Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN:1309-6575 *Kış* 2017 *Cilt: 8- Sayı: 4*
Winter 2017 *Volume: 8- Issue: 4*

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309 – 6575

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (EPOD) yılda dört kez yayınlanan hakemli ulusal bir dergidir. Yayımlanan yazıların tüm sorumluğu ilgili yazarlara aittir.

İletişim
e-posta: epod@epod-online.org
Web: http://epod-online.org

Journal of Measurement and Evaluation in Education and Psychology (EPOD) is a national refereed journal that is published four times a year. The responsibility lies with the authors of papers.

Contact
e-mail: epod@epod-online.org
Web: http://epod-online.o

Dizinleme / Abstracting & Indexing
DOAJ (Directory of Open Access Journals), TÜBİTAK Ulakbim Sosyal ve Beşeri Bilimler Veri Tabanı, Tei (Türk Eğitim İndeksi)

Hakem Kurulu / Referee Board

Yavuz AKPINAR (Boğaziçi Üni.)
Yeşim ÖZER ÖZKAN (Gaziantep Üni.)
Zekeriya NARTGÜN (Abant İzzet Baysal Üni.)

*Ada göre alfabetik sıralanmıştır. / Names listed in alphabetical order.

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Aralık 2017, Sayı: 8-4)

Journal of Measurement and Evaluation in Education and Psychology (December 2017, Issue: 8-4)

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(4);344-353

Anket Aracılığıyla Toplanan Veriye Güvenebilir Miyiz? : İdeal Cevaplama Süresi ve Cronbach's Alpha Yanılgısı*

Can We Trust The Data Collected Through Survey? : Ideal Response Time and Delusion of Cronbach's Alpha*

Volkan DOĞAN **

Öz
Sosyal ve davranışsal araştırma disiplinlerinde birçok araştırma anket verisine dayanmaktadır. İdeal olarak anketi cevaplayıcıların dikkatli şekilde cevap verdikleri varsayılmaktadır. Fakat yakın geçmişte araştırmaların sonuçları dikkatsiz cevaplayıcı tehditinin düşünülenden daha büyük olduğunu göstermektedir. Bu araştırma Türk kültürü kapsamında kesitsel olarak dikkatsiz anket cevaplayıcılarının oranının ne düzeyde olduğuna dair keşifsel bulgular sunmakta, ideal cevaplama süresi hesaplamaya yönelik yeni bir metodolojik yaklaşım önermekte ve Cronbach's Alpha yanılgısı kavramına değinmektedir. Bu amaçla, 154 Türk üniversite öğrencisinden sınıf ortamında ve 148 Türk yetişkinden çevrim içi ortamda toplanan anket verisi üzerinden analizler gerçekleştirilmiştir.

Anahtar Kelimeler: Anket verisi, veri kalitesi, dikkatsiz cevaplayıcılar, ideal cevaplama süresi.

Abstract
Much of the studies in the social and behavioral sciences is based on the data collected through survey. To obtain unbiased knowledge, survey respondents are required to answer survey items/questions in a careful way. However, past research demonstrated that research projects are under the threat of careless/inattentive respondents. This paper provides some evidences regarding the rate of careless/inattentive respondents within Turkish respondents. To this end, this paper comprises of two survey data—154 Turkish undergraduate students (face-to-face survey) and 148 online Turkish respondents (via Facebook)—to provide some descriptive evidences. Furthermore, a new approach toward calculating ideal response time was proposed in the current paper. Lastly, the delusion of Cronbach's Alpha was discussed.

Keywords: Survey data, data quality, careless respondents, ideal response time.

GİRİŞ

Davranışsal ve sosyal bilimlerde gerçekleştirilen araştırmaların çoğu cevaplayıcıların kendilerini ölçme araçları aracılığıyla rapor etmesine dayanmaktadır. Araştırmacılar, cevaplayıcıların kendilerini rapor etmesi neticesinde elde edilen veriyi analiz ederek araştırma modellerini ve hipotezlerini test etmektedirler. Peki ya veri toplama süreci neticesinde elde edilen veri iyi kalitede değil ise? İyi kalitede veri, iki farklı perspektiften tanımlanabilir. İlk perspektif, araştırmacının araştırma modeli ve hipotezlerini destekleyen bir veri seti olarak tanımlanan kaliteli veridir. İkinci perspektif ise araştırmanın katılımcılarının gerçekten dikkatli şekilde kendilerini rapor ettiği veri seti olarak tanımlanan kaliteli veridir. İkinci perspektif doğrultusunda kaliteli veri kavramına yaklaşım daha objektif bir yaklaşımdır. Bu ikinci perspektiften, kaliteli veri elde etmenin önündeki engeller ve bu süreçte dikkatsiz cevaplayıcıları tespit etmeyi sağlayacak metodolojik yöntemlere ilişkin geçmişte birtakım araştırmalar gerçekleştirilmiştir (Berinsky, Margolis ve Sances, 2014; Curran, 2016; Oppenheimer, Davis ve Davidenko, 2009). Ancak bu metodolojik araştırmaların sadece Batılı

* Bu çalışmanın bir kısmı 21. Pazarlama Kongresi'nde sunulmuştur.
** Arş. Gör. Dr., Eskişehir Osmangazi Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Eskişehir-Türkiye, e-posta: vodogan@ogu.edu.tr, ORCID ID: orcid.org/0000-0001-5440-9440.

DOI: 10.21031/epod.313692
Geliş Tarihi: 16.05.2017
Kabul Tarihi: 03.08.2017

toplumlardaki cevaplayıcılar kapsamında gerçekleştirildiği dikkat çekmektedir. Bu yüzden, bu araştırmanın amacı, Türk kültüründe gerçekleştirilen anket uygulamalarında dikkatsiz cevaplayıcıların ne oranda olduğunu keşifsel olarak tespit edebilmek ve kaliteli anket verisi elde edebilmek için anketlerde yer alan ifade/soru başına verilmesi gereken minimum süreye ilişkin keşifsel bir eşik değer belirlemeye çalışmaktır.

