maddelerin, adayların görme engeli durumu açısından değişen madde fonksiyonu gösterip göstermediğini test etmiştir. Şenel (2017) bilgisayar ortamında bireye uyarlanmış testlerin görme engelli öğrencilere uygunluğunu belirlemeye çalışmıştır. Çobanoğlu Aktan, Aksu ve Eser (2018) Türkiye ve Amerika'da engelli öğrenciler için yapılan merkezi sınavları yasal sorumluluklar, uygulama yöntemleri ve geçerlik açısından karşılaştırmıştır. Doğuş ve diğ. (2020) ise görme engelli kişilerin merkezi sınav düzenlemelerine dair görüşlerini incelemiştir. Alanyazında engelli adaylara ait sınav koşullarını, doğrudan bu sınavlarda görev alan akademisyen

görüşlerine dayalı olarak inceleyen bir çalışmaya rastlanmamıştır. Dolayısıyla çalışmanın literatüre katkı sağlayacağı düşünülmektedir.

Yöntem

Araştırma fenomenolojik desene göre yürütülmüştür. Çalışma grubu ölçüt örneklemeye göre belirlenmiş ve ÖSYM'nin uyguladığı sınavlarda en az üç defa engelli salon görevlisi olarak görev alan 12 akademisyen araştırmanın katılımcılarını oluşturmuştur. Çalışmanın verileri araştırmacılar tarafından geliştirilen ve açık uçlu dört maddeden oluşan bir anket formu ile toplanmıştır. Veri toplama sürecine başlanmadan önce Dicle Üniversitesi Sosyal ve Beşerî Bilimler Etik Kurulu Başkanlığı'na etik onay başvurusunda bulunulmuştur. Araştırmacılar tarafından sunulan dilekçeye 18.03.2020 tarih ve 34061 sayılı yazıyla cevap verilmiş ve çalışmanın bilimsel etiğe uygun olduğu bildirilmiştir. Etik kurul oluru alındıktan sonra veri toplama işlemine başlanmış ve veri doygunluğuna ulaşılan noktaya kadar sürece devam edilmiştir. Elde edilen veriler içerik analizi ile çözümlenmiştir. Analiz kapsamında öncelikle yanıt aranan alt problemler ve anket formundaki maddeler ile ilişkili olarak dört tema belirlenmiştir. İkinci aşamada ise katılımcıların görüşleri iki araştırmacı tarafından her bir tema altında incelenerek alt kategoriler oluşturulmuştur. Güvenirlik için iki araştırmacının yaptığı kodlamalar arasındaki tutarlığa bakılmıştır. Bu amaçla Miles ve Huberman (1994)'ın önerdiği formülden yararlanılmış ve her bir tema için kodlayıcılar arasındaki uyum sırasıyla .90, .80, 1.00 ve .96 olarak hesaplanmıştır. Ayrıca görüş ayrılığı yaşanılan kategorileri tartışıp kategori isimlerinde bütünlüğü sağlamak amacıyla kodlamayı yapan araştırmacılardan farklı ve daha önce ÖSYM'nin uyguladığı sınavlarda engelli salon görevlisi olarak görevlendirilmiş bir ölçme ve değerlendirme uzmanının görüşüne başvurulmuştur. Araştırmacılar ve ilgili uzman online olarak bir ara gelmiş ve kategoriler üzerinde tartışarak görüş birliğine varmaya çalışmıştır. Kategorilere son hali verildikten sonra katılımcı teyidi alınarak araştırmanın iç geçerliği (inandırıcılığı) arttırılmaya çalışılmıştır. Yine geçerlik kapsamında, bulgular sunulurken katılımcı görüşlerinden doğrudan alıntılara yer verilerek ayrıntılı betimleme yapılmıştır.

Sonuçlar

Araştırmada, katılımcıların ÖSYM'nin engelli adaylara yönelik sınav uygulamaları hakkındaki görüşleri olumlu yönler, sınırlı yönler, göreve ilişkin zorluklar ve öneriler şeklinde dört tema altında toplanmıştır. Olumlu yönler teması; "eğitimde fırsat eşitliği, uygulama koşulları, ortam ve görevli motivasyonu" kategorilerinden oluşmuştur. Sınırlı yönler teması altındaki kategoriler "görevli seçimi, madde yapısı, uygulama koşulları ve ortam" olarak belirlenmiştir. Göreve ilişkin zorluklar teması; "okuma, görevli seçimi, ayrıcalık talebi ve sınav kuralları" kategorilerinden meydana gelmiştir. Bununla beraber, bu tema altındaki görüşlerin bir kısmının ÖSYM'nin öngördüğü uygulamalardan çok adayların sınav sırasındaki tavırlarıyla ilgili olduğu saptanmıştır. Öneriler teması altındaki kategoriler ise "görevli seçimi, uygulama koşulları, madde yapısı, görevlilerin eğitime alınması, sınav kuralları ve ortam" olarak sıralanmıştır. Araştırmada ulaşılan bulgular ile Doğuş ve diğ. (2020)'nin görme engelli adayların merkezi

sınav düzenlemelerine ilişkin görüşlerini incelediği çalışmanın sonuçları arasında özellikle okuyucu niteliğine ilişkin görüşler açısından paralellikler olduğu tespit edilmiştir.

Kaynaklar

Çobanoğlu Aktan, D., Aksu, G. ve Eser, M. T. (2018). Türkiye ve Amerika'da engelli öğrenciler için yapılan geniş ölçekli sınavların yasal sorumluluklar, uygulama yöntemleri ve geçerlik açısından incelenmesi. *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, *14*(1), 69–83. https://doi.org/10.17860/mersinefd.322551

Doğuş, M., Aslan, C., & Çakmak, S. (2020). Görme engelli bireylerin merkezi sınav düzenlemelerine ilişkin görüşleri. *Eğitim ve Toplum Araştırmaları Dergisi*, *7*(1), 219–247. https://dergipark.org.tr/tr/pub/etad/issue/55359/697087

Karabay, E. (2016). *Canlı okuyucu ve bilgisayar destekli okumanın görme engelli öğrencilerin test başarıları üzerindeki etkilerinin karşılaştırılması* (Tez No. 431283) [Doktora tezi, Ankara Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Miles, M. B., and Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed). Los Angeles: Sage.

Özarkan, H. B., Kucam E. ve Demir, E. (2017). Merkezi ortak sınav matematik alt testinde değişen madde fonksiyonunun görme engeli durumuna göre incelenmesi. *Current Research in Education*, *3*(1), 24–34.

Şenel, S. (2015). Görme engelli öğrencilerin üniversite giriş sınavı deneyimleri. *Hacettepe Araştırmaları Dergisi*, *1*(1), 1–17.

Şenel, S. (2017). *Bilgisayar ortamında bireye uyarlanmış testlerin görme engelli öğrencilere uygunluğunun incelenmesi* (Tez No. 456700) [Doktora tezi, Ankara Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Arkadaş tercihlerinin belirlenmesi: Çok boyutlu ölçekleme uygulaması

Ceren Tunaboylu Demir ve Duygu Anıl

Anahtar kelimeler: Çok boyutlu ölçekleme, proxscal, yakınlık analizi

Giriş

İnsan sosyal bir varlıktır. Belirli bir toplumun içinde doğar ve o toplumun değerleri ile büyür. Moslow (1970) ihtiyaçlar hiyerarşisinde ilk aşamada nefes alma ve yeme-içme; ikinci aşamada kendini güvende hissetme gelirken, üçüncü aşamada ise başkaları ile ilişki kurmak, kabul edilmek ve bir yere ait olmak gelmektedir. Bir başka ifadeyle bir insanın kendini gerçekleştirebilmesi için besin ihtiyacını karşılamak ve güvende hissetmekten hemen sonra başkaları ile ilişki kurmak gelmektedir. Bireyin çevresiyle ilişki kurması ailede başlar, arkadaş grubuyla devam eder. Arkadaş ilişkilerinde başarılı olan bireyler sağlıklı bir hayat sürdürebilirler.

Türk Dil Kurumu Sözlüğünde "arkadaş" kelimesi 'birbirlerine karşı sevgi ve anlayış gösteren kimselerden her biri, yâren, yoldaş' ve 'bir ortamda birlikte bulunanlardan her biri' olarak tanımlanmaktadır. Kişilerin sosyal hayatta karşılaştıkları insanları arkadaş olarak kabul etmelerine ilişkin inançları arkadaşlık algısı olarak ifade edilmektedir. "Eğitim seviyesi yüksek insanlar güven verici insanlardır" düşüncesine sahip bir insanın arkadaşlık ilişkilerinde sahip olduğu arkadaşlık algısına örnek olarak verilebilir. Kişilerin zihinlerindeki genel beğeni/kabulleri açısından bir insanı diğerinin önüne koydukları zihinsel süreç ise arkadaş tercihlerini oluşturur. Örneğin "Eğitim seviyesi yüksek bireylerle arkadaş olmak istiyorum." Bir arkadaş tercih ifadesidir. Arkadaş seçimleri ise kişilerin mutlak arkadaş oldukları kişilerin ifade etmektedir. Seçim ve tercih arasındaki en önemli fark tercihlerin mutlak bir seçim gerektirmemesidir. Sonuç olarak algılar tercihlerin oluşmasını sağlarken tercihler de seçimlerin ortaya çıkmasını sağlar.

Zamanının çoğunu birlikte geçiren çocuklar arkadaşını "İyi Arkadaş" ve "Kötü Arkadaş" olarak iki kısma ayırmaktadırlar. İyi arkadaşın özelliklerinin başında "iyi olan", "sırdaş olan", "güvenilir olan" gelmektedir. Kötü özelliklerinin başında ise "kötü olan", "tembel olan" , "geçimsiz olan", "terbiyesiz olan" gelmektedir (Gündoğdu, 2003). Peki, biz arkadaşlarımızı neye göre, nasıl tercih ediyoruz? Arkadaş seçimlerimiz ve tercihlerimizde nelerden etkileniyoruz? Bu sorular ve cevaplar belki farkına bile varmadan hayatımızın akışı içerisinde süregelen sorular ve cevaplar.

Bu araştırmada sosyal hayatta karşılaştığımız bir insanı arkadaş olarak seçmemizde rol oynayan tercihlerimizin ölçeklenmesi amaçlamıştır. Araştırmada "arkadaş" bir ortamda birlikte bulunulan bir kişi; "yakın arkadaş" zamanınızın çoğunu beraber geçirmeyi tercih ettiğiniz, özlediğiniz ve sizin hakkınızda birçok şeyi bilen kişi; "romantik arkadaş" kadın/erkek ilişkisine dayalı duygusal bağ kurulan kişi; "iş arkadaşı" aynı iş ortamında bulunan, çalışma zorunluluğunuzun olduğu kişi ve "sosyal medya arkadaşı" ise yüz yüze iletişim kurulmayan sadece bir kanal aracılığıyla iletişim halinde olduğumuz kişi olarak tanımlanmıştır. Arkadaş tercihlerinde rol oynayan özellikler belirlenmiş ve gerek insanların birlikte bulundukları ortamdan kaynaklı gerekse iki kişinin birbirine yakınlık düzeyine bağlı olarak sınırlandırılan arkadaş türleri ve arkadaş özellikleri incelenmiştir.

Yöntem

Araştırmanın örneklemini Türkiye'de yaşayan arkadaş seçiminin önemli olduğunu düşünen 336 gönüllü kişi oluşturmaktadır. Veriler, araştırmacı tarafından geliştirilen anket ile toplanmıştır. Anketin ilk bölümünde katılımcıların demografik özeliklerini ve profillerini belirlenmek amacıyla öncelikle cinsiyet, yaş, medeni durum ve eğitim durumu ardından da arkadaş seçimini önemli bulup bulmadığı, yakın arkadaş olarak kabul ettiği kaç arkadaşı olduğu, haftada kaç gün arkadaşlarıyla görüştüğü ve son bir yılda kaç yeni arkadaş edindiğine ilişkin sorular yer almaktadır. İkinci bölümde ise katılımcılardan beş arkadaşlık türünü, kişisel yargıları doğrultusunda 10 karşılaştırma çifti itibarı ile benzerliklerine göre 1 = Hiç Benzer Değil, 7 = Çok Benzer arasında olan bir ölçek üzerinden değerlendirmeleri istenmiştir. Arkadaş türleri alan yazın incelemesi doğrultusunda arkadaş, yakın arkadaş, romantik arkadaş, iş arkadaşı ve sosyal medya arkadaşı olarak belirlenmiştir. Son olarak da katılımcılardan arkadaş özelliklerini önce genel sonra her bir arkadaşlık türü için ayrı ayrı önemli olup olmadığı sorgulanmıştır. Arkadaş özellikleri araştırmacı tarafından yapılan ön görüşmeler neticesinde "bencillik, cimrilik, dış güzellik, eğitim seviyesi, eğlenceli olmak, ekonomik durum, entelektüellik, güler yüzlülük, güvenirlik, kibarlık, pozitiflik, sorumluluk sahibi olma, uyum sağlayabilmek, yalan söylememek ve yardımsever olmak" olarak belirlenmiştir. Katılımcılar her bir arkadaş türünü, her bir özellik açısından 1 = Çok Önemsiz ile 7 = Çok Önemli arasında 7'li likert ölçek ile puanlamışlardır.

Verilerin analizinde çok boyutlu ölçekleme (ÇBÖ) yöntemlerinden biri olan PROXSCAL Yakınlık Analizi (Proximity Analysis) kullanılmıştır. ÇBÖ, değişkenler arasında gözlenen benzerlik ve farklılıklara dayalı değişkenlerin çok boyutlu uzayda gösterimini elde edip veri yapısını anlamlandırmaya dayalı çok değişkenli bir yöntemdir (Manly, 1994; Davidson ve Sireci, 2000). Benzerlik ve farklılıkları değişkenler arasında hesaplanan uzaklık değerlerine göre belirlenmektedir. PROXSCAL Yakınlık Analizi ise nesneler/olgular arasında var olan bir gizli yapıyı ortaya çıkarmayı amaçlar. Benzerlik ve farklılık ölçüleri için 'yakınlık' terimi kullanılır (Kruskal ve Wish, 1978). Yakınlık, iki nesnenin birbirine benzerlik/farklılığını gösterir. Sonuçlar SPSS 24.0 paket programında yer alan PROXSCAL uygulaması kullanılarak elde edilmiştir.

Bulgular

Bu bölümde öncelikle benzerlik verisi (1=hiç benzemiyor, 7=çok benziyor), farklılık verisine dönüştürülmüş (1=çok benziyor, 7=hiç benzemiyor) ve analize dâhil edilmiştir. Kruskal ve Wish (1978), veri benzerlik verisi ise orijinal veri değerlerini toplanan tüm puanlardan daha yüksek bir sabitten çıkarılarak dönüştürülmesini önermektedir (Giguère, 2006). Farklılık matrislerinde yüksek sayılar daha büyük benzemezliği/farklılığı ifade etmektedir ve psikolojik uzayda algılamayı kolaylaştırır. Tablo 3'te arkadaş türlerine göre ortalama farklılık matrisi verilmiştir.

Tablo 3

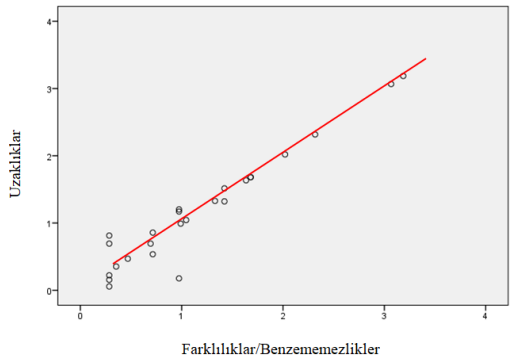
Arkadaş Çeşitlerine Göre Ortalama Farklılıklar Matrisi

	Arkadaş	Yakın Arkadaş	Romantik Arkadaş	İş Arkadaşı	Sosyal Medya Arkadaşı
Arkadaş	0.00				
Yakın Arkadaş	4.04	0.00			
Romantik Arkadaş	4.69	3.65	0.00		
İş Arkadaşı	4.44	5.02	5.76	0.00	
Sosyal Medya Arkadaşı	5.24	5.69	5.81	4.93	0.00

Tablo 3 incelendiğinde en çok benzerlik gösteren arkadaş türü yakın arkadaş ile romantik arkadaş iken en az benzerlik gösteren arkadaş türünün romantik arkadaş ile sosyal medya arkadaşlığı olduğu görülmektedir. Arkadaş türlerine ilişkin farklılık matrisine uygulanan yakınlık analizi sonucunda elde edilen iki boyutlu konfigürasyon haritası için Kruskal'ın Stres-1 değeri 0.154 olarak hesaplanmıştır. Bu değer, Kruskal'ın uygunluk ölçütlerine göre iyi uyum göstermektedir. Tucker'ın Uygunluk Katsayısı 0.98 olarak belirlenmiştir, mükemmel uyuma işaret etmektedir. Şekil 1'de farklılık matrisine ilişkin Sheapard diyagramı verilmiştir.

Şekil 1

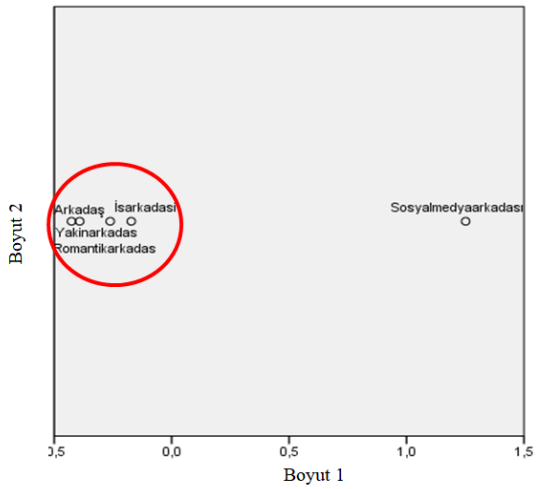
Sheapard Diyagramı



Şekil 1 incelendiğinde noktaların doğrusal bir şekilde yerleştiği, model-veri uyumunun iyi olduğu söylenebilir. Şekil 2'de farklılık matrisinin iki boyutlu konfigürasyon haritası verilmiştir.

Şekil 2

Farklılık Matrisinin Konfigürasyon Haritası



Şekil 2 incelendiğinde sosyal medya arkadaş türünün diğer arkadaş türlerinden farklılaştığı görülmektedir. Konfigürasyon haritalarında benzer olan nesneler birbirine daha yakın olarak konumlanırken farklı olan nesneler uzaklaşmaktadır. Tablo 4'te arkadaş tercihlerinde 15 farklı özelliğe ilişkin ortalama puanları yer almaktadır.

Tablo 4

Arkadaş Tercihlerine İlişkin Özelliklerin Ortalama Puanları

No	Özellik	Genel	Arkadaş	Yakın arkadaş	Romantik arkadaş	İş arkadaşı	Sosyal medya arkadaşı
1	Bencillik	6.0	6.0	6.2	6.3	5.7	4.0
2	Cimrilik	5.5	5.5	5.7	6.1	4.7	3.1
3	Dış güzellik	2.2	2.6	2.6	4.6	2.1	2.4
4	Eğitim seviyesi	4.3	4.4	4.4	5.6	5.0	4.2
5	Eğlenceli olmak	5.3	5.5	5.7	6.0	4.9	5.1
6	Ekonomik durum	2.4	2.8	2.9	4.3	2.8	2.3
7	Entelektüellik	4.3	4.5	4.7	5.2	4.4	4.5
8	Güler yüzlülük	6.2	6.1	6.3	6.4	6.0	4.6
9	Güvenirlik	6.9	6.7	6.8	6.8	6.5	5,6
10	Kibarlık	6.1	6.1	6.1	6.5	6.1	5.4
11	Pozitiflik	5.9	6.0	6.1	6.4	6.1	5.4
12	Sorumluluk sahibi olma	6.1	6.2	6.4	6.7	6.6	4.8
13	Uyum sağlayabilmek	6.1	6.2	6.4	6.6	6.4	5.2
14	Yalan söylememek	6.6	6.6	6.8	6.8	6.6	5.8
15	Yardımsever olmak	6.1	6.1	6.3	6.5	6.1	5.0

Tablo 4 incelendiğinde tüm arkadaş türleri açısından en önemli tercih sebepleri güvenirlik ve yalan söylememek olarak belirlenmiştir. Güvenirliği, sorumluluk sahibi olma, uyum sağlayabilme ve kibarlık

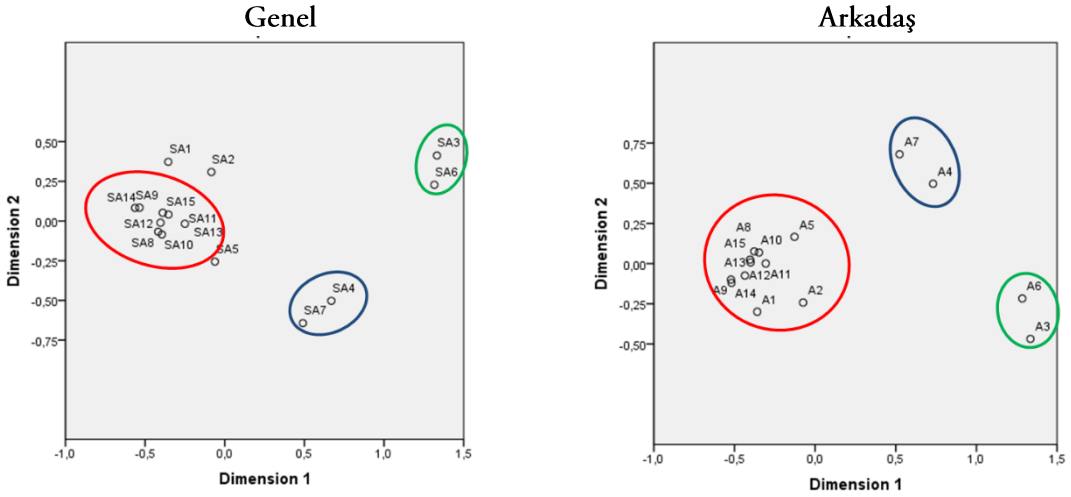
takip etmektedir. Yakın arkadaş ve romantik arkadaş kategorilerinde bu özellikler maksimum ortalamaya sahip olduğu görülmektedir. Dış güzellik ve ekonomik durum tüm arkadaş türü kategorilerinde alt sırada yer almaktadır. Arkadaş özellikleri bakımından romantik arkadaş türü en fazla ortalamaları sağlarken sosyal medya arkadaşı katılımcıların arkadaş özelliklerinde en az etkileyen özellikler olduğunu göstermektedir.

Arkadaş özelliklerinin iki boyutlu yapıda model veri uyumları incelendiğinde genel değerlendirmenin Stres-1 değeri 0,041 olarak hesaplandığı; arkadaş, yakın arkadaş, romantik arkadaş, iş arkadaşı ve sosyal medya arkadaşı için Stres 1 değerinin 0,028 ile 0,057 arasında değiştiği belirlenmiştir. Bu değerlerin Kruskal'ın tolerans değerleri tablosuna göre çok iyi uyumu ifade ettiği söylenebilir.

Arkadaş tercihlerinde rol oynayan arkadaş özelliklerinin genel değerlendirmesi ve araştırma kapsamında incelenen arkadaş türlerinde (arkadaş, yakın arkadaş, romantik arkadaş, iş arkadaşı ve sosyal medya arkadaşı) modellen iki boyutlu konfigürasyon haritaları Şekil 3'te verilmiştir. Konfigürasyon haritalarında aynı sayıyla kodlanan her bir nokta aynı arkadaş özelliğini göstermektedir.

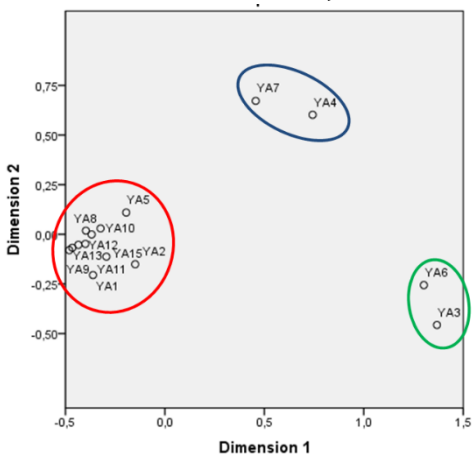
Şekil 3

Arkadaş Türlerine Göre Konfigürasyon Haritaları



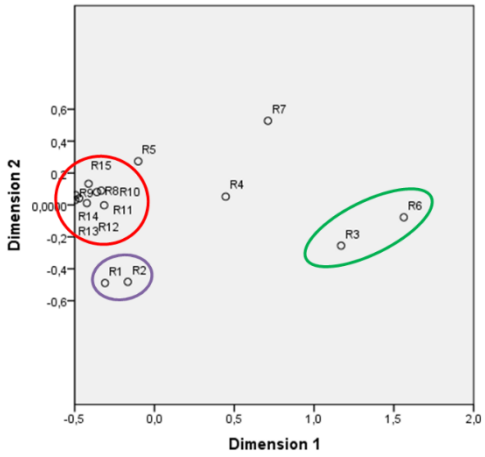
Stres-1 = 0,041 Stres-1 = 0,030

Yakın Arkadaş



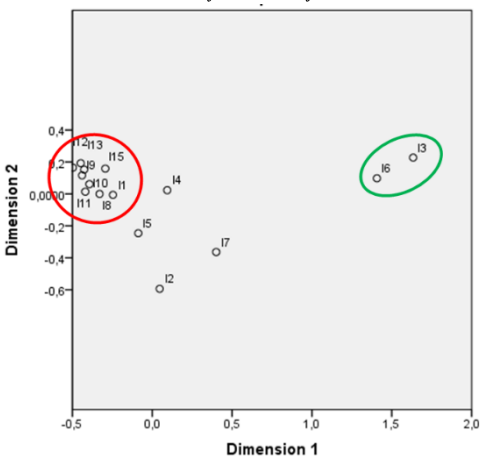
Stres-1 = 0.028

Romantik Arkadaş



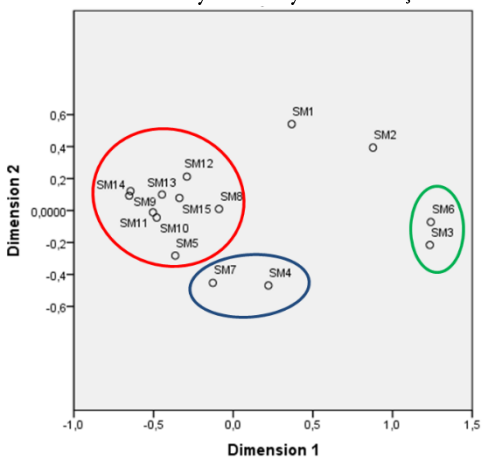
Stres-1= 0.035

İş Arkadaşı



Stres-1 = 0.057

Sosyal Medya Arkadaşı



Stres-1 = 0.045

Şekil 3'te arkadaş özelliklerinin önce genel, sonra her bir arkadaş türüne göre ayrı ayrı konumlandırılan arkadaş özelliklerine ilişkin konfügürasyon haritaları yer almaktadır.

"Dış Güzellik" ve "Ekonomik Durum" kategorilerinin her arkadaşlık türünde farklılaştığı ve birbirine yakın konum aldığı görülmektedir. "Eğitim seviyesi" ve" Entellektüellik" özellikleri de benzer şekilde farklılaşmakta ve birbirine yaklaşmaktadır.

Tüm arkadaş türleri açısından bakıldığında "güvenirlik", "pozitiflik", "sorumluluk sahibi olma", uyum sağlayabilme" ve "yalan söylememe" en çok tercih edilen arkadaş özellikleri arasında yer almaktadır.

Sonuçlar

Araştırmada verilerin model-veri uyumunun sağlandığı görülmektedir. Elde edilen verilerin yorumlanmasında konfigürasyon haritalarının kullanılmasının yorumlamayı kolaylaştırdığı söylenebilir. Veri türü açısından da her hangi bir sınırlılığının bulunmadığından çok boyutlu ölçekleme analizi yapılacak araştırmalarda kullanılmasını araştırmacılara önerebiliriz.

Arkadaş tercihlerinin belirlenmesinde en aranılan özellikler "güvenirlik" ve "yalan söylememe" olduğu belirlenmiştir. Bu iki özellik birbiri ile benzerlik göstermekte kişilerin arkadaş tercihlerinde en çok aradığı özelliğin güven olduğunu ortaya koymaktadır.

"Dış Güzellik" ve "Ekonomik Durum" arkadaş tercihlerinde önemli bir tercih sebebi olarak görülmemiş, hatta kişiler tarafından önemsiz olarak kategorize edilmiştir. Ancak romantik arkadaş türünde "dış güzellik" tercih edilme nedenlerinde artış gösterirken sosyal medya arkadaşlığında tercih nedenlerinde gerileme gösterdiği görüşmüştür. Bu durum kişilerin arkadaş tercihlerinde bulundukları yakınlık derecesi ile ilgili olduğu şeklinde yorumlanabilir.

İş arkadaşı kategorisinde "cimrilik" ve "entellektüellik" diğer arkadaş türü kategorilerinin aksine bir tercih sebebi değildir. Ayrıca iş arkadaşlığı mecburiyete dayalı bir arkadaşlık olmadığından diğer arkadaşlık türlerinden de farklılaşmıştır.

"Arkadaş" ve "Yakın arkadaş" kategorilerinde tercih edilen arkadaş özellikleri benzerlik göstermekte olup farklılaşma tercih edilme yoğunluğunda görülmektedir.

Romantik arkadaş kategorisinde tercih edilen arkadaş özelliklerinin tercih edilme yoğunlukları artarken sosyal medya arkadaşları kategorisinde arasında daha fazla yakınlık bulunurken söz konusu özelliklerin tercih edilme yoğunluklarında azalma gözlenmektedir. Bu durumu bu özeliklerin sosyal medya arkadaşlığında tercih edilme koşulunun güçlü olmadığı şeklinde yorumlayabiliriz.

Kaynaklar

Davidson, M. L., & Sireci, S. G. (2000) Multidimensional Scaling. In Tinsley, H. E., and Brown, S. D. (Eds.). *Handbook of applied multivariate statistics and mathematical modeling* (pp. 323-352). Academic press.

Giguère, G. (2006). Collecting and analyzing data in multidimensional scaling experiments: A guide for psychologists using SPSS. *Tutorials in Quantitative Methods for Psychology*, *2*(1), 26-37. https://doi.org/10.20982/tqmp.02.1.p026

Güngen, Y., Tokyürek, Ş., & Şanlı, N. (2002). *Ev ve ailede yaşam yönetimi*. Pegem Akademi Yayıncılık.

Gündoğdu, R. (2003). *İlköğretim 3., 4. ve 5. sınıf çocuklarının arkadaşlık konusundaki görüşleri ve arkadaşlık seçimlerini etkileyen etmenler* (Tez No: 125656) [Yüksek lisans tezi, Çukurova Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Kandur, H. (2018). *Tercih veri modellerinde çok boyutlu ölçekleme* (Tez No: 532476) [Doktora tezi, Marmara Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Kılavuz, F. (2015). *Ortaokul 6. 7. ve 8. Sınıf öğrencilerinin arkadaşlık hakkındaki tutumları ve arkadaşlık seçimlerini etkileyen etmenler* (Tez No: 410475) [Yüksek lisans tezi, Niğde Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling sage university papers series.* Sage Publications. https://dx.doi.org/10.4135/9781412985130

Manly, F., and Bryan, J. (1994). *Multivariate Statistical Methods.* Chapman&Hall.