Anket aracılığıyla veri toplama, davranışsal ve sosyal bilimlerde geçmişten günümüze değin en sıklıkla kullanılan metodolojik uygulamalardan bir tanesidir. Özellikle World Wide Web teknolojisinin gelişmesiyle birlikte, anket verisi toplama araştırmacılar tarafından daha da yaygın kullanılmaya başlanmıştır (Gosling, Vazire, Srivastasa ve John, 2004). Anket verisinin giderek daha kolay toplanabilir olması avantajların yanı sıra bazı sorunları da beraberinde getirmiştir. Bu sorunların en başında, dikkatsiz cevaplayıcılardan toplanmış olan kalitesiz anket verisi gelmektedir. Dikkatsiz cevaplayıcılar, anket ve deneysel araştırma neticesinde elde edilecek olan veriye yanlı bulgulara sebep olacak şekilde zarar verebilmektedir. Örneğin; normalde aralarında mükemmel düzeyde pozitif ilişki (r = 1.00) beklenen iki değişken arasındaki ilişki, dikkatsiz cevaplayıcılardan dolayı çok daha düşük düzeyde ilişkili (r = 0.30) tespit edilebilmektedir (Woods, 2006). Üstelik bu yanlı bulgulara ulaşmak için, dikkatsiz cevaplayıcıların toplam cevaplayıcılar içindeki oranının %5 kadar az olması bile yeterlidir (Huang, Liu ve Bowling, 2015).

Dikkatsiz cevaplayıcıları tespit etmeye yönelik birçok metodolojik araştırma literatürde yer almaktadır (Curran, 2016). Bu yöntemlerden en dikkat çekenler; "talimatsal manipülasyon kontrolü (instructional manipulation check)" (Oppenheimer, Davis ve Davidenko, 2009), "cevap süresi (response time)" (Huang, Curran, Keeney, Poposki ve Deshon, 2012), "sahte ifade (bogus-item)" (Beach, 1989) yöntemleridir. Geçmiş araştırmalar "talimatsal manipülasyon kontrolü (instructional manipulation check)" ve "sahte ifade (bogus-item)" yöntemlerinden faydalanarak dikkatsiz anket ve/veya deney cevaplayıcılarının oranlarına ilişkin keşifsel tespitlerde bulunmuşlardır. Bu keşifsel araştırmalar neticesinde dikkatsiz cevaplayıcıların oranları; %3.3 - %59.1 arası (Anduiza ve Galais, 2017), %8 - %11 arası (Hauser ve Schwarz, 2015), %63 (Berinsky, Margolis ve Sances, 2014), %46 (Oppenheimer, Davis ve Davidenko, 2009) ve %11 (Meade ve Craig, 2012) olarak tespit edilmiştir. Bu keşifsel bulgular, anket aracılığıyla toplanan verilerin kalitesinin dikkatsiz cevaplayıcıların varlığından dolayı tehdit altında olduğunu göstermektedir. Ayrıca bu geçmiş araştırmaların Batılı gelişmiş ülkelerde yaşayan cevaplayıcılar kapsamında gerçekleştirildiği, gelişmekte olan ülkelerde gerçekleştirilmiş bir geçmiş araştırma olmadığı dikkat çekmektedir.

Dikkatsiz cevaplayıcıları tespit etmeye yönelik bir diğer yaklaşım ise cevaplayıcıların anket cevaplarken geçirdikleri zamanı (süre) hesaplamayı içermektedir. Çünkü kötü veri kalitesinin başlıca sebebi olan dikkatsiz cevaplayıcıların temel karakteristiği, anket ifadeleri/soruları üzerinde çok az zaman harcıyor olmalarıdır (Curran, 2016; Niessen, Meijer ve Tendeiro, 2016; Oppenheimer, Davis ve Davidenko, 2009). Bir diğer deyişle, dikkatsiz cevaplayıcıların anket veya deneyi kısa sürede tamamladıkları yönünde bulgulara ulaşılmıştır. Özellikle, dikkatsiz cevaplayıcıları tespit etmeye yönelik farklı yöntemlerin performansı karşılaştırıldığında dikkatsiz cevaplayıcıları en etkili tespit eden yöntemin cevaplama süresi yöntemi olduğu bulunmuştur (Niessen, Meijer ve Tendeiro, 2016). Ancak dikkatsiz cevaplayıcıları dikkatli cevaplayıcılardan ayırabilmek için cevap süresi açısından spesifik bir eşik değer belirlenmesine yönelik bir araştırma günümüze değin gerçekleştirilmemiştir. Sadece Huang, Curran, Keeney, Poposki ve Deshon (2012), dikkatsiz cevaplayıcıları tespit etmeye yönelik olarak ifade/soru başına iki saniye şeklinde bir öneride bulunmuşlardır. Bir diğer deyişle, ifade/soru başına iki saniyeden az zaman harcayanlar dikkatsiz cevaplayıcılar olarak tespit edilebilir önerisinde bulunulmuştur.

Araştırmanın Amacı

Dikkatsiz cevaplayıcıların tespitine ilişkin geçmiş araştırmalar, Türkiye'de anket cevaplayıcıları kapsamında gerçekleştirilmiş bir dikkatsiz cevaplayıcı oranı belirleme araştırmasının olmadığını ve ayrıca kaliteli veri elde edebilmek için ifade/soru başına harcanması beklenen bir eşik değer süresi belirlemeye yönelik metodolojik bir araştırmanın bugüne değin yapılmadığını göstermektedir.

Bu kapsamda bu araştırmada aşağıda yer alan sorulara cevap aranmaktadır:
a) Türk kültüründe uygulanan anketlerde dikkatsiz cevaplayıcı oranı ne düzeydedir?
b) Bu oran, anketin hangi yöntem ile (çevrimiçi-çevrimdışı) toplandığına bağlı olarak farklılaşır mı?
c) Kaliteli veri elde edebilmek için ankette yer alan ifade başına cevaplayıcının ne kadar süre harcamış olması beklenmelidir?
d) İçsel tutarlılığın bir göstergesi olan Cronbach's Alpha katsayısının yüksek olması verinin kalitesine ilişkin bir problemin işaretçisi olabilir mi?

YÖNTEM

Bu araştırmanın yöntemine ilişkin detaylara aşağıda yer alan alt başlıklar kapsamında değinilmiştir.

Araştırmanın Türü

Bu araştırma kesitsel (cross-sectional) niteliğe sahip bir araştırmadır. Bu yüzden belirli bir anda katılımcıların veri toplama aracı olan anket içerisinde yer alan ifadelere cevapları kayıt altına alınmıştır. Ayrıca mevcut duruma ilişkin bir kesit sunulmaya çalışıldığı için bu araştırma, tanımlayıcı (descriptive) bir araştırmadır. Araştırmanın yöntemine ilişkin detaylı bilgilere ilerleyen alt başlıklarda yer verilmiştir.

Çalışma Grubu

İlk uygulama neticesinde 154 Esogü öğrencisinden, ikinci uygulama neticesinde ise 148 cevaplayıcıdan (online ortamda) veri toplanmıştır. Bir diğer deyişle, toplam 302 veri sınıf ortamında ve online ortamda toplanmıştır.

İşlem (Prosedür)

Bu araştırma iki farklı veri toplama sürecini kapsamaktadır. İlki, Eskişehir Osmangazi Üniversitesi (ESOGÜ) İ.İ.B.F. ve Turizm Fakültesi öğrencilerini üzerinde sınıf ortamında gerçekleştirilen anket uygulamasıdır. İkincisi, akademikpersonel.org'un Facebook sayfası aracılığıyla gerçekleştirilen anket çağrısı doğrultusunda çevrimiçi (online) ortamda gerçekleştirilen anket uygulamasıdır. Özellikle, öğrenciler kapsamında gerçekleştirilen uygulamada öğrencilerin anket cevaplama süresince birbirleri ile iletişime geçmemesi gerektiği uyarısında bulunulmuştur. İlk ve ikinci veri toplama sürecinde kullanılan anketin içeriği aynıdır. Tek farklılık, çevrimiçi (online) anket uygulamasında cevaplayıcıların her bir sayfada geçirdikleri süre kodlar aracılığıyla (Qualtrics) saniye cinsinden hesaplanmıştır.

Uygulanan anket sırasıyla; dikkatsiz cevaplayıcıları belirlemeye yönelik "talimatsal manipülasyon kontrol (instructional manipulation check) (IMC)" (Ek-1) sayfasını, dikkat sorusunu, Türk kültürü kapsamında geliştirilmiş olan ve 18 ifade ile beş boyuttan oluşan lüks tüketim eğilimi ölçeği (LTE; Doğan, Özkara ve Doğan, 2016) ifadeleri ile bir adet kontrol ifadesini içermektedir. Kontrol ifadesi olarak "Doğa üstü güçlere sahip bir dinozorum" ifadesi kullanılmıştır.

Veri Toplama Araçları

Dikkatsiz cevaplayıcı ölçümü

Dikkat sorusuna ve kontrol ifadesine verilen cevap doğrultusunda olmak üzere iki farklı şekilde dikkatsiz cevaplayıcılar belirlenmiştir. Girişte yer alan IMC, araştırmacı tarafından dikkatsiz cevaplayıcıları tespit edebilmek için Oppenheimer, Davis ve Davidenko (2009)'nin yaklaşımından yararlanarak tasarlanmıştır. IMC içinde cevaplayıcıların dikkatli cevaplayıcı olduklarını

gösterebilmek için bir sonraki soruya araştırmacı tarafından belirlenmiş "okudum" cevabını vermeleri istenmiştir.

Öte yandan, anket ifadeleri arasında konumlandırılan kontrol ifadesine ("Doğa üstü güçlere sahip bir dinozorum") verilen cevaba göre de katılımcılar dikkatli veya dikkatsiz cevaplayıcı olarak sınıflanmıştır. İlgili kontrol ifadesine *kesinlikle katılmıyorum* veya *katılmıyorum* dışında cevap verenler dikkatsiz cevaplayıcılar olarak belirlenmiştir.

Anket içinde yer alan LTE ölçeği (Doğan, Özkara ve Doğan, 2016) ifadeleri ile kontrol ifadesine cevaplar 7-point-scale (*1 = Kesinlikle Katılmıyorum, ..., 7 = Kesinlikle Katılıyorum*) aracılığıyla ölçülmüştür. LTE ölçeği, 18 ifade ve 5 boyuttan oluşmaktadır. Boyutlar; eşsizlik (α = .84), pahalılık (α = .77), sembolik anlam (α = .72), ihtiyaç dışı arzulama (α = .64) ve azınlığa ait olma (α = .80) şeklindedir.

Kaliteli veri ölçümü

Kaliteli veri ölçümü amacına yönelik olarak LTE'nin bağımsız değişken, kontrol ifadesinin bağımlı değişken, anket ifadelerinde geçirilen sürenin düzenleyici (moderator) değişken olduğu doğrusal regresyon modeli kurgulanmıştır. LTE değişkeni beş alt boyutun toplam puanı anılarak işlemselleştirilmiştir. LTE'nin beş alt boyutunun toplamının alınmasının sebebi, alt boyutların ilişkili olup yansıtıcı (reflective) ölçüm modeli kapsamında bir araya gelerek LTE'yi oluşturdukları yönünde ölçüm modelinin tasarlanmış olmasıdır. Nitekim bu varsayım bulgular başlığı altında raporlanan açıklayıcı faktör analizi sonuçları ile de desteklenmektedir. Bu regresyon modelinin kurulmasının ardında yatan mantık şöyledir; kurulan regresyon modelindeki β katsayısının anlamlı olmaktan anlamsız olmaya geçtiği bir spesifik düzenleyici değişken değeri var ise bu değer verinin kalitesiz olmaktan kaliteli olmaya geçtiği değer olarak yorumlanabilir. İlgili analiz "Conditional Process Model-1" uygulanarak (Hayes, 2013) ve ayrıca post-hoc nitelikteki "Johnson-Neyman" tekniği aracılığıyla gerçekleştirilmiştir.