Tunaboylu, C. (2019). *Çok boyutlu ölçekleme yöntemlerinin farklı benzetim koşullarında karşılaştırılması* (Tez No: 556267) [Yüksek lisans tezi, Ankara Üniversitesi]. Yükseköğretim Kurulu Tez Merkezi.

Karma veriler ne zaman sürekli kabul edilebilir?

İbrahim Uysal, Abdullah Faruk Kılıç ve Nuri Doğan

Anahtar kelimeler: Karma veri, açımlayıcı faktör analizi, polikorik korelasyon matrisi, pearson korelasyon matrisi

Giriş

Açımlayıcı faktör analizi gerçekleştirilerek ölçeklerin faktör yapılarına ilişkin bir çıkarımda bulunulmaktadır. Ancak açımlayıcı faktör analizi teorisi ve yöntemleri sürekli veriler üzerine kurgulanmıştır (Jöreskog ve Moustaki, 2001). Eğitim bilimleri ya da sosyal bilimler alanında ölçekler genellikle Likert tipinde (örn. kesinlikle katılmıyorum, katılmıyorum, kararsızım, katılıyorum, kesinlikle katılıyorum) hazırlanmakta ve sıklıkla 2-7 aralığında bir kategori sayısı seçilerek geliştirilmektedir. Dolayısıyla elde edilen veriler sıralıdır. Faktör analizinde sıralı veriler ile çalışılırken sıralı verilere özgü faktör analizi yöntemlerinin kullanılması gerektiği literatürde belirtilmektedir (Robitzsch, 2020). Bunun sebebi sürekli tanımlanan sıralı verilerle gerçekleştirilen açımlayıcı faktör analizi sonuçlarında faktör yüklerinin olduğundan düşük kestirilebilmesidir. Yani polikorik korelasyon matrisi Pearson korelasyon matrisine göre daha tutarlı sonuçlar göstermektedir (Holgado-Tello ve diğ., 2010). Bu durumun altında yatan neden ise verinin sürekli tanımlandığı Pearson korelasyon matrisi ile gerçekleştirilen açımlayıcı faktör analizinde eşit aralık ölçeğinde ölçüm yapılmasına ve doğrusal bir ilişkinin var olmasına (faktör yükleri ile gözlenen değişkenlerin kestirilen değerleri arasında) dair varsayımdır (Baglin, 2014). Sıralı verilerde bu varsayımlar ihlal edilmektedir (Timmerman ve Lorenzo-Seva, 2011). Bu nedenle açımlayıcı faktör analizi gerçekleştirilirken 5 ve daha az sayıda kategori içeren ölçeklerde kategorik olarak tanımlanan (tetrakorik ve polikorik) korelasyon matrisinden yararlanılması önerilmektedir (Rhemtulla ve diğ., 2012). Rhemtulla ve diğ. (2012) 6-7 kategori bulunduğunda ise sürekli ve kategorik veri matrislerinin benzer sonuçlar gösterdiğini belirtmiştir. Her ne kadar literatürde bu konuda bilgi olsa da araştırmacılar artan değerlerle oluşturulan (örn. 1-hiç, 2-bazen, 3-her zaman) sıralı verilerde çalışırken değişkenlerin kategorik doğasını reddetme eğilimine girerek sürekli yöntemlere yönelebilmektedir. Ancak bilinmelidir ki sıralı verilerde Pearson korelasyon matrisi ile kestirim yapılması gerçekte var olmayan çok boyutluluğa ve yanlı faktör yüklerine yol açabilecektir. Dahası verilerin güvenirliği azaldıkça temel bileşenler analizinde sonuçların yanıltıcılığı artmaktadır (Bernstein ve Teng, 1989). Yine de açımlayıcı faktör analizi

gerçekleştirilirken çoğunlukla Pearson korelasyon matrisi kullanılmaktadır (Holgado-Tello ve diğ., 2010).

Ölçeklerde genellikle derecelendirme sayısı aynıdır. Ancak bazı durumlarda ölçekler farklı derecelendirmeler içerebilmektedir. Yani aynı ölçek içerisinde hem 3 hem de 4 kategorili maddeler yer alabilmektedir. Ancak veri içerisinde kategori sayıları değişkenlik gösterdiğinde (buradan sonra karma veri olarak anılacaktır) verinin ne zaman sürekli kabul edilerek açımlayıcı faktör analizi gerçekleştirileceğine ilişkin incelenen literatürde bir simülasyon çalışmasına rastlanmamıştır. Simülasyon çalışmalarında gerçek faktör yükleri ve kestirilen faktör yükleri bilinmekte, bu durum ise faktör yüklerinin ne kadar doğru kestirildiğine ilişkin çıkarım yapılmasını kolaylaştırmaktadır. Ancak insanlardan elde edilen verilerde gerçek faktör yükünü bilmek mümkün değildir. Dolayısıyla araştırmanın simülasyon temelinde gerçekleştirilmesi daha uygun görünmektedir.

Araştırmanın amacı kategori sayıları değişkenlik gösteren veri kümeleri oluşturarak örneklem büyüklüğü, ölçme modeli, ortalama faktör yükü koşulları altında polikorik korelasyon matrisi ve Pearson korelasyon matrisi ile açımlayıcı faktör analizi gerçekleştirerek elde edilen faktör yüklerini göreli yanlılık açısından karşılaştırmaktır. Bu sayede verilerin hangi durumlarda sürekli kabul edilerek açımlayıcı faktör analizi gerçekleştirilebileceğine ilişkin çıkarım sağlanacaktır. Göreli yanlılık değeri düşük korelasyon matrisinin seçimi gerçek verilerle yürütülen araştırmalarda faktör yüklerinin doğru kestirilebilmesi için oldukça önem taşımaktadır. Dahası gerçekçi olmayan boyutların ortaya çıkmasını engellemek açısından doğru korelasyon matrisinin seçilmesi önemlidir.

Yöntem

Açımlayıcı faktör analizinde karma veriler için hangi durumlarda verilerin sürekli kabul edileceğinin belirlenmesinin hedeflendiği bu araştırma bir Monte Carlo simülasyonudur. Araştırmada simülasyon koşulları; örneklem büyüklüğü (100, 200, 500 ve 1000), madde sayısı (16 ve 24 madde), ölçme modeli (tek boyutlu, iki boyutlu [boyutlar arası korelasyon = .30]), ortalama faktör yükü (.40 ve .70) ve kategori sayısı (3 kategori – 4 kategori, 3 kategori – 5 kategori, 4 kategori – 5 kategori, 5 kategori – 7 kategori, 7 kategori – 8 kategori) olarak belirlenmiştir. Simülasyon tasarımı iki farklı kategoriye sahip maddeler üzerine gerçekleştirilmiştir. Toplamda örneklem büyüklüğü 4 koşul, madde sayısı 2 koşul, ortalama faktör yükü 2 koşul, kategori sayıları 5 koşul olmak üzere (4x2x2x2x5 = 160) 160 simülasyon koşulunda çalışılmış olup her bir koşul için 1000 replikasyon yapılmıştır.

Veri üretimi için R yazılımında bulunan (R Core Team, 2020) *lavaan* (Rosseel, 2012) paketi kullanılmıştır. Veriler kategorik hale getirilirken belirlenen kesme noktalarından (threshold) yararlanılmıştır. Karma veriler oluşturulurken maddeler %50 oranında bölünmüştür. Örneğin 16 maddelik koşul için maddelerin 8'i 3 kategorili, kalan 8'i ise 4 kategorili olacak şekilde kategorik hale getirilmiştir. Ayrıca iki boyutlu yapılar için birinci boyutta madde sayısının yarısı kadar madde yer almıştır. Diğer bir deyişle iki boyutlu yapılarda faktör başına madde sayısı 8 ve 12 olarak belirlenmiştir (simülasyon koşullarındaki madde sayılarının yarısı). Açımlayıcı faktör analizi *psych* (Revelle, 2020)

paketiyle gerçekleştirilmiştir. Açımlayıcı faktör analizinde faktör çıkarma yöntemi olarak temel eksenler (principal axis) yöntemi kullanılmıştır.

Pearson ve polikorik korelasyon matrislerinden elde edilen faktör yüklerini değerlendirmek için göreli yanlılık değeri kullanılmıştır. Göreli yanlılık replikasyonlar sonucunda elde edilen ortalama faktör yükünün tanımlanan ortalama faktör yükünden farkının tanımlanan faktör yüküne bölünmesiyle elde edilmektedir. Göreli yanlılık değerinin -.10 ile .10 aralığında olması kabul edilebilir düzeyde yanlı kestirimlerin olduğu anlamına gelmektedir (Flora ve Curran, 2004; Moshagen ve Musch, 2014). Bu nedenle bu çalışmada |GY|<.10 kriteri kullanılmıştır.

Sonuçlar

Tek boyutlu yapılarda kategori sayısı değişkenlik gösteren testler için Pearson korelasyon matrisinin kabul edilebilir aralıkta yanlı olduğu gözlenmiştir. Ancak polikorik korelasyon matrisiyle gerçekleştirilen açımlayıcı faktör analizi sonucunda özellikle küçük örneklem büyüklüğünde yakınsamanın sağlanmadığı koşullar olmuştur. Ancak polikorik korelasyon matrisiyle yakınsama sağlanabildiği durumlarda Pearson korelasyon matrisiyle elde edinen sonuçlardan daha az yanlı sonuçlar elde edilmiştir. Bu bulgulara göre tek boyutlu yapılardaki karma verilerde Pearson korelasyon matrisi kullanılarak açımlayıcı faktör analizinin gerçekleştirilebileceği söylenebilir.

İki boyutlu karma verilerde Pearson korelasyon matrisinin bazı koşullarda kabul edilenin üzerinde yanlı olduğu belirlenmiştir. Örneklem büyüklüğü küçük iken bazı koşullarda polikorik korelasyon matrisinin de kabul edilebilir değerin üzerinde yanlı olduğu sonucuna ulaşılmıştır. İki boyutlu yapılarda ortalama faktör yükünün yüksek olduğu koşullarda göreli yanlılık değerleri kabul edilebilir aralığın uç değerlerine daha yakındır.

Genel olarak karma verilerde kategori sayıları arttıkça kestirim Pearson korelasyon matrisi ile de yapılsa polikorik korelasyon matrisi ile de yapılsa kabul edilebilir derecede yanlıdır. Korelasyon matrislerinin göreli yanlılık değerleri oldukça benzer olmakla birlikte koşullara göre Pearson ya da polikorik korelasyon matrisinin daha düşük yanlılık gösterebileceği belirlenmiştir. Araştırmacılara kategori sayılarının az olduğu karma verilerde özellikle küçük örneklemlerde çalışıyorlarsa polikorik korelasyon matrisi ile kestirim yapmaları önerilebilir.

Kaynaklar

Baglin, J. (2014). Improving your exploratory factor analysis for ordinal data: A demonstration using FACTOR. *Practical Assessment, Research, and Evaluation, 19*(5), 1-15. https://doi.org/10.7275/dsep-4220

Bernstein, I. H., and Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 105*(3), 467-477. https://doi.org/10.1037/0033-2909.105.3.467

Flora, D. B., and Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466–491. https://doi.org/10.1037/1082-989X.9.4.466

Holgado–Tello, F. P., Chacón–Moscoso, S., Barbero–García, I., and Vila-Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity, 44*, 153-166. https://doi.org/10.1007/s11135-008-9190-y

Jöreskog, K. G., and Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research, 36*(3), 347-387. https://doi.org/10.1207/S15327906347-387

Moshagen, M., and Musch, J. (2014). Sample size requirements of the robust weighted least squares estimator. *Methodology, 10*(2), 60–70. https://doi.org/10.1027/1614-2241/a000068

R Core Team. (2020). *R: A language and environment for statistical computing* (version 4.1.0) [Computer software]. https://www.r-project.org/

Revelle, W. (2020). *psych: Procedures for psychological, psychometric, and personality research* (version 2.0.12) [Computer software]. https://cran.r-project.org/package=psych

Rhemtulla, M., Brosseau-Liard, P. E., and Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354-373. https://doi.org/10.1037/a0029315

Robitzsch, A. (2020). Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. *Frontiers in Education, 5*, 589965. https://doi.org/10.3389/feduc.2020.589965

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Timmerman, M. E., and Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological methods, 16*(2), 209–220. https://doi.org/10.1037/a0023353

Madde tepki kuramı model uyumsuzluğunun test eşitleme bağlamında pratik sonuçları

Sibel Aydoğan, Tuba Gündüz ve Sebahat Gören

Anahtar kelimeler: Madde Tepki Kuramı, test eşitleme, model uyumsuzluğu, eşitleme hatası, ölçek dönüştürme yöntemleri.

Giriş

Madde Tepki Kuramı (MTK) lojistik regresyon temelli kestirimler yapan ve Klasik Test Kuramının (KTK) tek modeline karşı birçok farklı model alternatifi sunan bir kuramdır. Bu da araştırmacılara mevcut veriye en uygun modeli seçerek MTK'nın önemli avantajlarından (parametre değişmezliği vb.) faydalanma imkânı sağlayabilir. MTK'nın önemli avantajlarından faydalanmak için kuramın gerektirdiği varsayımların karşılanması gerekir. Bunlar kısaca uygun boyutluluk, yerel bağımsızlık ve model veri uyumunun sağlanması olarak sıralanabilir. Bu varsayımlar birbiriyle yakından ilişkilidir ve birinin sağlanması diğeri için de temel oluşturur denebilir.

Model-veri uyumu kapsamında araştırmacıların uygulayacağı ve bu model veriye kesinlikle uyuyor ya da uymuyor denilebilecek belirli bir prosedürün varlığından bahsetmek güçtür. Neredeyse bütün istatistiksel modellerde olduğu gibi mükemmel uyum sağlayan model çok nadir bulunur ve "Uyum her zaman bir derece meselesidir (Tendeiro ve Meijer, 2015)". Bununla birlikte Madde Tepki Kuramı modellerinde model uyumsuzluğunu belirlemek için birçok metod önerilmiştir (Swaminathan ve diğ., 2007) ve uyumun belirlenmesinde izlenen yollar kısaca gözlenen değerler ve MTK modeli altında beklenen değerler arasındaki farkla açıklanmaktadır.

İç içe geçmiş nested modellerde daha az kısıtlanmış model daha fazla kısıtlanana göre daha iyi uyum gösterme eğiliminde olacaktır. Örneğin 1, 2 ve 3 parametreli lojistik modeller (PLM) -2LL temelinde değerlendirildiğinde daha az kısıtlı olan 3PLM diğerlerine göre, 2PLM ise 1PLM'ye göre daha iyi uyum sağlama eğiliminde olacaktır. Bununla birlikte operasyonel olarak iyi uyum göstermiş bir model belirli uygulamalarda pratik olarak uygulanmayabilir. Daha az uyumlu model bazı pratik sonuçları sebebiyle tercih edilebilir. Bu sebepler arasında, model basitliği, yazılıma ulaşabilirlik, fiyat/zaman etkisi, daha karmaşık modelin daha fazla katılımcı gerektirmesi, eklenen parametrelerin replikasyonlar altında daha az kararlı olabilmesi (Molenaar, 1997) yer alabilir (Zhao ve Hambleton, 2017). Dahası bir uygulamanın başında belirli bir modele karar verilince bu çalışma kapsamında yapılacak olan diğer işlemlerde de örneğin madde

kalibrasyonu, test eşitleme ve puanlama için ardışık yıllarda aynı model kullanılabilir (Zhao ve Hambleton, 2017). Uyumsuzluğun derecesi de uygulamadan uygulamaya gizil değişkenin dağılımındaki değişim ve madde parametrelerindeki kayma (drift) nedeniyle değişebilir (Park ve diğ., 2016). Bu sebeplerle model uyumsuzluğunun pratik sonuçları değerlendirilmesi gereken konulardır.

MTK model uyumsuzluğunun pratik sonuçlarının önemi araştırılması gereken bir konudur. Bu konu ile ilgili alanyazında çalışmalar yapılmasına rağmen, konuya hak ettiği önem verilmemiştir (Hambleton ve Han, 2005; Zhao ve Hambleton, 2017).

Sinharay ve Haberman (2014) model uyumsuzluğunun pratik sonuçlarını çeşitli deneysel veri setleri üzerinde araştırmış ve uyumsuzluğun belirli maddelerde gözlenmesi koşuluyla her zaman pratik olarak kayda değer sonuçlar sunmadığını belirlemiştir. Meijer ve Tendeiro (2015) ise uyumsuz kişi ve maddelerin testten çıkarılmasının testi alan kişilerin yetenek düzeyleri sırasına etkisinin seçilen MTK modeline göre farklılık gösterdiği sonucuna ulaşmıştır. Zhao ve Hambleton (2017) yaptıkları araştırmada model uyumsuzluğunun sonuçlarının seçilen ölçekleme yöntemine göre farklılık gösterdiği sonucuna ulaşırken, bu bağlamda FCIP yöntemi yetenek farkından en az etkilenen yöntem olarak bulgu verdiği, SL'nin model uyumsuzluğuna karşı daha dayanıklı olduğu sonucuna ulaşılmıştır.

Ayrıca MTK temelli yapılan eşitleme uygulamaları geleneksel yöntemlere göre birtakım avantajlar sunmaktadır. Bunlar arasında; uç puanların eşitlenmesi için daha kullanışlı çözümler sunması, eşitlenecek test formlarının seçilmesinde esnekliklerinin olması, madde düzeyinde ön eşitleme imkânı tanıması ve bireyin farklı güçlük düzeyine sahip test formlarından herhangi birini almasının yetenek kestirimini değiştirmemesi yer alabilir (Cook ve Eignor, 1991; Hambleton ve Swaminathan, 1985). MTK ve KTK ile eşitleme yöntemlerini karşılaştıran çalışmalar incelendiğinde, genellikle MTK eşitleme yöntemlerinin daha kararlı sonuçlar verdiği görülmüştür (Han ve diğ., 1997; Yang ve Houang, 1996). Bu nedenle MTK test eşitlemede yaygın olarak kullanılmaktadır (Kim ve Lee, 2006).

Özet olarak bu çalışmada model uyumsuzluğunun ölçek dönüştürme yöntemleri üzerindeki etkisi farklı örneklem büyüklüklerinde incelenecektir. Alanyazında model uyumsuzluğunun pratik sonuçlarına odaklanan çalışmalar olmakla birlikte, farklı örneklem büyüklüklerinde model uyumsuzluğunun farklı ölçek dönüştürme yöntemleri üzerindekini etkisini inceleyen çalışmaya rastlanılmamıştır. Çalışma bu yönüyle değerli görülmektedir.

Yöntem

Bu araştırmada bazı koşullar kontrol altına alındığında ortaya çıkan sonuçları incelemek amaçlandığında çalışma, simülatif veriler üzerinden yürütülecektir. Araştırmada eşitleme deseni olarak "denk olmayan gruplarda ortak test deseni" kullanılacaktır.

Araştırmada Form-X ve Form-Y olmak üzere iki kategorili çoktan seçmeli (1-0) maddelerden oluşan iki form oluşturulacaktır. Her bir formdaki madde sayısı 40; ortak madde sayısı 10 olarak belirlenecektir. Araştırma verileri WINGEN3'te, grupların yetenek dağılımları her bir grup için standart

normal dağılım ($\theta \sim N(0,1)$) kullanılarak üretilecektir. Elde edilecek cevap örüntüleri 3PLM'ye uygun olarak üretilecektir. Bunun için ortak maddeler ve Form-X için ayırıcılık parametresi BILOG-MG'deki madde ayırıcılık parametrelerinin varsayılan dağılımı olan ortalaması 0 standart sapması 0.5 olan log-normal dağılımdan, güçlük parametresi ortalaması 0 standart sapması 1 olan normal dağılımdan üretilecektir. Şans parametresi ise beş seçenekli testler için pratikte sık rastlanabilecek aralık olan 0.20 - 0.30 aralığında uniform dağılımdan üretilecektir. Test eşitleme farklı formlar arasındaki güçlük farklılıklarını gidermek amacıyla yapıldığı için Form-Y için güçlük parametre ortalaması 0.5 olarak değiştirilecektir.

Araştırmada etkisi incelenen simülasyon faktörleri; model uyumsuzluğu, örneklem büyüklüğü ve test eşitlemede kullanılan yöntemlerdir. İlk olarak 3PLM'ye göre türetilen verinin 1PLM'ye ve 2PLM'ye göre de parametre kestirimi yapılarak uyumsuzluk sonuçları incelenecektir. Araştırmaya dâhil edilen ikinci faktör örneklem büyüklüğüdür. Bu araştırmada her bir grup için 500, 1000 ve 3000 olmak üzere üç koşul olması planlanmıştır. Her bir koşul için 25 replikasyon yapılacaktır.

Verilerin analizinde öncelikle 500, 1000 ve 3000 birimlik veri setleri için madde ve yetenek parametreleri 1, 2 ve 3PLM'ye göre kestirilecektir. Parametre kestiriminde BILOG-MG programı, yetenek parametre kestiriminde ise EAP yöntemi tercih edilecektir. Madde ve yetenek parametreleri kestirildikten sonra IRTEQ programı yardımıyla eşitleme için gerekli olan A ve B katsayıları dört farklı yöntem için (ortalama-ortalama, ortalama standart sapma, Stocking-Lord, Haebara) elde edilecektir. Son aşamada ise farklı koşullar için elde edilen eşitleme hataları hesaplanacaktır.

Sonuçlar

Farklı örneklem büyüklüklerinde (500,1000, 3000) ve model-veri uyumsuzluğu durumunda ölçek dönüştürme yöntemlerinden (OO, OS, SL, HA) elde edilen eşitleme hataları hesaplanacaktır. Örneklem büyüklüğü yeterli düzeye ulaşmadığında (N=500) tüm yöntemlerde daha fazla hatalı sonuçlar üretmesi beklenmektedir. Model uyumsuzluğu karşısında en dayanıklı yöntemin SL olduğu sonucuna ulaşılması beklenmektedir. 3PLM için yeterli örneklem büyüklüğüne ulaşılması koşulunda ise (N=3000) bütün ölçek dönüştürme yöntemleri ile yapılan eşitleme sonucunda en az hataya ulaşılması beklenmektedir.

Sonuç olarak model-veri uyumsuzluğu olduğu durumda farklı örneklem sayılarında hangi ölçek dönüştürme yöntemlerinin kullanılması detaylı bir şekilde önerilecektir. Bu çalışma gösterecektir ki model uyumsuzluğu eşitleme süreci için önemli olmakla birlikte tek ve birinci şart değildir. Modelin de belirli varsayımlarının olduğu dikkate alınarak (örneğin; örneklem büyüklüğü) eşitleme uygulamaları gerçekleştirilmelidir.

Kaynaklar

Cook, L. L., and Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. *Educ. Meas. Issues Pract.* 10, 37–45. https://doi.org/10.1111/j.1745-3992.1991.tb00207.x

Hambleton, R. K., and Han, N. (2005). Assessing the fit of IRT models to educational and psychological test data: A five step plan and several graphical displays. In W. R. Lenderking, and D. Revicki (Eds.), *Advances in health outcomes research methods, measurement, statistical analysis, and clinical applications* (pp. 57–78). Degnon Associates.

Hambleton, R. K. ve Swaminathan, H. (1985). *Item response theory: Principles and applications.* Nijhoff Publishing.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 31*(5), 457–459. https://doi.org/10.1177/0146621607299271

Han, T., Kolen, M. J. ve Pohlmann, J. (1997). A comparison among IRT true-and observed score equating and traditional equipercentile equating. *Applied Measurement in Education, 10*(2), 105-121. https://doi.org/10.1207/s15324818ame1002_1

Kim, S. ve Lee, W. (2006). An extention of four IRT linking methods for mixed-format tests. *Journal of Educational Measurement, 43*(1), 53-76. https://www.jstor.org/stable/20461809

Meijer, R. R., and Tendeiro, J. N. (2015). *The Effect of Item and Person Misfit on Selection Decisions: An Empirical Study. Law School Admission Council Research Report, RR* 15-05. http://www.lsac.org/docs/defaultsource/research-(lsac-resources)/rr-15-05.pdf

Molenaar, I. W.(1997). Lenient or strict application of IRT with an eye on practical consequences. In J. Rost and R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences*, (pp. 38-49). Waxman.

Park, Y. S., Lee, Y.S., and Xing, K. (2016). Investigating the impact of item parameter drift for item response theory models with mixture distributions. *Frontiers in Psychology*, 7:255. https://doi.org/10.3389/fpsyg.2016.00255

Sinharay, S., and Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice,* 33, 23–35. https://doi.org/10.1111/emip.12024

Swaminathan, H., Hambleton, R. K., and Rogers, H. J. (2007). *Assessing the fit of item response theory models.* In C. R. Rao, and S. Sinharay (Eds.), *Handbook of Statistics* (Volume: 26, pp. 683–718). Elsevier Publishing.

Tendeiro, J. N., and Meijer, R. R. (2015). *How Serious is IRT Misfit for Practical Decision Making? Law School Admission Council Research Report,* RR15-04. http://www.lsac.org/docs/default-source/research-(lsacresources)/rr-15-04.pdf

Yang, W. L., and Houang, R. T. (1996, April, 8-12). *The effect of anchor length and eqauating method on the accuracy of test equating comparisons of linear and IRT-based equating using an anchor-item design* [Paper presentation]. American Educational Research Association, New York, United States.

Zhao, Y., and Hambleton R. K. (2017) Practical consequences of ıtem response theory model misfit in the context of test equating with mixed-format test data. *Frontiers in Psychology*, 8:484. https://doi.org/10.3389/fpsyg.2017.00484

Geçmişten 21. Yüzyıla zihinsel süreçlerin ölçülmesinde yaşanan değişim

Ömer Kutlu

İnsanlık tarihi boyunca genç nüfusun eğitilmesi, toplumsal yaşamda gerekli olan görevlerin ve rollerin kazandırılması toplumların öncelikleri arasında yer almıştır (Harari, 2015; 2016). Eğitim toplumların varlıklarını sürdürebilmeleri için vazgeçilmezdir. Eğitimin başarısı ve bu başarının izlenmesi eğitimin değerlendirilmesini de beraberinde getirmiştir. İnsanın psikolojik özellikleriyle ilgili merak, Platon, Aristo ve diğer Yunan düşünürlerine kadar uzanmaktadır. Bu düşünürler, *bellek, öğrenme, güdü, algı, rüyalar ve akıl dışı davranışlar* gibi insan doğasıyla ilgili günümüzdeki psikologların da ilgilendiği pek çok konu hakkında düşünmüşlerdir. 19. yüzyılın son çeyreğine kadar filozoflar insan doğasını birtakım kurgulara, sezgilere ve kendi sınırlı kişisel deneyimlerine dayalı genellemeler yoluyla incelemişlerdir. Daha sonra, biyoloji ve fizik alanlarında başarıları önceden kanıtlanmış bilimsel araç ve yöntemlerin insan doğasına ilişkin sorunlara uyarlanmasıyla büyük bir değişiklik meydana gelmiştir (Schultz ve Schultz, 2011).

Psikolojik ölçmelerin 1800'lü yıllardan itibaren tarihi incelenecek olursa, son 200 yıl içinde insan davranışlarının ölçülmesi ve değerlendirilmesi süreçlerinde önemli değişmeler yaşanmıştır. Bir yandan zekâ, tutum, algı, kişilik, yetenek ve başarı gibi psikolojik yapıların ölçülmesindeki ilerlemeler (Anastasi ve Urbina, 1997; Cronbach, 1990) bir yandan da Klasik Test Kuramı (KTK) ve Madde Tepki Kuramı (MTK) gibi ölçme kuramlarındaki zenginleşmeler (Crocker ve Algina, 1986; Gulliksen, 1950; Kline, 1986; Lord ve Novick, 1968) alanın değişmesinde ve gelişmesinde etkili olmuştur. 1850'li yıllara doğru Horace Mann tarafından uyarlanan yazılı sınavların öğrencileri bir düzene soktuğu, geniş bir içerik kapsamını ölçtüğü, söz konusu şans ögesini azalttığı ve sınav görevlileri tarafındaki kayırmacılık olasılığını ortadan kaldırdığı düşüncesini ön plana çıkarmıştır (Domino ve Domino, 2006). 1850'li yıllardan itibaren yazılı sınavlara dayalı sorulardan oluşan ve öğrenmelerin ayrıntılarını ölçen sınavlar yaygınlaşmıştır. Öğrencilerin büyük miktarda bilgiyi ezberlemelerini gerektiren okul anlayışı için bu sınavlar önemli görülmüştür (Raban, 2008).

1900'lü yılların ortalarına doğru güvenirlik üzerinde yapılan çalışmalar, aynı özelliği ölçen madde sayısı ve birey sayısı arttığında güvenirliği etkileyen rastlantısal hatanın azaldığı ve güvenirliğin arttığını ortaya koymuştur. Bu durum çok sayıda maddeden oluşan test uygulamalarının okul ortamlarında yaygınlaşmasına yol açmıştır. Öğrenci başarısının değerlendirilmesinde derslerin kazandırmayı

hedeflediği davranışlar farklı madde türleriyle ölçülmeye başlanmış, testlerin kapsam geçerliğini sağlamada konu boyutunu temsil eden fazla sayıda maddenin kullanılması önemli görülmüştür. Bu durum kapsam geçerliğinin diğer boyutu olan bilişsel düzeyleri de etkilemiştir. Bilişsel sınıflamanın üst basamaklarını oluşturan *uygulama, analiz* ve *değerlendirme* gibi düşünme düzeylerine verilen yanıtların zaman alması ve madde sayılarını azaltması test geliştiricileri çoğunlukla hatırlama ve anlama gibi bilişsel düzeylerde madde yazmaya zorlamıştır.

Geçerlik kavramı üzerinde yoğun çalışmaların yapıldığı 1940'lı yıllarda testin geçerliğini ortaya koyacak ölçütün testin dışından ya da testin içinden olup olamayacağı da tartışılmıştır. Sireci, (1998) geçerlik üzerine ilk çalışmaları yapan bilim insanlarının geçerlik değerlendirmelerinin yalnızca istatistiksel bakışıyla yapılmasının çok kısıtlayıcı olduğu yönündeki kaygılarını dile getirmiştir. Bununla ilgili olarak, Kelly'nin 1927 yılında testlerin geçerlik değerlendirmelerini desteklemek için uzman yargılarının da kullanılması önerisini dile getirmiştir. Rulon ise 1946 yılında, testin geçerliği hakkında bilgi verecek ölçütün testin içeriğiyle ve amacıyla ilgili olduğunu vurgulamıştır.

Öğrencilerin okullardan edindikleri temel bilgileri doğal ve toplumsal yaşam durumlarında istenen düzeyde kullanamamaları özellikle 1980'li yıllardan itibaren öğrenme ve ölçme yaklaşımlarının sorgulanmasına yol açmıştır. Bu tartışmalarda okulların yaşam için gerekli bilişsel, içsel ve kişilerarası becerileri yeterince kazandıramadığı ele alınmıştır (UNDP ve UNICEF, 1990; Marzano, 1992; Popham, 2007; Haladyna, 1997; Kutlu ve diğ., 2017; Kutlu ve Kula-Kartal, 2018).