Doğrusal regresyon modeli kurulmadan önce regresyon analizinin varsayımı olan normal dağılım varsayımı her bir değişken için sınanmıştır. LTE (çarpıklık = .322, basıklık = -.075), kontrol ifadesi (çarpıklık = .559, basıklık = .147) ve anket ifadelerinde geçirilen süre (çarpıklık = .694, basıklık = .994) değişkenlerinin çarpıklık ve basıklık değerleri -1 ile +1 aralığında (Hair, Black, Babin, Anderson ve Tatham, 1998) tespit edildiği normallik varsayımının ihlal edilmediği çıkarsanmıştır. Ayrıca varyans enflasyon faktör (variance inflation factor, VIF) değeri 1.087 olarak tespit edilmiştir. VIF değerinden anlaşıldığı üzere regresyon modelinin bir diğer varsayımı olan çoklu-doğrusallık (multi-collinearity) sorunu olmadığı da ortaya konmuştur.

Verilerin Analizi

Bu araştırmanın veri toplama aracı olan anket içerisinde yer alan LTE ölçeğinin her bir alt boyutunun geçerliliği Cronbach's alpha katsayısı aracılığıyla test edilmiştir. LTE ölçeğinin yapısal geçerliliğine yönelik olarak ise açıklayıcı faktör analizi uygulanmıştır. Açıklayıcı faktör analizi SPSS 18.0 programı kullanılarak gerçekleştirilmiştir. Kaliteli veri için gerekli olan ifade başına cevaplama süresi hesabı için ise doğrusal regresyon analizi ve ardından 'Conditional Process Model-1' analizi (Hayes, 2013) uygulanmıştır. Bu süreçte 'Johnson-Neyman' tekniğinden faydalanılarak spesifik olarak ifade başına cevaplama süresi hesaplaması gerçekleştirilmiştir. Son olarak dikkatli ve dikkatsiz cevaplayıcılar arası LTE ölçeği alt boyutlarının α değerleri karşılaştırılması ise Feldt (1969) tarafından geliştirilen hesaplama yöntemi ile hesaplanmıştır.

BULGULAR

IMC'ye verilen cevaplar doğrultusunda, sınıf ortamında gerçekleştirilen anket uygulaması neticesinde öğrencilerin 104 (%67.5)'ünün dikkatsiz cevaplayıcı olduğu, çevrimiçi (online) ortamda gerçekleştirilen anket uygulaması neticesinde ise cevaplayıcıların 139 (%95.3)'unun dikkatsiz cevaplayıcı olduğu tespit edilmiştir. Böylece IMC'ye verilen cevaplar doğrultusunda, toplanan toplam 302 veriden 245 (%81.1)'inin dikkatsiz cevaplayıcılardan elde edilmiş olduğu tespit edilmiştir. Ayrıca dikkatsiz cevaplayıcı oranının iki farklı anket toplama yöntemi arasında anlamlı şekilde farklılaştığı tespit edilmiştir (*Phi* = -.35, *p* = .001).

Dikkatsiz cevaplayıcıları belirlemenin bir diğer yöntemi olarak anket ifadeleri arasında konumlandırılan kontrol ifadesine ("Doğa üstü güçlere sahip bir dinozorum") verilen cevaplar doğrultusunda, sınıf ortamında gerçekleştirilen anket uygulaması neticesinde öğrencilerin 44 (%28.6)'ünün dikkatsiz cevaplayıcı olduğu, çevrimiçi (online) ortamda gerçekleştirilen anket uygulaması neticesinde ise cevaplayıcıların 21 (%14.2)'inin dikkatsiz cevaplayıcı olduğu tespit edilmiştir. Böylece kontrol ifadesine verilen cevaplar doğrultusunda, toplanan toplam 302 veriden 65 (%21.5)'inin dikkatsiz cevaplayıcılardan elde edilmiş olduğu tespit edilmiştir.

Yukarıda görüldüğü üzere, iki farklı dikkatsiz cevaplayıcı tespit yöntemi ışığında belirlenen dikkatsiz cevaplayıcı sayılarının tutarsız olduğu görülmektedir. Nitekim, IMC ve kontrol ifadesi yöntemlerinin tutarlı şekilde dikkatsiz cevaplayıcı sayısını belirleme performansı sergilemediği de phi-coefficient hesaplaması neticesinde desteklenmiştir (*Phi* = .07, *p* = .242). Bu bulgu, anket içerisinde yer alan ölçek ifadelerine ve kontrol sorusuna dikkatli şekilde cevap vermiş fakat girişte yer alan IMC sayfasında dikkat sergilemeyen cevaplayıcılar olabileceğine işaret etmektedir. Bu gerekçeden dolayı ilerleyen analiz sürecinde dikkatsiz-dikkatli cevaplayıcı ayrımı anket içerisinde yer alan kontrol ifadesi (Doğa üstü güçlere sahip bir dinozorum) doğrultusunda yapılmıştır.

Çevrimiçi (online) ortamda gerçekleştirilmiş olan anket uygulamasına katılan cevaplayıcıların IMC sayfasında geçirdikleri süre 0.82 ile 122.53 saniye arasında değişmektedir (*M* = 13.22, *SD* = 25.45). Anket ifadelerinin yer aldığı sayfada cevaplayıcıların geçirdikleri süre ise 4.61 ile 444.23 saniye arasında değişmektedir (*M* = 118.84, *SD* = 86.70). Ayrıca IMC sayfasında geçirilen süre ile anket ifadeleri sayfasında geçirilen süre arasında pozitif anlamlı ilişkili olduğu görülmüştür (*r* = .27, *p* = .001).