Kaynaklar

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Prentice Hall/Pearson Education.

Crocker, L. M., & Algina, L. (1986). *Introduction to classical and modern test theory.* Holt, Rinehart and Winston.

Cronbach, L. J. (1990). *Essentials of psychological testing.* Harper & Row.

Domino, G., & Domino, M. L. (2006). *Psychological testing: An introduction* (2nd ed.). Cambridge University Press.

Gulliksen, H. (1950). *Theory of mental tests* (1st ed.). John Willey & Sons.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking.* Viacom Company.

Harari, Y. N. (2015). *Hayvanlardan tanrılara SAPIENS: İnsan türünün kısa bir tarihi.* Kolektif Kitap.

Harari, Y. N. (2016). *Homo Deus: Yarının kısa bir tarihi.* Kolektif Kitap.

Kline, P. (1986). *A handbook of test construction.* Methuen & Co. Ltd.

Kutlu, Ö., Doğan, C. D. ve Karakaya, İ. (2017). *Ölçme ve değerlendirme: Performansa ve portfolyoya dayalı durum belirleme* (5. Baskı). Pegem Akademi Yayıncılık.

Kutlu Ö., & Kula-Kartal, S. (2018). The prominent student competences of the 21st century education and the transformation of classroom assessment. *International Journal of Progressive Education 14*(6), 70-82.

Lord F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Addison-Wesley Publishing Company.

Popham, J. W. (2007). Formative assessment: False pathway to proficiency? Properly implemented formative assessments can improve student learning, even if they don't measure proficiency. *Leadership Compass, 4*(3), 1-2.

Raban, S. (2008). Introduction. In S. Raban (Ed.). *Examining the world: A history of the university of Cambridge local examinations syndicate* (pp. 1-11). Cambridge University Press.

Schultz, D. P., & Schultz, S. E. (2011). *A history of modern psychology* (10th ed.). Cengage Learning.

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research, 45*, 83-117. https://doi.org/10.1023/A:1006985528729

UNDP., & UNICEF (1990, March 5-9). *World declaration on education for all and framework for action to meet basic learning needs.* World Conference on Education for All: Meeting Basic Learning Needs. Jomtien, Thailand.

21. Yüzyılda öne çıkan bilişsel, içsel ve kişilerarası beceriler

Seval Kula Kartal

Okulların amacı, farklı özelliklere sahip bireylerin bilgili, sorumlu ve içinde yaşadığı toplumla ilgilenen yetişkinler olarak yetişmesine katkı sağlamaktır. Toplumun büyük çoğunluğunda, böyle bireyler yetiştirmenin zorluğuna ilişkin farkındalık yüksektir. Ancak, bu zorluğun, çocukların bilişsel becerilerinin yanında içsel (intrapersonel) ve kişilerarası (interpersonal) becerilerinin öğrenme sürecine katılarak, sistemli ve özenli dikkat gösterilerek aşılabileceğine ilişkin farkındalık düşüktür (Elias ve diğ., 1997). Farklı araştırmacılar, öğrencilerin okulda geçirdikleri zamanın önemli kısmının onların bilişsel özelliklerini geliştirmeye ayrıldığını ifade etmektedir (Cooper ve Cefai 2009; Gardner, 1983; Goleman, 1995; 2006). Ancak okulların bilgili, sorumlu ve ilgili yetişkinler yetiştirmesi bireye bir bütün olarak yaklaşmasını gerektirmektedir. Bu ise, okulların bireylerde yalnızca bilişsel yeterlik alanında değil, içsel ve kişilerarası yeterlik alanlarındaki becerilerini keşfetme ve geliştirme olanakları sağlamasını ön plana çıkarmaktadır.

Okulların temel aldığı bilişsel yeterlik alanındaki gelişimin istenilen düzeye ulaşabilmesi için bireylerin içsel ve kişilerarası becerilerinin geliştirilmesi büyük önem taşımaktadır. Öğrencilerin sosyal ve duygusal gereksinimlerinin karşılanmasının, öğrencileri öğrenmeye daha hazır duruma getirerek onların öğrenme kapasitelerini arttırdığı bilinmektedir [Collaborative for Academic, Social and Emotional Learning (CASEL), 2003]. Yapılan çalışmalar bireylerin bilişsel gelişimleriyle içsel ve kişilerarası gelişimleri arasındaki ilişkisel doğayı destekleyen bulgular ortaya koymuştur. Örneğin, Wang ve diğ., (1990) tarafından yapılan çalışma, içsel ve kişilerarası becerilerle ilgili değişkenlerin akademik başarıyla güçlü bir ilişkiye sahip olduğunu göstermiştir. Benzer biçimde Durlak ve diğ., (2011), öğrencilerin bu becerilerini geliştirmek amacıyla uygulanan programların, öğrencilerin akademik başarılarını arttırdığını ortaya koymuştur. Grolnick ve Ryan (1987) tarafından yapılan çalışmada, içsel güdülenme düzeyleri yüksek öğrencilerin kavramsal öğrenme açısından daha yüksek puanlar elde ettiği bulunmuştur. McConney ve Perry'nin (2010) çalışmasında düşük sosyoekonomik düzeye sahip öğrenci grubu içerisinde özyeterlik düzeyi yüksek olan öğrencilerin, düşük olan öğrencilere göre daha yüksek matematik başarısına sahip oldukları görülmüştür. PISA 2006 uygulamasında, sosyoekonomik bakımdan dezavantajlı öğrenci grubu içerisinde başarı düzeyi yüksek olan öğrencilerin içsel güdülenme düzeylerinin, başarı düzeyi düşük olan öğrencilerden daha yüksek olduğu bulunmuştur (OECD, 2011). Strayhorn (2013) çalışmasında öğrencilerin kararlılık düzeylerinin akademik başarılarındaki varyansın

%26'sını açıkladığını ortaya koymuştur. Bu çalışmayı destekler biçimde, PISA 2012 Türkiye verileri üzerinde yapılan bir araştırmada, tüm konu alanlarında kararlılık ve problem çözmeye açıklık değişkenlerinin başarıyı anlamlı biçimde yordadığı görülmüştür (Kutlu ve diğ., 2017).

Öğrencilerin farklı içsel özellikleri üzerinde çok önemli çalışmalar yapmış ve bu yapıların kuramsal temellerinin geliştirilmesinde katkılarda bulunmuş bilim insanları da bu yeterliklerin öğrencilerin okul ve yaşam başarısı üzerinde olumlu etkileri olduğunu belirtmektedir (Bandura, 1977; 1982; Deci ve Ryan 1985; 2000; Marzano, 1992; Schunk ve diğ., 2014). Hem kuramsal hem de görgül araştırmalar, öğrencilerin okul başarısını arttırmada içsel ve kişilerarası becerilerin geliştirilmesinin önemli olduğunu ortaya çıkarmış özellikle 1980'li yıllardan itibaren 2000 yılına kadar bu konuda çeşitli çalışmalar yapılmıştır. Bu aşamada, farklı araştırmacılar tarafından bazı yönleriyle ortak bazı yönleriyle farklılaşan tanımlar ve çerçeveler geliştirilmiştir. Özellikle Gardner (1983), Goleman (1995), Salovey ve Mayer (1990) tarafından geliştirilmiş çerçeveler bu alandaki öncü çalışmalardır. Ayrıca, bu çerçeveler yayınlandığı tarihten sonraki uygulamalarda sıklıkla temel alınan çerçeveler içerisinde yer almıştır. 2000 yılı sonrasında bu alandaki hem kuramsal hem de uygulamalı çalışmaları yürüten çok sayıda farklı kuruluş oluşmuştur. Bu kuruluşlar da kendi çerçevelerini tanımlamıştır. Bu çerçeveler içerisinde, CASEL ve NESET (Network of Experts working on the Social dimension of Education and Training) tarafından geliştirilen çerçeveler oldukça kapsamlıdır (CASEL, 2003; Cefai ve diğ., 2018). İçsel ve kişilerarası beceri alanında önemli katkı getiren bir kuruluş da OECD (2019) olmuştur. OECD 2018 yılında ilk kez geniş ölçekli bir sosyal ve duygusal beceri çalışması (Social and Emotional Skills Study -SESS-) gerçekleştirmiş ve bu konuda dünyanın farklı ülkelerinden içsel ve kişilerarası becerilere ilişkin elde edilen veriye dayalı bir çerçeve geliştirmiştir.

Bu çalışmada, öncü araştırmacılar ve kuruluşlar tarafından oluşturulan altı farklı içsel ve kişilerarası beceri çerçevesindeki (CASEL, 2003; Gardner, 1983; Goleman, 1995; OECD, 2019; Salovey ve Mayer, 1990) ortak beceriler incelenmiş ve 21. yüzyılda öne çıkan temel beceriler ortaya konulmuştur. Tüm bu çalışmalar incelendiğinde bilişsel, içsel ve kişilerarası becerilerin çoğunlukla birbirinden bağımsız ölçüldüğü görülmüştür.

Okullarda öğrenilen temel bilgi ve becerilerin gerçek yaşam durumlarında kullanımı bir yandan öğrencilerin anlama, problem çözme, eleştirel düşünme, yaratıcı düşünme gibi bilişsel yeterliklerini, diğer yandan da içsel ve kişilerarası yeterliklerini birlikte kullanmasını gerekli kılmaktadır. Birçok bilim insanı tarafından da üst düzey düşünme süreçlerinde gelişimin, bu üç yeterlik alanının birlikte kullanımına bağlı olduğu ileri sürülmüştür (Haladyna, 1997; Kubiszyn ve Borich, 2003; Kutlu ve Kula-Kartal, 2018; Marzano ve Heflebower, 2012). Bu nedenle okul öğrenmelerinde öğrenme çıktıları, öğrencilerin bu üç yeterlik alanına ait becerileri aynı anda kullanacakları öğrenme-öğretme etkinliklerine dayandırılmalı ve aynı zamanda ölçme ve değerlendirme uygulamalarında da maddeler bu üç beceriyi gerçekçi durumlardan yararlanarak ölçecek biçimde oluşturulmalıdır.

Bu çalışmada üç temel yeterlik alanının bütünselliği kuramsal temellere dayalı olarak ele alınacaktır.

Kaynaklar

Bandura, A. (1977). Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191-215. https://doi.org/10.1037/0033-295X.84.2.191

Bandura, A. (1982). *Self-efficacy mechanism in human agency.* American Psychologist.

CASEL (2003). *Safe and sound: An educational leader's guide to evidence-based social and emotional learning programs.* Collaborative for Academic, Social, and Emotional Learning.

Cefai, C., Bartolo, P. A., Cavioni., V., & Downes, P. (2018). *Strengthening social and emotional education as a core curricular area across the EU. A review of the international evidence, NESET II report.* Publications Office of the European Union.

Cooper, P., & Cefai, C. (2009). Introducing emotional education. *The International Journal of Emotional Education, 1*(1), 1-7. https://www.um.edu.mt/library/oar/handle/123456789/58521

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior.* Springer.

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self determination of behavior. *Psychological Inquiry, 11*(4), 227-268.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development, 82,* 405-432.

Elias, M. J., Zins, J. E., Weissberg, R. P., Frey, K. S., Haynes, N. M., Kessler, R., Shwab-Stone, M. E., & Shriver, Y. (1997). *Promoting social and emotional learning: Guideline for educators.* Association for Supervision and Curriculum Development.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences.* Basic Books.

Goleman, D. (1995). *Emotional intelligence.* Bantam Books, Inc.

Goleman, D. (2006). *Social intelligence: The new science of human relationships.* Bantam Books.

Grolnick, W., & Ryan, R. M. (1987). Autonomy in children's learning: An experimental and individual difference investigation. *Journal of Personality and Social Psychology, 52*(5), 890-898.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking.* Viacom Company.

Kubiszyn, T., & Borich, G. (2003). *Educational testing and measurement.* John Wiley & Sons, Inc.

Kutlu Ö., & Kula-Kartal, S. (2018). The prominent student competences of the 21st century education and the transformation of classroom assessment. *International Journal of Progressive Education 14*(6), 70-82. https://doi.org/10.29329/ijpe.2018.179.6

Kutlu, Ö., Kula-Kartal, S., & Şimşek, N. T. (2017). Identifying the relationships between perseverance, openness to problem solving, and academic success in PISA 2012 Turkey. *Journal of Educational Sciences Research, 7*(1), 263-274.

Marzano, R. J. (1992). *A different kind of classroom: Teaching with dimensions of learning.* Association for Supervision and Curriculum Development.

Marzano, R. J., & Heflebower, T. (2012). *Teaching and assessing 21st century skills.* Marzano Research.

McConney, A., & Perry, L. B. (2010). Socioeconomic status, self-efficacy, and mathematics achievement in Australia: A secondary analysis. *Educational Research for Policy and Practice, 9,* 77-91.

OECD (2011). *Against the odds: Disadvantaged students who succeed in school.* Organization for Economic Co-operation and Development.

OECD (2019). *Assessment framework of the OECD study on social and emotional skills.* Organization for Economic Co-operation and Development.

Salovey, P., & Mayer, J. D. (1990). Emotional intelligence. *Imagination, Cognition and Personality, 9*(3), 185-211.

Schunk, D. H., Meece, J. L., & Pintrich, P. R. (2014). *Motivation in education: Theory, research and applications*. Pearson Education, Inc.

Strayhorn, T. L. (2013). What role does grit play in the academic success of black male collegians at predominantly white institutions? *Journal of African American Studies*, *18,* 1-10.

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1990). What influences learning? A content analysis of review literature. *The Journal of Educational Research, 84*(1), 30-43.

21. Yüzyılda bilişsel, içsel ve kişilerarası becerilerin ölçülmesi ve durum belirleme

Özge Altıntaş

Anahtar kelimeler: durum belirleme, 21. yüzyıl becerileri, bilişsel beceriler, içsel beceriler, kişilerarası beceriler, öğrenci başarısı, geribildirim

Giriş

Eğitim, dünyadaki birçok ülke tarafından savaş, terör, açlık, çevre kirliliği, küresel ısınma, çocuk ve kadına yönelik şiddet, gelir adaletsizliği, uyuşturucu kullanımı, hakların ihlali gibi birçok sorunun çözümünde önemli bir araç olarak görülmektedir. Bazı bilim insanları ise, yanlış eğitim politikalarının bu sorunlara yol açabileceği üzerinde durmaktadır (Haladyna, 1997; Kutlu ve diğ., 2017; Popham, 2000). Birçok görüş okullarda yürütülen eğitimin yetersizliğini ve okul uygulamalarını eleştirmektedir (Gardner, 2019; Harari, 2018; Wagner, 2015). Okul eğitiminin yetersizliğini eleştiren çalışmalar günümüz okullarının 21. yüzyılın gereklerine ve genç kuşağın özelliklerine uymadığına vurgu yapmaktadır (Marzano ve Heflebower, 2012).

Okullar uzun yıllardır öğrencilerde çoğunlukla bilişsel davranışların gelişimine odaklanmışlardır. UNESCO (2013-4) raporu, eğitimdeki düşük kalitenin, öğrenmeyi okula gidenler için bile engellediğini, ilkokul çağındaki çocukların üçte birinin okula gitseler de temel becerileri öğrenemediklerini dile getirmektedir. Bu olumsuz koşulların değiştirilmesi isteği eğitim-öğretim süreçlerinde güncellemeler yapılmasını kaçınılmaz kılmaktadır. 2000'li yıllardan itibaren uygulanan PISA, PIRLS ve TIMSS gibi uluslararası öğrenci başarısını belirleme çalışmaları eğitimde karar vericilere önemli veriler sunmakta ve eğitimde alınması gereken önlemlere işaret etmektedir. Bunun yanında, biyoloji, tıp ve psikoloji bilimlerinin bulguları da insan davranışlarının doğasını daha ayrıntılı anlamamızı sağlayan bilgiler sunmaktadır (Andreasen, 2005; Asbury ve Plomin, 2013; Demirsoy, 2018; Kılıç, 2019).