Tüm veri kapsamında, anket içinde yer alan LTE ölçeği ifadelerine açıklayıcı faktör analizi uygulanmıştır. Faktör analizi sonuçları (Tablo 1), eşsizlik boyutundan bir ifade (E1) ile pahalılık boyutundan bir ifadenin (P2) ait oldukları faktörlere yeterince yüklenmediklerini ortaya koymuştur. Bu durumunun sebebinin cevaplayıcılar arasında önemli oranda dikkatsiz cevaplayıcı olduğu düşünülerek açıklayıcı faktör analizi testi dikkatsiz cevaplayıcılar ve dikkatli cevaplayıcılar kapsamında olmak üzere iki ayrı şekilde de gerçekleştirilmiştir. Dikkatli cevaplayıcılar kapsamında gerçekleştirilen açıklayıcı faktör analizi sonuçları, beklendiği üzere, beş faktörlü bir çözüm ortaya koymuştur (Tablo 1). Üstelik her ifadenin ait olduğu faktöre en az .70 düzeyinde yüklendiği görülmüştür. Fakat dikkatsiz cevaplayıcılar kapsamında gerçekleştirilen açıklayıcı faktör analizi sonuçlarında ise eşsizlik boyutundan bir ifadenin (E1), pahalılık boyutundan bir ifadenin (P2) ve azınlığa ait olma boyutundan bir ifadenin (A2) ait olduğu boyutlara yeterince yüklenmediği görülmüştür. Ayrıca dikkatsiz cevaplayıcılar kapsamında gerçekleştirilen açıklayıcı faktör analizi sonuçları dört boyutlu bir çözüm önermiştir (Tablo 1). Bu bulgular, özellikle dikkatsiz cevaplayıcılar kapsamında LTE ölçeğinin beklendiği gibi çalışmadığı ve bu yüzden veri bir araya getirildiğinde iki ifadenin ait oldukları faktörlere yeterince yüklenemedikleri yönünde çıkarım sağlamaktadır. Dolayısıyla orijinalinde olduğu üzere LTE ölçeği beş boyuta sahip olacak şekilde işlemselleştirilmiştir.

Tablo 1. Açıklayıcı Faktör Analizi Sonuçları

LTE ifadeleri	Dikkatsiz cevaplayıcılar		Dikkatli cevaplayıcılar		Tüm veri	
E = eşsizlik, P = pahalılık, SA = sembolik anlam, İDA = ihtiyaç dışı arzulama, AAO = azınlığa ait olma	Faktör yükü	Model sonuç	Faktör yükü	Model sonuç	Faktör yükü	Model sonuç
E1. Bir ürünü/hizmeti diğerlerinden farklı olduğu için satın alırım.	.471		.863		.437	
E2. Satın alma kararlarımda ilgili ürünün/hizmetin benzersiz özelliklere sahip olmasını gözetirim.	.775		.872		.860	
E3. Diğer ürünlerden/hizmetlerden farklı özelliklere sahip ürünlere karşı ilgi duyarım.	.778		.814		.816	
E4. Satın aldığım bir ürünün/hizmetin yalnızca bana özel olmasını arzularım.	.709		.852		.798	
P1. Pahalı ürünleri/hizmetleri satın almaktan mutluluk duyarım.	.708		.798		.884	
P2. Ucuz ürünleri/hizmetleri bulmak benim için çok önemli değildir	.419		.880		.480	
P3. Ucuz ürünleri/hizmetleri satın almayı tercih etmem.	.797		.855		.774	
P4. Pahalı bir ürünü/hizmeti ucuz bir ürüne/hizmete tercih ederim.	.798		.898		.744	
SA1. Bir ürünün/hizmetin fonksiyonel özelliklerinden ziyade sembolik özelliklerini önemserim.	.710	KMO = .665, p = .001, özdeğer = 9.291, açıklanan varyans = %61.25	.802	KMO = .81, p = .001, özdeğer = 11.232, açıklanan varyans = %77.54	.865	KMO = .80, p = .001, özdeğer = 10.828, açıklanan varyans = %67.67
SA2. İçinde yaşadığım toplumda lüks sembolik anlama sahip ürün/hizmeti satın alırım.	.702		.799		.846	
SA3. Ürünleri/hizmetleri satın alırken etrafımdaki insanlar için ne ifade ettiğini göz önünde bulundururum.	.785		.780		.796	
SA4. Bir ürünün/hizmetin başkaları için ne ifade ettiği benim için önemlidir.	.721		.864		.736	
İDA1. Hiç ihtiyacım olmadığı halde sadece istediğim için alışveriş yaparım.	.723		.803		.897	
İDA2. Arzuladığım bir ürüne/hizmete ihtiyacımın olup olmadığını sorgulamam.	.714		.847		.866	
İDA3. Fiziksel ihtiyaç duymaktan daha çok duygusal ihtiyaç duyduğum ürünleri/hizmetleri satın alırım.	.707		.805		.739	
AAO1. Birçok kişinin satın alabileceği bir ürünü/hizmeti satın almaktan hoşlanmam.	.734		.834		.876	
AAO2. Sahip olduğum bir ürüne/hizmete diğer birçok insanın sahip olması beni rahatsız eder.	.434		.799		.844	
AAO3. Satın aldığım ürünler/hizmetler aracılığıyla kendimi azınlık bir gruba ait hissetmekten hoşlanırım.	.701		.765		.793	

Bir diğer yandan, LTE ölçeğinin beş boyutu olan; 'eşsizlik', 'pahalılık', 'sembolik anlam', 'ihtiyaç dışı arzulama' ve 'azınlığa ait olma' boyutlarının içsel tutarlılık değerleri olan Cronbach's alpha (α) katsayıları dikkatli ve dikkatsiz cevaplayıcılar için ayrı ayrı hesaplanmış ve karşılaştırma yapılmıştır. Her bir boyutun α değerinin dikkatsiz cevaplayıcılar kapsamında daha yüksek olduğu tespit edilmiştir (Ek-2). Ancak α değerlerinin dikkatsiz ve dikkatli cevaplayıcılar arasında istatistiksel olarak anlamlı şekilde farklılaşmadığı da Feldt (1969) testi neticesinde gözlemlenmiştir (W değeri = 0.976, p = .477). Her ne kadar α değerleri arası istatistiksel anlamlı farklılık olmasa da, α değerinin dikkatsiz cevaplayıcılar kapsamında görece daha yüksek olması, α değerinin yüksek olmasının her

zaman verinin kaliteli olduğuna dair bir çıkarım sağlamadığı ve hatta yanıltıcı olabildiği çıkarımını sağlamaktadır. Bu durum Cronbach's alpha yanılgısı'dır.