Sözü edilen değişmeler dikkate alındığında 21. yüzyılın, doğayla ve toplumla daha barışık yaşayacak bireylerin yetiştirilmesine gereksinim duyduğu görülmektedir. Bu yüzyıl, okul öğrenmeleri için üretilmiş eski bilgilerden de yararlanarak yeni bilgiler ortaya koyabilmelidir. Öğrencilerin başarısında

etkili olduğu düşünülen bilişsel, içsel (intrapersonel) ve kişilerarası (interpersonal) yeterlikler de ölçme ve durum belirleme sürecine dahil edilebilmelidir.

Öğrencilerin kendi öğrenmelerini yönetebilmelerine olanak sağlayan; problem çözme, analitik düşünme, akıl yürütme, eleştirel düşünme, yaratıcılık gibi ***bilişsel***; özgüven, özyeterlik, güdülenme, problem çözmeye açıklık, kararlılık, esneklik, uyum gibi ***içsel***; işbirliği yapma, iletişim kurma, zamanı yönetme, sorumluluk alma, ikna etme, hesap verebilir olma, üretkenlik gibi ***kişilerarası*** özelliklerinin de ölçme sürecinin içine çekilmesi iyi bir ölçme ve durum belirleme anlayışının vazgeçilmez koşulu olarak kabul edilmektedir. Okul öğrenmelerinde, günlük yaşamda ve çalışma yaşamında ön plana çıkan bu beceriler 21. yüzyıl becerileri olarak adlandırılmaktadır (Collins, 2014; Kutlu ve Kula-Kartal, 2018; Kyllonen, 2012; Marzano ve Heflebower, 2012; Payne ve Kyllonen, 2012).

Öğrencilerin üst düzey düşünme becerilerine ilişkin gelişimlerinin bilişsel, içsel ve kişilerarası yeterliklerin birlikte kullanımına bağlı olduğu birçok bilim insanı tarafından ileri sürülmüştür. Yapılan araştırmalar bu anlayışı destekleyen yaklaşımların doğruluğunu gösteren kanıtlar ortaya koymuştur. Bunun için geribildirim vermeyi önceleyen ve bu üç yeterlik alanına ilişkin becerileri aynı anda kullanmayı gerektiren performans görevleri ile öğrencinin kendi yanıtını yapılandırmasını sağlayan açık uçlu maddelere dayalı ölçme ve durum belirleme yaklaşımları yaygınlaştırılmalıdır (Durlak ve diğ., 2011; Grolnick ve Ryan, 1987; Marzano, 1992). Bu nedenle bu çalışmada, 21. yüzyılda ön plana çıkan ölçme ve durum belirleme anlayışı, kuramsal temellere dayalı olarak madde örnekleriyle, puanlama ve geribildirim verme yaklaşımıyla ele alınmıştır.

Kaynaklar

Andreasen, N. C. (2005). *The creating brain: The neuroscience of genius.* Dana Press.

Asbury, K., and Plomin, R. (2013). *G is for genes: The impact of genetics on education and achievement.* John Wiley & Sons, Inc.

Collins, R. (2014). Skills for the 21st Century: Teaching higher-order thinking. *Curriculum & Leadership, 12*(14). http://www.curriculum.edu.au/leader/teaching_higher_order_thinking,37431.html?issueID=12910

Demirsoy, A. (2018). *Biyolojik saat: Belleğin ve davranışların evrimi.* Asi Kitap.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., and Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, *82*(1), 405-432. https://doi.org/10.1111/j.1467-8624.2010.01564.x

Gardner, H. (2019). *Eğitilmemiş zihin.* Alfa Basım Yayım Dağıtım.

Grolnick, W., and Ryan, R. M. (1987). Autonomy in children's learning: An experimental and individual difference investigation. *Journal of Personality and Social Psychology, 52*(5), 890-898. https://doi.org/10.1037/0022-3514.52.5.890

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking.* Viacom Company.

Harari, Y. N. (2018). *21. yüzyıl için 21 ders.* Kolektif Kitap.

Kılıç, T. (2019, 24 Aralık). *Yeni bilim ve kültürün kaynağı: Bağlantısal bütünsellik.* [Video]. YouTube. https://www.youtube.com/watch?v=rLu2zXWvAX8&feature=youtu.be

Kutlu, Ö., Doğan, C. D. ve Karakaya, İ. (2017). *Ölçme ve değerlendirme: Performansa ve portfolyoya dayalı durum belirleme* (5. Baskı). Pegem Akademi Yayıncılık.

Kutlu, Ö. ve Kula-Kartal, S. (2018). The prominent student competences of the 21st century education and the transformation of classroom assessment. *International Journal of Progressive Education*, *14*(6), 70-82. https://doi.org/10.29329/ijpe.2018.179.6

Kyllonen, P. C. (2012, May 7-8). *Measurement of 21st century skills within the common core state standards* [Paper presentation]. Invitational Research Symposium on Technology Enhanced Assessment. K12 Centre at ETS. https://www.ets.org/Media/Research/pdf/session5-kyllonen-paper-tea2012.pdf

Marzano, R. J. (1992). *A different kind of classroom: Teaching with dimensions of learning.* Association for Supervision and Curriculum Development.

Marzano, R. J. & Heflebower, T. (2012). *Teaching and assessing 21st century skills.* Marzano Research.

Payne, D. & Kyllonen, P. C. (2012, January 11-13). *The role of noncognitive skills in academic success.* [Paper presentation]. 21st Century Knowledge and Skills: The New Curriculum and the Future of Assessment Center for Enrollment Research, Policy, and Practice (CERPP). Marriott Downtown, Los Angeles, United States. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.418.8555&rep=rep1&type=pdf

Popham, J. W. (2000). *Modern educational measurement: Practical guidelines for educational leaders* (3rd ed.). Allyn and Bacon.

UNESCO. (2013-4). *Öğretme ve öğrenme: Herkes için kaliteli eğitimi gerçekleştirmek.* Herkes İçin Eğitim (EFA) Küresel İzleme Raporu. https://www.unesco.org.tr/Content_Files/Content/Sektor/Egitim/EFA2013-4.pdf

Wagner, T. (2015). *21st century schools.* https://www.21stcenturyschools.com/tony-wagner.html

21. Yüzyılda davranışsal değerlendirme ve değerlendirme merkezi uygulamaları

Eren Suna

Eğitim bilimleri kapsamında yapılan değerlendirme çalışmaları, öğrenci özelliklerinin mevcut durumunu belirlemek ya da onların gelecekteki başarıları hakkında kestirim yapmak amacıyla gerçekleştirilmektedir. Bu değerlendirmeler için farklı ölçme araç ve yöntemleri kullanılmakta, bu araç ve yöntemlerden elde edilen sonuçlara dayalı kararlar verilmektedir. Geçmişten günümüze eğitim bilimleri alanında kullanılan ölçme ve durum belirleme yaklaşımlarının çoğunlukla birey algılarına dayalı olması, bu araç ve yöntemlerin geçerliğine dair sorgulamalara neden olmaktadır (Arnold ve Feldman, 1981; van Berkel ve diğ., 2020). Bu durum özellikle bilişsel beceriler dışında kalan içsel ve kişilerarası becerilerde yoğunlaşmaktadır. Bireysel algıya dayalı yanıtlanan araçların bireyleri tanımak adına önemli katkı sağladığı kabul edilmekle beraber, başta sosyal beğenilirlik olmak üzere sonuçların geçerliğini tehdit eden birçok ögeden etkilendiği yapılan çalışmalarda ortaya konulmuştur (Holtgraves, 2004; Krietchmann ve diğ., 2019; Van de Mortel, 2008). Dolayısıyla bu tür araçlar kullanılarak mevcut duruma ya da geleceğe dair sonuçların gerçek durumu hangi ölçüde yansıttığına dair alanyazında tartışmalar devam etmektedir.

Davranışsal değerlendirme, ifade edilen geçerlik sorununu azaltmak, bireylerin davranışlarını gerçekçi ortamlarda değerlendirerek yordama geçerliğini artırmayı ve daha gerçekçi bir değerlendirme süreci sunmayı amaçlamaktadır (Klimoski ve Brickner, 1987). Davranışsal değerlendirme günümüzde eğitim, sağlık, insan kaynakları yönetimi vb. birçok alanda sık kullanılır duruma gelmiştir (Carney ve diğ., 2016; Steele-Schernoff ve Kratochwill, 2003). Bu değerlendirme türünde veri toplama süreci gözlem, doğrudan izleme, davranışsal görüşme vb. birçok farklı yolla gerçekleştirilebilmektedir (Fernández-Ballesteros, 2001; Richard ve Haynes, 2002). Bu değerlendirme türünde bireylerin dikkate alınan özellikleri çeşitli ölçme araç ve yöntemleriyle göstermesi sağlanmakta ve değerlendirme doğrudan bireyin gerçekçi davranışları üzerinden gerçekleştirilmektedir. Davranışların bu yolla doğrudan değerlendirilmesi, daha az hatalı ölçme sonuçlarının elde edilmesine ve yordama geçerliğinin artmasına olanak sağlamaktadır.

Değerlendirme merkezi uygulamaları, davranışsal değerlendirme kapsamında en sık kullanılan değerlendirme yaklaşımları arasında yer almaktadır (Povah ve Thornton, 2011). Bir "yöntemler bütünü" olarak tanımlanan değerlendirme merkezi, bireylerin değerlendirilecek özelliklerinin ve değerlendirme

süreçlerinin ayrıntılı olarak planlandığı, davranışların gösterileceği ortamlara benzer ortamlarda gerçekleştirilen, değerlendirmelerin bağımsız ve uzman puanlayıcılar tarafından gerçekleştirildiği bir süreçtir (Lievens ve Klimoski, 2001; Cristiansen ve diğ., 2013). Değerlendirme merkezinde katılımcılar farklı becerilerinin ölçüldüğü denemelere katılır, bu denemelerin her birinde gerçek yaşam ortamlarına benzer bir ortamda kendilerine kapsamlı bir vaka sunulur, verilen vakaya yönelik tepkileri uzman değerlendiriciler tarafından puanlanır (Povah ve Thornton, 2011; Thornton ve Lievens, 2019). Değerlendiriciler puanlama yaparken vakaya özel geliştirilen dereceli puanlama anahtarını kullanır, davranışlara yönelik ek gözlemlerini de kayda geçirir. Vakaya özgü bir hata ya da yanlılık oluşmaması dolayısıyla her bir özelliğin birden fazla denemede değerlendirilmesi bir gerekliliktir. Değerlendiriciden kaynaklı hata ve yanlılıkları önlemek için de her bir özelliğin birden fazla değerlendirici tarafından puanlanması önemlidir. Tüm üstünlüklerin yanı sıra değerlendirme merkezi denemeleri, geleneksel kişisel algıya dayalı değerlendirme süreçleriyle de desteklenebilmektedir. Diğer bir anlatımla, değerlendirme merkezinde gerçekleştirilen davranışsal değerlendirmeler, çoğunlukla öz, akran ve grup değerlendirme formları gibi çeşitli gözlem araçlarından elde edilen bulgularla birlikte yorumlanabilmektedir. Bu durum, bireye ilişkin daha nesnel ve birey algısını daha gerçekçi yansıtan sonuçlar elde etmeye olanak sağlamaktadır. Dolayısıyla değerlendirme merkezi sonuçlarının bireyin davranışlarını yordamada önemli bir üstünlük sağladığı, öğrencilerin başarısı hakkında sunacağı bu geçerlik türüne ait kanıtların daha güçlü olacağı görülmektedir (Klimoski ve Brickner, 1987; Winfred ve diğ., 2003) Bu bağlamda, değerlendirme merkezleri öğrencilere gerçekçi ortamlar ve vakalar sunarken bir taraftan da öğrenci başarısı hakkında geçerli ve güvenilir sonuçlar sunacaktır.

Değerlendirme merkezi uygulamalarının Türkiye'deki kullanımı çoğunlukla özel sektöre çalışan seçimi, terfi süreçleri ve iş süreçlerinin planlanmasıyla sınırlı kalmıştır (Yelboğa, 2012). Bu sınırlılık, özellikle eğitim alanında değerlendirme merkezinin sunduğu ve ayrıntılı olarak açıklanan üstünlüklerden mahrum kalınmasına yol açmaktadır. Özellikle okullar ve bölgeler arası farklılıkların bulunduğu Türkiye için değerlendirme merkezleri, öğrenme ve öğretme süreçlerine yol gösterici önemli katkılar getirecektir. Eğitimde bilişsel beceriler kadar içsel ve kişilerarası becerilerin de öncelik kazandığı, öğrenci başarısının gelişimi hakkında önemli değerlendirme kararlarının verildiği ve eğitim süreçlerinde 21. yüzyıl becerilerinin öne çıktığı yeni yüzyılda değerlendirme merkezi uygulamalarının rolü de artacaktır.

Bu çalışmada, davranışsal değerlendirme yaklaşımı ve değerlendirme merkezi uygulamalarının 21. yüzyıl becerilerinin ölçülmesinde ve güncel eğitim süreçlerinde nasıl etkin kullanılabileceği ele alınarak tartışılacaktır.

Kaynaklar

Arnold, H. J., & Feldman, D. C. (1981). Social desirability response bias in self-report choice situations. *Academy of Management Journal, 24*(2), 377-385.

Carney, P. A., Palmer, R. T., Fuqua Miller, M., Thayer, E. K., Estroff, S. E., Litzelman, D. K., Biagioli, F. E., Teal, C. R., Lambros, A., Hatt, W. J., & Satterfield, J. M. (2016). Tools to assess behavioral

and social science competencies in medical education: A systematic review. *Journal of the Association of American Medical Colleges, 91*(5), 730-742.

Christiansen, N. D., Hoffman, B. J., Lievens, F., & Speer, A. B. (2013). Assessment centers and the measurement of personality. In Christiansen, N., & Tett, R. (Eds.) *Handbook of personality at work.* Routledge.

Fernández-Ballesteros, R. (2001). *International encyclopedia of the social & behavioral sciences.*

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin, 30*(2), 161-172. https://doi.org/10.1177/0146167203259930

Klimoski, R., & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology, 40*(2), 243-260.

Kreitchmann, R. S., Abad, F. J., Ponsoda, V., Dolores, M. N., & Morillo, D. (2019). Controlling for response biases in self-report scales: Forced-choice vs. psychometric modeling of Likert items. *Front. Psychol, 15.* https://doi:/10.3389/fpsyg.2019.02309

Lievens, F., & Klimoski, R. J. (2001). Understanding the assessment center process: Where are we now? In C.L. Cooper & I.T. Robertson (Eds.) *International review of industrial and organizational psychology* (vol. 16., pp. 245-286). John Wiley & Sons, Ltd.

Povah, N., & Thornton, G. C. (2011). *Assessment centers and global talent management.* Gower.

Richard, D. C. S., & Haynes, S. N. (2002). *Encyclopedia of psychotherapy.* Academic Press.

Steele-Shernoff, E. S., & Kratochwill, T. (2003). The application of behavioral assessment methodologies in educational settings. In Haynes, S. N., & Heiby, E. H. (Eds). *Comprehensive handbook of psychological assessment.* John Wiley & Sons.