Son olarak, kaliteli veri elde edebilmek için cevaplayıcıların ifade başına en az kaç saniye harcamış olmasını tespit edebilmek amacıyla "Conditional Process Model-1" analizi ve sonrasında "Johnson-Neyman" teknik testi uygulanmıştır. Test, %95 güven aralığı ve 10,000 örneklem sayısı bootstrapping yaklaşımıyla gerçekleştirilmiştir. Modelde; LTE bağımsız değişken, kontrol ifadesi bağımlı değişken, anket ifadelerini cevaplama süresi düzenleyici değişken olarak kurgulanmıştır. Analiz sonuçları; anket ifadelerini cevaplama süresi 361.73 saniyeye kadar LTE'nin kontrol ifadesini anlamlı etkilediği, bu cevaplama süresinden sonra anlamsız etkilediğini göstermiştir. Bu kapsamda, mantık dışı etkinin anlamlı olması verinin kalitesizliğine yönelik bir işaret iken, anlamsız olması ise verinin kaliteli olduğuna yönelik bir işaret olabilir. Toplam 18 ölçek ifadesine cevaplayıcıların cevap verdikleri göz önünde bulundurulduğunda (361.73 / 18 = 20.09) ifade başına yaklaşık 20 saniyeden daha fazla harcayanların kaliteli veri sağlamış olabileceği yönde bir varsayım söz konusu olur.

SONUÇLAR ve TARTIŞMA

Bu araştırmada, çevrimiçi (online) ve sınıf ortamında uygulanmış olan iki farklı anket verisi kapsamında anket cevaplayıcıları arasında ne oranda dikkatsiz cevaplayıcı olduğu sorusuna cevap aranmıştır. Ayrıca kaliteli veri elde edebilmek için çevrimiçi (online) ortamda cevaplayıcıların anket başına ne kadar süre harcamış olması gerektiği sorusuna da cevap aranmıştır.

Toplam 302 Türk cevaplayıcı kapsamında gerçekleştirilmiş olan bu araştırma neticesinde, anket cevaplayıcıları içindeki dikkatsiz cevaplayıcı oranının %14.2 ile %95.3 arasında değiştiği yönünde bulgulara ulaşılmıştır. Bu oranın, Batılı kültürlerde gerçekleştirilmiş olan geçmiş çalışmalar neticesinde elde edilen dikkatsiz cevaplayıcı oranlarına kıyasla oldukça yüksek olduğu görülmektedir (Anduiza ve Galais, 2017; Berinsky, Margolis ve Sances, 2014; Hauser ve Schwarz, 2015; Meade ve Craig, 2012). Ayrıca bu araştırma neticesinde dikkatsiz cevaplayıcı oranının çevrimiçi (online) ortamda toplanmış olan anket verisinde daha yüksek olduğu yönünde bulgulara da ulaşılmıştır. Bu bulgular, Türk kültürü kapsamında çevrimiçi (online) ve sınıf ortamında uygulanan anketlerin veri kalitesine ilişkin önlemler alınması gerektiğine işaret etmektedir. Dikkatsiz cevaplayıcı oranının %5 gibi düşük düzeyde olmasının bile araştırma sonuçlarının yanlı olmasına sebep olabildiği (Huang, Liu ve Bowling, 2015) göz önünde bulundurulduğunda, Türk ulusal literatüründe bulguları anket verisine dayanan araştırmaların sonuçlarının yanlı sonuçlar içerme riski ile karşı karşıya olduğu söylenebilir.

Öte yandan, anket uygulamalarında dikkatsiz cevaplayıcıları tespit etmeye yönelik olarak anket cevaplama süresi açısından bir metodolojik araştırma uluslararası literatürde yer almamaktadır. Sadece Huang, Curran, Keeney, Poposki ve Deshon (2012) tarafından ifade/soru başına iki saniye önerisinde bulunulmuştur. Bu araştırma neticesinde ise, kaliteli veri elde edebilmek için cevaplayıcının ifade başına yaklaşık 21.28 saniye harcamış olması yönündeki varsayımsal bulgu Johnson-Neyman teknik uygulanarak ulaşılmıştır. Böylece kaliteli veri için ifade başına harcanması gereken sürenin tespitine yönelik keşifsel bir bulgu literatüre kazandırılmıştır.

Ayrıca IMC sayfasında geçirilen süre ile anket içerisinde yer alan ifadelerde geçirilen süre arasında güçlü bir korelasyon olmaması (r = .27, p = .001) dikkatsiz cevaplayıcı belirleme yaklaşımına bağlı olarak farklı sonuçlara ulaşılabileceği çıkarımını sağlamaktadır. Bir diğer deyişle, IMC sayfasında az zaman geçiren fakat dikkatli şekilde anket içinde yer alan ifadelere cevap veren katılımcılar bulunabilir. Bu yüzden gelecek araştırmalarda dikkatsiz cevaplayıcıların sınıflandırılmasına yönelik olarak IMC'de geçirilen zaman ile anket içindeki ifadelerde geçirilen zaman eş anlı şekilde gözönünde bulundurularak yeni bir dikkatsiz cevaplayıcı sınıflandırılması geliştirilebilir.

Ulaşılan en ilginç bulgu ise LTE ölçek boyutlarının α katsayılarının dikkatsiz cevaplayıcılar kapsamında daha yüksek olduğu bulgusudur. Bu durum, α katsayısının verinin metodolojik açıdan kalitesine ilişkin araştırmacıları yanıltabileceğine dikkat çekmektedir. Bu durum 'Cronbach's alpha yanılgısı'dır. Bu durumun sebebinin, dikkatsiz cevaplayıcıların anket ifadelerine tutarlı şekilde fakat

dikkatsizce cevaplayabilmesinin olduğu düşünülmektedir. Bu noktada, uluslararası metodoloji literatüründe yea-saying (Bachman ve O'Malley, 1984) olarak bilinen, ifadelerin çoğuna aynı cevabı verme eğiliminin dikkatsiz cevaplayıcılar arasında yaygın olabileceği fikri göz önünde bulundurulmalıdır.