Thornton, G. C., & Lievens, F. (2019). Theoretical principles relevant to assessment center design and implementation. In Schlebusch, S., & Roodt, G. (Eds). *Assessment centres: Unlocking people potential for growth* (2nd Edition).

Winfred A. Jr., Day, E. A., Mcnelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. Personnel Psychology, *56*(1), 125-153.

van Berkel, N., Goncalves, J., Hosio, S., Sarsenbayeva, Z., Velloso, E., & Kostakos, V. (2020). Overcoming compliance bias in self-report studies: A cross-study analysis. *International Journal of Human-Computer Studies, 134,* 1-12.

Van de Mortel, T. F. (2008). Faking it: Social desirability response bias in self-report research. *Australian Journal of Advanced Nursing, 25*(4), 40-48.

Yelboğa, A. (2012). Değerlendirme merkezi uygulamalarının Türkiye'deki organizasyonlarda kullanımına ilişkin bir araştırma. *İstanbul Üniversitesi İşletme Fakültesi İşletme İktisadı Enstitüsü Yönetim Dergisi, 23*(72), 8-24.

Sağlık alanında ölçme ve değerlendirme: Tıp fakültesi uygulamaları

S. Ayhan Çalışkan

Giriş

Tıp ve Sağlık Bilimleri Eğitimi alanının önemli bileşenlerinden biri olan tıp fakülteleri, mezuniyet öncesi eğitim programı sonunda pratisyen hekim, mezuniyet sonrası eğitim programı sonunda uzman hekim ve yan dal uzmanı hekim yetiştiren eğitim kurumlarıdır. Tıp fakültelerinde sürüdürülen eğitimin amacı; toplumun sağlık sorunlarını önceleyerek çözmeye çalışan, çalışma yaşamı boyunca öğrenmeyi ve gelişmeyi benimseyen ve uygulayan, bilgi ve yeteneklerini güncelleyen hekim yetiştirmektir (Swanwick, 2018). Tıp fakültelerinde bu amaca erişmek için yürütülen eğitim etkinlikleri diğer lisans eğitimleriyle hem benzer özellikler taşımakta hem de sağlık bilimlerinin öznesi olan insan ve sağlık kavramlarının birleşimiyle ortaya çıkan karmaşık yapıyı destekleyen farklılıklar sergilemektedir. Benzer biçimde söz konusu eğitim etkinliklerinin ölçme ve değerlendirilmesi de benzerlik ve farklılıklar göstermektedir. Tıp eğitimi ile öğrenenlere kazandırılması hedeflenen i) bilgi, ii) beceri ve iii) tutumlar veya profesyonel davranışların ölçme ve değerlendirilmesi, belki de tıp eğiticilerini en çok zorlayan alanların başında gelmektedir. Hekimin sağlık hizmeti sunumunda sergilediği karmaşık süreci değerlendirmek için gereken tüm veriyi tek bir ölçme değerlendirme yönteminin sağlaması beklenemez. Bu nedenle Miller, hekim yetkinliğinin (competency) ölçme ve değerlendirilmesine ilişkin bir çerçeve model ortaya koymuş (Miller, 1990) ve yıllar içinde bu model sık başvurulan ve geliştirilen bir kılavuz haline dönüşmüştür (şekil 1) (Cantillon ve Wood, 2010; Miller, 1990).

Şekil 1

Miller'in yetkinlik piramidi modeli (2, 3).

Miller GE. The assessment of clinical skills/competence/performance (1990)'dan uyarlanmıştır.

Miller'ın yetkinlik piramidi modeli temelinde tıp eğitiminde kullanılan ölçme değerlendirme yöntemleri; hekimin bilgisi *(knowledge)*, bu bilginin nerede ve nasıl klinik nedenselleştirme *(clinical reasoning)* kullanılacağı (knows how), bu bilgi birikimi ve klinik nedenselleştirmeyi de kapsayan psikomotor becerilerini gösterebilmesine *(shows how)* ve tüm bunları sağlık hizmeti sunulan gerçek ortamlarında yapmasına *(does)* ilişkin verilere ulaşmamızı sağlamalıdır (Miller, 1990). Bu nedenle tıp eğitiminde kullanılan ölçme ve değerlendirme yöntemleri oldukça geniş bir çeşitlilik göstermektedir. Söz konusu bu geniş yelpazede yer alan ölçme ve değerlendirme yöntemleri içinde biçimlendirici *(formative)* ve karar verdirici *(summative)* yöntemler yer almaktadır (Koşan, 2016). Bu bildiride tıp fakültesi ölçme ve değerlendirmesi uygulamalarına bir örnek olarak Ege Üniversitesi Tıp Fakültesi Simüle Hasta ile eğitim etkinliklerin ölçme ve değerlendirmesinde kullanılan yöntemin sunulması amaçlanmıştır.

Yöntem

Simüle Hasta (SH), sağlık çalışanlarının eğitimi, sınanması veya iletişim ve muayene beceri uygulamaları için ya bir hasta senaryosunu ya da kendi öykü ve fizik muayene bulgularını sunmak üzere eğitilmiş sağlıklı veya belirli bir hastalığı/bulgusu olan kişidir. Simüle Hastalar; gerçek hastayı, deneyimli bir klinisyenin bile ayırt edemeyeceği biçimde sergilemek (simüle etmek) üzere eğitilirler. Simüle Hastalar gerçek hastanın yalnızca öyküsünü değil, beden dili, fizik bulguları, duygusal ve kişilik özelliklerini de ortaya koyarlar (Barrows, 1987).

Ülkemizde de birçok tıp fakültesinde SH ile eğitim verilmektedir. Ege Üniversitesi Tıp Fakültesi 2003-2004 akademik yılından itibaren 2. ve 3. sınıf eğitim programında; öğrencilerin, iletişim, hasta öyküsü alma ve fizik muayene becerilerinin geliştirilmesi için SH ile eğitimler sürdürmektedir. Bu eğitimlere öğrenciler dört kişilik küçük gruplar halinde katılmaktadırlar. Öğrenciler, planlanan saatlerde SH ile buluşturularak bu eğitimler için özel olarak tasarlanmış ve bir poliklinik veya hasta muayene odasını tüm donanımına sahip SH laboratuvarı odalarında görüşmelerini yapmaktadır. Bir öğrenci doktor görevi ile görüşme yaparken diğer öğrenciler dikkat dağıtmayacak biçimde görüşmeyi gözlemlemektedir. Görüşme sonunda öğrenciler SH ve akranlarından geri bildirim alır. Görüşmenin tüm aşamaları video ile kayıt altına alınmaktadır. Görüşmeyi yapan öğrenci bu video kaydını alıp izleyerek Fakülte Öğretim Yönetim Sisteminde (Moodle) ödevini hazırlar. Ödevde; hastanın öykü ve fizik muayene bulguları yanında yapılandırılmış ve yarı yapılandırılmış başlıklara yanıt vererek özdeğerlendirmesini yapar, güçlü ve zayıf yönleri ile bir sonraki görüşme için öğrenme hedeflerini yazar. Gruptaki dört öğrencinin görüşmeleri tamamlanmasından bir süre sonra, sorumlu öğretim üyesi ile bir geri bildirim oturumu yapılır. Sorumlu öğretim üyesi bu oturuma SH görüşme videolarını izleyerek gelir ve öğrencilere SH görüşme performansları ile ilgili yazılı ve/veya sözlü yapıcı geri bildirim verir. Oturumda ayrıca, SH görüşmesinde yer alan olgulara ve yönetimine ilişkin klinik akıl yürütme tartışması yapılır. Öğrenciler bir sonraki görüşmeleri için; SH, arkadaşları ve öğretim üyesinden aldıkları geri bildirimler ışığında ve öğrenme hedefleri doğrultusunda hazırlanırlar.

Ege Üniversitesi Tıp Fakültesinde SH ile eğitim etkinliklerinin ölçme ve değerlendirmesi aşağıdaki ölçütlere göre yapılmaktadır:

Tablo 1

Görüşmelere, geri bildirim oturumlarına devam, aktif katılım ve katkı

Katılım durumu	Kendi görüşmeniz	Meslektaşlarınızın görüşmeleri			Geri bildirim oturumu	Toplam
		1.	2.	3.		
Puan	5	5	5	5	10	30

Tablo 2

Anamnez, fizik muayene özeti formu, öğrenme ve gelişim hedefleri ve hasta ile görüşme özdeğerlendirme formlarının tamamlanması

Formların tamamlanma durumu	Zamanında		Gecikmeli		Geri bildirim oturumuna dek tamamlanmadı
	Yeterli	Düzeltme gerekli	Yeterli	Düzeltme gerekli	
Puan	20	15	10	5	0

Tablo 3

Eğitici gözlemi

		Çok iyi	İyi	Orta	Yetersiz	Çok yetersiz
Değerlendirme başlığı	Puan	10	8	6	4	2
1. Hasta ile iletişim becerisi						
2. Anamnez alma becerisi						
3. Fizik Muayene becerisi						
4. Klinik akıl yürütme sürecine katkısı						
5. Ekip arkadaşlarının görüşmelerine katkısı						

Simüle hasta görüşmelerinden başarılı kabul edilebilmek için yukarıda açıklanan ölçütlere göre en az 80 puan almak gereklidir.

*(*Covid-19 pandemi süreci öncesine ait ölçme ve değerlendirme ölçütleridir. Pandemi sürecinde hem SH görüşme yöntemi hem de ölçme ve değerlendirme ölçütlerinde değişiklik yapılmıştır.)*

Sonuçlar

Sağlık alanında ölçme ve değerlendirme uygulamalarının tıp fakültesindeki yansımalarının tartışıldığı bu bildiride Ege Üniversitesi Tıp Fakültesi Simüle Hasta ile eğitim etkinliklerin ölçme ve değerlendirmesinde kullanılan yöntem sunulmuştur. Bu yöntem her ne kadar karar verdirici yapıda ve sonuç odaklı görünse de, öğrencilerin eğitimler kapsamında hazırladıkları ödevler, bireysel performanslarına ilişkin yansıtma *(reflection)*, özdeğerlendirme ve özdenetimli öğrenme *(self directed*

learning) boyutları da olan ve aynı zamanda sürecin değerlendirildiği biçimlendirici bileşenleri de içermektedir. Simüle Hasta ile eğitimi ölçme ve değerlendirme yönteminin bu çok kaynaklı, çok boyutlu ve eksiklerin tamamlanmasına izin veren yapısı, öğretim üyelerinin yapıcı geri bildirimleri ile bütünleştiğinde öğrencilerin başarı ve programa ilişkin beğenilerini çok yüksek düzeylere taşımaktadır.

Kaynaklar

Barrows, H. S. (1987). *Simulated (standardized) patients and other human simulations.* Health Sciences Consortium.

Cantillon, P. and Wood, D. (Eds.) (2010). *ABC of learning and teaching in medicine* (2nd ed.). BMJ Books.

Koşan, A. A. ve Çalışkan, S. A. (2016). Tıp eğitimi programlarında ölçme ve değerlendirme. In İ. Sayek (Ed.), *Tıp eğiticisi el kitabı* (1. Baskı, pp. 215–240). Güneş Tıp Kitabevleri.

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Acad Med., 65*(9), 63–67. https://doi.org/10.1097/00001888-199009000-00045

Swanwick, T. (2018). Understanding medical education. In T. Swanwick, K. Forrest, and B. C. O'Brien (Eds.), *Understanding medical education: Evidence, theory, and practice* (pp. 1–6). Wiley. https://onlinelibrary.wiley.com/doi/full/10.1002/9781119373780.ch1

Sağlık alanında ölçme değerlendirme uygulamaları: Hemşirelik fakültesi uygulamaları

Hale Sezer

Giriş

Hemşirelik, bir toplum hizmeti olarak uzun yıllar varlığını korumuş, insanların sağlığını geliştirmek ve hastalandığında bakımını sağlamak isteği ile ortaya çıkmıştır (Ergol, 2011). Temel sağlık hizmetlerinin başarısı, sağlık bakım sistemindeki değişikliklerin hastanın bakım standardını arttıracak şekilde kullanılmasına, bakım verecek hemşire ve diğer sağlık personelinin iyi yetiştirilmesine bağlıdır. Hemşirelik eğitimi en az dört yıl veya 4600 saatlik teorik ve klinik eğitimi kapsar. Teorik eğitimin süresi toplam sürenin en az üçte biri, klinik eğitimin süresi ise toplam eğitimin yarısı kadardır (HUÇEP, 2014). Sağlık hizmet sunumunda hemşirelerden beklenen temel yetkinliklerin çerçevesi, ülkemizde ilk defa "Hemşirelikte Temel Yetkinlikler Kılavuzu" şeklinde Sağlık Hizmetleri Genel Müdürlüğü tarafından hazırlanmıştır (Sağlık Bakanlığı, 2021). Bu yetkinlikler profesyonellik, etkili iletişim, kanıta dayalı uygulama, bakım yönetimi, kalite iyileştirme, ekip çalışması ve işbirliği, mesleki liderlik olarak tanımlanmıştır (Sağlık Bakanlığı, 2021). Mezuniyet öncesi eğitimde hemşirelik öğrencilerinin bu yetkinlikleri kazanarak mezun olması hedeflenmektedir. Hemşirelik lisans eğitimi süresince tanımlanan "Hemşirelikte Temel Yetkinliklere" ulaşabilmek için müfredat içerisinde yer alan dersler ile belirli yeterliliklerin kazandırılması gerekmektedir. Hemşirelik öğrencilerinin bu yeterliklere farklı ölçme ve değerlendirme yöntemleri kullanılarak ulaşıldığı belirlenebilir. Hemşirelik eğitimindeki yeterliklerin belirlenmesinde Miller'ın Yeterlikler piramidinden yararlanılmaktadır (Miller, 1990). Miller piramidinin dört düzey belirlemesine göre öğrencilerin değerlendirilmesinde kullanılacak ölçme araçlarında "bilir", "nasıl yapılacağını bilir", "gösterir", "yapar" boyutunda güvenilir ve geçerli öğrenci başarısını ölçebilen farklı ölçme araçlarının kullanılması önerilmektedir (Miller, 1990). Hemşirelik eğitiminde Miller primadine göre "bilir" düzeyinde sözlü ve yazılı yoklamalar, çoktan seçmeli sorular ve raporlardan yararlanılmaktadır. "Nasıl olduğunu bilir" düzeyinde ise vaka sunumları, proje ödevleri, vakaya dayalı çoktan seçmeli sorular, CORE sınavları olarak adlandırılan kliniğe yönelik mantık yürütme sınavları gerçekleştirilir. "Gösterir" basamağında objektif yapılandırılmış klinik sınavlar (OSCE), simülasyon, simüle hasta, kontrol listeleri ve rubrikler kullanılmaktadır. "Yapar" düzeyi uygulama alanlarında hemşirelik uygulamalarını gerçekleştirilirken gözlem, portfolyo, akran değerlendirmesi, öz değerlendirme, mini klinik sınav ve 360 derece değerlendirmeden yararlanılmaktadır (HUÇEP, 2014;

Tengiz ve Şahin, 2014). Bu değerlendirmeler dışında mezuniyet aşamasına gelmiş hemşirelik öğrencilerinin klinik uygulama sırasında sergilediği hemşirelik yeterliklerinin değerlendirilmesinde farklı modellerden de yararlanılmaktadır. Bunlardan biri "Uygulamanın Yapılandırılmış Gözlem ve Değerlendirilmesi" olarak tanımlanan SOAP modelidir. Bu derlemede mezuniyet aşamasındaki hemşirelik öğrencilerinin yeterliklerinin belirlenmesinde kullanılan SOAP modelinin sunulması amaçlanmaktadır.