Bu araştırmanın en önemli kısıtı, deneysel bir araştırma olmamasıdır. Gelecek araştırmalarda, cevaplayıcıları dikkatsiz niteliğe kavuşturacak manipülasyonlar geliştirilip deneysel araştırma kapsamında dikkatsiz ve dikkatli cevaplayıcılar arasında anket cevaplama tarzları açısından ne tür farklılıklar olduğu sorusuna cevap aranmalıdır. Öte yandan anket içerisinde yer alan ölçek ifadelerinde harcanan sürenin sadece dikkatsizlik düzeyi ile ilgili değil cevaplayıcının yaşı ve okuma hızı ile de ilgili olabileceği göz önünde bulundurulmalıdır. Fakat bu araştırma kapsamında cevaplayıcının yaşı ve okuma hızına ilişkin bir ölçüm yapılmamıştır. Bu nedenle ilgili faktörler analiz sürecinde kontrol altında tutulmamıştır. Gelecek araştırmalarda cevaplayıcının yaşının ve okuma hızının kontrol altında tutularak ifade başına cevaplama süresi hesaplaması yapılması önerilmektedir.

KAYNAKÇA

Anduiza, E., & Galais, C. (2017). Answering without reading: IMCs and strong satisficing in online surveys. *International Journal of Public Opinion Research*, basım aşamasında.

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly*, *48*(2), 491-509.

Beach, D. A. (1989). Identifying the random responder. *The Journal of psychology*, *123*(1), 101-103.

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered survey. *American Journal of Political Science*, *58*(3), 739-753.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4-19.

Doğan, V., Özkara, B. Y., & Doğan, M. (2016). *Luxury consumption tendency: Conceptualization, scale development, and validation*. Paper presented at the 2016 annual meeting of the American Marketing Association, Atlanta, GA.

Feldt, L. S. (1969). A test of the hypothesis that cronbach's alpha or kuder-richardson coefficent twenty is the same for two tests. *Psychometrika*, *34*(3), 363-373.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*(2), 93-104.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (1998). *Multivariate data analysis*. Upper Saddle River, New Jersey: Prentice Hall.

Hauser, D. J., & Schwarz, N. (2015). It's a trap! Instructional manipulation checks prompt systematic thinking on 'tricky' tasks. *SAGE Open*, *5*(2), 1-6.

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

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & Deshon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99-114.

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*(3), 299-311.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological methods*, *17*(3), 437-455.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use?. *Journal of Research in Personality*, *63*, 1-11.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867-872.

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*(3), 186-191.

EXTENDED ABSTRACT

Introduction

Much of the knowledge in the social and behavioral sciences based on survey data. The central assumption of survey research is that respondents are reporting themselves in a careful way. Researchers have begun to question carelessness/inattentiveness levels of respondents. A great deal of research demonstrated that rate of careleness/inattentive respondents ranged from 3.3% to 63% (Anduiza & Galais, 2016; Berinsky et al., 2014; Hauser & Schwarz, 2015; Meade & Craig, 2012; Oppenheimer et al., 2009), implicating that research projects based on survey design are under the threat of careleness/inattentive respondents problem.

Researchers have not stayed silent to the threat of careleness/inattentive respondents and developed some methods to detect careleness/inattentive respondents. Some of these methods are Instructional Manipulation Check (IMC), bogus-item, response time, long-string analysis, even-odd index, Mahalanobis distance, semantic antonyms/synonyms index, pscyhometric antonyms index, and explicit instructed response item (Curran, 2016; Niessen et al., 2016).

With the advent of crowdsourcing platforms, collecting survey data have got easier. Furthermore, participants recruited from crowdsourcing platforms are easily participate online surveys. Thanks to online survey platforms (e.g., Qualtrics), reseachers are currently able to measure respondents' survey completion time. This development provides an advantage to compare survey completion time of careless respondents and careful respondents. However, to the best of our knowledge, ideal survey response completion time has yet to be examined in the current paper. In addition, all the past research regarding to identification of the rate of careless/inattentive respondents have conducted in Western and developed countries/cultures. The current papar is also first attemp to provide descriptive statistics about the rate of careless/inattentive respondents among Turkish culture.

Method

This paper uses two different cross-sectional survey data. First data comes from 154 Turkish undergradute students and second data comes from 148 adults. Former data were collected via face-to-face survey in the classroom, whereas latter data were collected via online-survey. Also, online participants were recruited from akademikpersonel.org's Facebook page.

Results and Discussion

Results of the current paper showed that face-to-face survey data includes %67.5 careless/inattentive respondents, whereas online survey data includes %95.3 careless/inattentive respondents. On the other hand, according online survey data, respondents' survey completion time ranged from 4.61 sec to 444.243 sec.

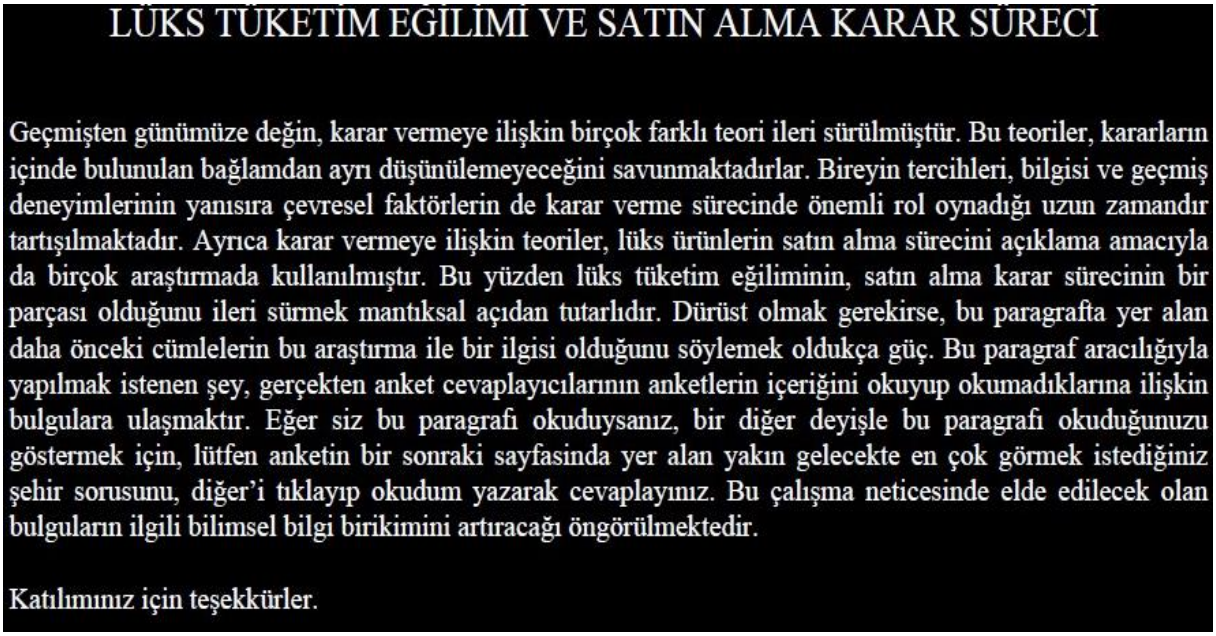
One of the interesting finding of the current study is that careless respondents had better internal consistency performace than did careful respondents. Put another way, careless respondents inflate Cronbach's Alpha coefficient. This problem is coined as 'delusion of Cronbach's Alpha'.