Yöntem

Bu derlemede, mezuniyet aşamasındaki hemşirelik öğrencilerinin yeterliklerinin belirlenmesinde kullanılan SOAP modelinin tanımlanması, SOAP modeline uygun bir değerlendirme planlanması ve kullanılması ile ilgili bilgiler verilmesi ve modelin yürütülmesi sırasında ortaya çıkabilecek sorunlara çözüm önerileri sunulması amaçlanmaktadır.

Sonuçlar

Hemşirelik öğrencilerinin klinik yeterliliklerinin değerlendirilmesi, eğitimcileri uzun bir süre boyunca geçerlilik ve güvenilirlik sorunlarıyla karşı karşıya bırakmıştır (Levett-Jones ve diğ., 2011). SOAP, hemşirelik öğrencilerinin öğrenmesini motive eden, eleştirel düşünmeyi teşvik eden ve mezunların profesyonel uygulamaya hazır olduklarını teyit eden, tam gün uygulama odaklı bir klinik yeterlilik değerlendirme yaklaşımıdır (Levett-Jones ve diğ., 2011). SOAP birçok değerlendirme modelinden farklıdır; 'kontrol listesi' yoktur; bağlamsal olarak duyarlıdır; öğrencinin gözlemlenen davranışlarından daha fazlasını anlamaya çalışır; ayrıca öğrencinin uygulamasını bilgilendiren bilgi, değer ve tutumları da inceler. SOAP modeli, hemşirelik öğrencilerinin klinik bir bağlamda üstlendikleri klinik bilgi, beceri, davranış, tutum ve değerlerin 6 saatlik bütünsel bir değerlendirmesidir. SOAP modelini kullanan eğitim programlarında öğrencilerin programlarını tamamlayabilmeleri ve mezun olabilmeleri için SOAP'ta 'yeterli' bir derece almaları gerekmektedir. SOAP modelinde gözleme dayalı bir performans değerlendirmesi gerçekleştiği için değerlendiricilerin iyi belirlenmesi ve SOAP modelinde değerlendirme ile ilgili eğitim almaları gerekmektedir. SOAP modelinde değerlendiriciler klinik deneyimleri nedeniyle seçilen üniversitede çalışan hemşireler olmalıdır. Değerlendiriciler, sınavın amacının ve sürecinin tanıtıldığı, standart hasta ile öğrenci görüşme videolarını kullanarak değerlendirmeyi uygulama fırsatı sağlanan iki günlük bir eğitime katılırlar. Eğitim süresince diyalog ve problem çözmeyi teşvik etmek için varsayımsal durumlar sunulur. Tartışma konuları, profesyonel muhakemeyi, sürecin titizliğini ve adaleti sağlamayı içerir. SOAP modeline göre klinik değerlendirme gözlem, mülakat ve formatif-summatif geribildirimi içeren üç aşamadan oluşmaktadır. Sınavının gerçekleştirilmesinde öğrencilerin olağan hasta bakımı faaliyetlerine katıldıkları iki-üç saatlik bir gözlem süresi boyunca, öğrencinin ayrı hemşirelik davranışlarının her biri, bir durum, eylem, sonuç formatı kullanılarak değerlendirici tarafından sırayla belgelenir. Mülakat kısmında açık uçlu sorulardan yararlanılarak değerlendirici gözlemlenen öğrencinin bilgi, beceri, tutumlarını ve değerlerini ortaya çıkarmaya çalışır. Mülakat süresince öğrencinin eleştirel düşünme ve klinik akıl yürütme becerilerine ilişkin kanıt arar. Mülakat sonrası elde edilen veriler ile

öğrencinin hemşirelik yeterlik standartlarını karşılayıp karşılamadıklarına bakılır. Son olarak öğrenci için hem formatif hem de summatif geribildirim oturumu gerçekleştirilir. Bu çözümle oturumu sırasında, öğrencinin klinik güçlü yönleri ve gelişmeyi gerektiren alanları açıkça tanımlanır. Öğrencinin eleştirel öz refleksiyonu desteklenerek kendisi için iyileştirme ihtiyacını belirlemesi sağlanır. Geribildirim oturumu sonrasında öğrenci "yeterli", "yeterlik alanının düzeltilmesi bekleniyor" ya da "yeterli değil düzeltme ve yeniden değerlendirme gerekir" şeklinde bir sonuç almaktadır. SOAP sınavından yeterli olmayan öğrenciler iyileştirme sürecine katılırlar. Öğrenci, iyileştirme faaliyetleri içerisinde öğrenim sözleşmeleri, ek beceri uygulamaları, kritik bir olay hakkında yansıtıcı bir rapor yazma, kanıt araştırma ve uzman bir hemşire ile zaman geçirme gibi faaliyetlerde bulunulduktan sonra tekrar değerlendirilmeye alınır (Levett-Jones ve diğ., 2011). Sonuç olarak SOAP modeliyle değerlendirme de hemşirelik öğrencilerinin klinik uygulamasının eleştirel bir şekilde incelenmesi ve yansıtılması şansını verir. Öğrencilerin sınav süresince güçlü yönleri ve geliştirilmesi gereken yeterlik alanları belirlenir. Yeterliğe dayalı klinik performans değerlendirmesi gerçekleştirilmiş olur.

Kaynaklar

Ergol, S. (2011). Türkiye'de yükseköğretimde hemşirelik eğitimi. *Journal of Higher Education and Science*, *1*(3), 152–155. https://doi.org/10.5961/jhes.2011.022

HUÇEP. (2014). *Hemşirelik ulusal çekirdek eğitim programı (HUÇEP)*. https://www.yok.gov.tr/Documents/Kurumsal/egitim_ogretim_dairesi/Ulusal-cekirdek-egitimi-programlari/hemsirelik_cekirdek_egitim_programi.pdf

Levett-Jones, T., Gersbach, J., Arthur, C., & Roche, J. (2011). Implementing a clinical competency assessment model that promotes critical reflection and ensures nursing graduates' readiness for professional practice. *Nurse Education in Practice*, *11*(1), 64–69. https://doi.org/10.1016/j.nepr.2010.07.004

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine*, *65*(9), 63–67. http://winbev.pbworks.com/f/Assessment.pdf

Sağlık Bakanlığı (2021). *Hemşirelikte temel yetkinlikler kılavuzu*. Sağlık Hizmetleri Genel Müdürlüğü.

Tengiz, F. İ., ve Şahin, H. (2014). Klinikte eğitimde yeni bir ölçme yöntemi : Mini klinik değerlendirme. *Tıp Eğitimi Dünyası*, *13*(39), 13–18. https://dergipark.org.tr/tr/download/article-file/199277

7. Uluslararası Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi Sonuç Bildirgesi

Gazi Üniversitesi iş birliğiyle gerçekleştirilen 7. Uluslararası Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi'nin bildiri sunumları, çağrılı konuşmacıların sunumları, çalıştaylar ve panel çalışmaları tamamlanmıştır. 2008 yılında Ankara Üniversitesi ev sahipliğinde başlayan kongremizin yedincisi gerçekleştirilmiştir. İçinde bulunduğumuz salgın süreci dolayısıyla 2020 yılında yapılması planlanan kongremiz bir yıl ertelenerek 2021 yılında ilk defa çevrim içi olarak düzenlenmiştir.

Kongremize farklı üniversitelerde görev yapmakta olan akademisyenler, ölçme ve değerlendirme alan uzmanları, öğretmenler, ölçme ve değerlendirme merkezlerinde görev yapan uzmanlar ve lisansüstü program öğrencileri katılmıştır. Kongremiz toplam 164 kişinin katılımıyla gerçekleştirilmiştir.

Gazi Üniversitesi iş birliği ve ev sahipliğinde düzenlenen kongrenin 1 Eylül'de çevrim içi ortamda gerçekleşen açılış, Doç. Dr. Ayfer Sayın'ın moderatörlüğünde saygı duruşu ve İstiklal Marşı'nın okunmasıyla başladı. Kongrenin açılış konuşmalarını, EPODDER Yönetim Kurulu Başkanı Prof. Dr. Nuri Doğan, Gazi Eğitim Fakültesi Dekanı Prof. Dr. Mahmut Selvi, Gazi Üniversitesi Rektörü Prof. Dr. Musa Yıldız ve Milli Eğitim Bakan Yardımcısı Prof. Dr. Petek Aşkar gerçekleştirdi.

7. Uluslararası Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi'nin bilim kurulunda 5 farklı ülkede görev yapmakta olan toplam 76 öğretim üyesi görev almıştır. Dört gün süren kongre; çağrılı konuşmacılar, çalıştaylar, paneller ve sözlü sunumlar olmak üzere dört ana başlıkta planlanmıştır. Kongrede Prof. Dr. Akihito KAMATA, Prof. Dr. Bruno D. ZUMBO, Prof. Dr. Fatma BIKMAZ, Prof. Dr. İnayet AYDIN, Prof. Dr. Jimmy de la TORRE, Doç. Dr. Okan BULUT, Doç. Dr. Önder SÜNBÜL, Doç. Dr. Sedat ŞEN ve Doç. Dr. Serkan ARIKAN olmak üzere dokuz çağrılı konuşmacı sunumlarını gerçekleştirmişlerdir. Kongrede Prof. Dr. Akihito KAMATA, Prof. Dr. Nuri DOĞAN, Prof. Dr. Tuncay ÖĞRETMEN, Eğitimci Selim DAŞÇIOĞLU ve Eğitimci Zeynep UZUN üç farklı çalıştay, Gazi Eğitim Fakültesinde yüz yüze gerçekleştirilmiştir. "21. Yüzyıl Becerileri İçin Ölçme ve Durum Belirleme Yaklaşımı", "Tıp Eğitiminde Ölçme ve Değerlendirme", "Sağlık Alanında Ölçme Değerlendirme Uygulamaları" ve "Geçiş Sistemleri" başlıklı olmak üzere dört farklı panel düzenlenmiştir.

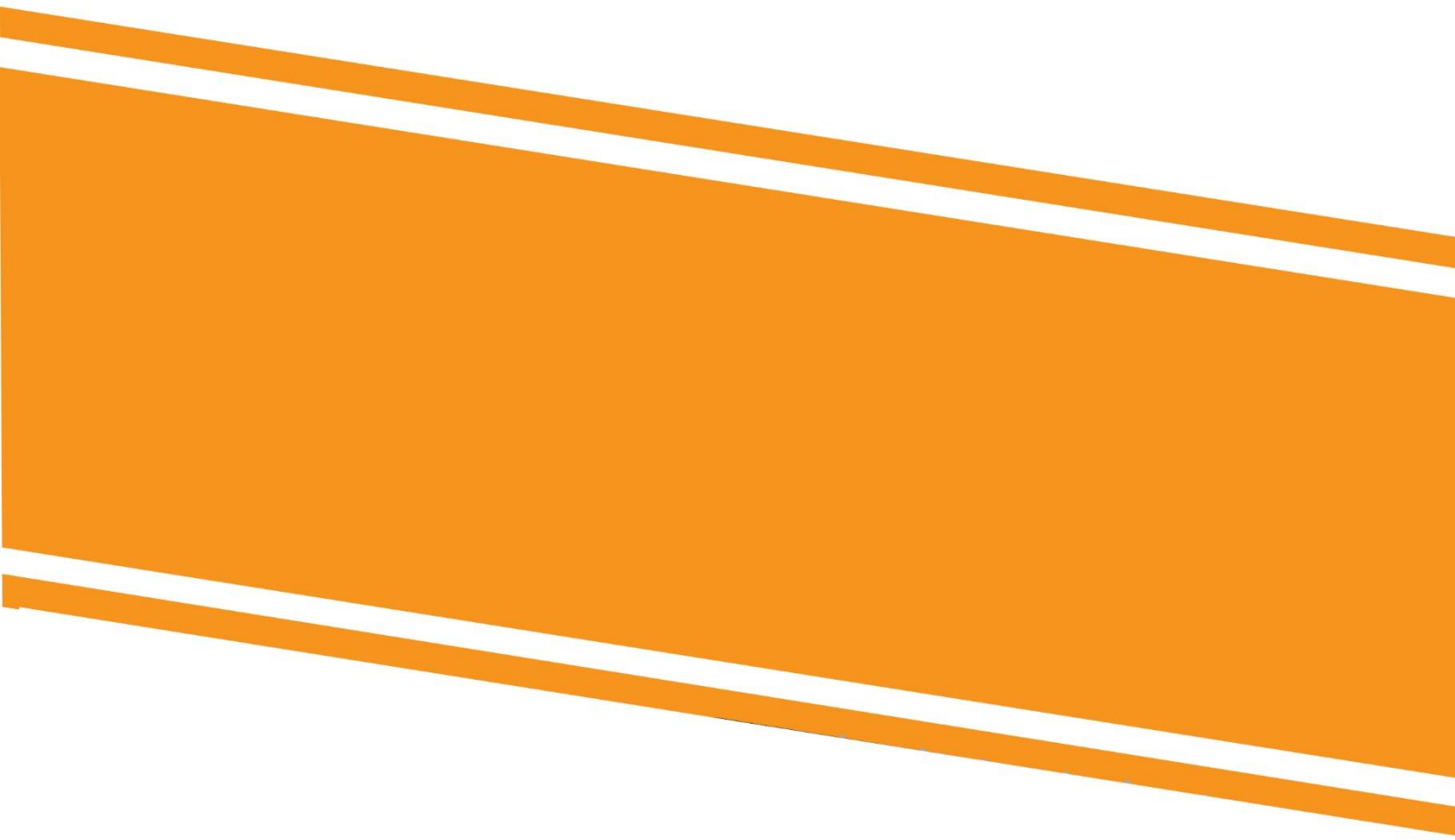
Konferansımızda 8 paralel oturumda düzenlenen toplam 24 farklı oturumda toplam 122 bildiri sunulmuştur. Oturumların birinde İngilizce, diğer oturumlarda Türkçe sunum dilinde bildiri sunumları gerçekleştirilmiştir. Bildirilerde yöntemsel ve uygulamalı çalışmalara ağırlık verildiği gözlemlenmiştir. Bildiriler sırasıyla ölçme ve değerlendirmede yeni yaklaşımlar, kuram temelli geçerlik çalışmaları, geniş ölçekli testler, araştırma deseni olarak ölçme ve değerlendirme uygulamaları, okullarda ölçme ve

değerlendirme uygulamaları, 21. yy becerilerinin ölçülmesi, uzaktan eğitimde ölçme ve değerlendirme, ölçme ve değerlendirmede etik konu başlıklarını içermektedir. Bilişsel tanı modelleri, yapay sinir ağları, meta-analiz, test geliştirme, geçerlik ve güvenirlik kanıtları, TIMSS ve PISA sonuçlarının incelenmesi vb. başlıklar öne çıkmaktadır.

Kongre sonucunda;

- Okullarda öğrenme-öğretme ve ölçme ve değerlendirme etkinliklerinin bilişsel, içsel ve kişiler arası yeterliklerini birlikte kullanmasını gerekliğine yönelik çalışmalar yapılması,
- Değerlendirme merkezi uygulamalarının ölçme ve değerlendirme alan uzmanlarınca desteklenmesi,
- Sağlık alanı gibi alanlarda da ölçme ve değerlendirme uygulamalarının desteklenmesi,
- Ölçme ve değerlendirme ve araştırma süreçlerinde etik konusuna ağırlık verilerek etik kurulların faaliyetlerinin denetlenmesi ve desteklenmesi,
- Geniş ölçekli testlerden elde edilen ya da büyük simülasyon verilerinin kullanımında veri kullanım prosedürlerine dikkat edilecek faaliyetler gerçekleştirilmesi,
- Kademeler arası geçişte yaşanan sorunların hâlen devam ettiği, bu sorunlara ilişkin ortak bir çalışma yapılması,
- Merkezi sınav yapan kurumların bağımsız denetleme kurullarınca denetlenmesi,

önerilmiştir.




EPODDER
GAZİ ÜNİVERSİTESİ
Gazi
1926