Floodlight analysis also conducted to identify specific cut-off response time value that separates careless respondents from careful respondents. Results suggested that respondents who spent less than 21.28 sec per-item should be classified as careless respondents. This classification is the first systematic initiation toward response time.

Results of the current paper implicates that the rate of careless respondents among Turkish respondents is quite high compared to the rate of careless respondents among Western cultures. Crucially, systematic approach to calculate ideal survey response time was developed for the first time.

Ekler

Ek-1: Dikkatsiz Cevaplayıcı Tespitine Yönelik Tasarlanan Instructional Manipulation Check



LÜKS TÜKETİM EĞİLİMİ VE SATIN ALMA KARAR SÜRECİ

Geçmişten günümüze değin, karar vermeye ilişkin birçok farklı teori ileri sürülmüştür. Bu teoriler, kararların içinde bulunulan bağlamdan ayrı düşünülemeyeceğini savunmaktadırlar. Bireyin tercihleri, bilgisi ve geçmiş deneyimlerinin yanısıra çevresel faktörlerin de karar verme sürecinde önemli rol oynadığı uzun zamandır tartışılmaktadır. Ayrıca karar vermeye ilişkin teoriler, lüks ürünlerin satın alma sürecini açıklama amacıyla da birçok araştırmada kullanılmıştır. Bu yüzden lüks tüketim eğiliminin, satın alma karar sürecinin bir parçası olduğunu ileri sürmek mantıksal açıdan tutarlıdır. Dürüst olmak gerekirse, bu paragrafta yer alan daha önceki cümlelerin bu araştırma ile bir ilgisi olduğunu söylemek oldukça güç. Bu paragraf aracılığıyla yapılmak istenen şey, gerçekten anket cevaplayıcılarının anketlerin içeriğini okuyup okumadıklarına ilişkin bulgulara ulaşmaktır. Eğer siz bu paragrafı okuduysanız, bir diğer deyişle bu paragrafı okuduğunuzu göstermek için, lütfen anketin bir sonraki sayfasinda yer alan yakın gelecekte en çok görmek istediğiniz şehir sorusunu, diğer'i tıklayıp okudum yazarak cevaplayınız. Bu çalışma neticesinde elde edilecek olan bulguların ilgili bilimsel bilgi birikimini artıracağı öngörülmektedir.

Katılımınız için teşekkürler.

Ek-2: LTE Ölçeğinin Boyutlarının α Değerleri

		Eşsizlik (α = .84)	**Pahalılık (α = .77)**	**Sembolik anlam (α = .72)**	**İhtiyaç dışı arzulama (α = .64)**	**Azınlığa ait olma (α = .80)**
IMC	Dikkatsiz cevaplayıcılar	.85	.79	.72	.66	.89
	Dikkatli cevaplayıcılar	.75	.68	.71	.62	.85
Kontrol ifadesi	Dikkatsiz cevaplayıcılar	.86	.79	.71	.67	.84
	Dikkatli cevaplayıcılar	.84	.76	.70	.63	.81

Parantez içinde yer alan α değerleri tüm veri seti kapsamında hesaplanmıştır.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(3);354-372

Çok Kategorili Parametrik ve Parametrik Olmayan Madde Tepki Kuramı Modellerinin Karşılaştırılması*

Comparison of Polytomous Parametric and Nonparametric Item Response Theory Models

Özge BIKMAZ BİLGEN ** Nuri DOĞAN ***

Öz

Bu araştırmanın amacı çok kategorili maddeler için Parametrik Madde Tepki Kuramı (PMTK) kapsamındaki Aşamalı Tepki Modeli (ATM) ve Parametrik olmayan Madde Tepki Kuramı (PoMTK) kapsamındaki Monoton Homojenlik Modeli (MHM) ile yapılan kestirimlere örneklem büyüklüğü, örneklem dağılımı, testte yer alan madde sayısı, testte yer alan maddelerin yanıt kategorisi sayıları bağımsız değişkenlerinin etkilerini incelemektir. Bu amaca ulaşabilmek için araştırma; örneklem büyüklüğü, örneklem dağılımı, madde sayısı, maddenin kategori sayısı değişkenleri ile belirlenen 192 simülasyon koşulu desenlenen temel bir çalışma olarak gerçekleştirilmiştir. Örneklem büyüklüğü (N=100, 250, 500, 1000), örneklem dağılımı (normal dağılım, çarpıklık katsayısı -1,0 olan dağılım), madde sayısı (10, 20, 40, 80), maddenin yanıt kategorisi sayısı (3, 5, 7) koşulları için ATM ve MHM ile yapılan kestirimler sırasıyla model veri uyumları, güvenirlik değerleri, madde parametrelerinin standart hataları hesaplanarak incelenmiştir. Araştırma sonucunda ATM'de model veri uyumu hesaplanırken değerlerin değişken artışından etkilenmesi, tek başına yorumlanamaması bu değerlerin karşılaştırılması ve genellenmesini zorlaştırmaktadır. MHM'de model veri uyumunun pratik olarak hesaplanması, başka bir değere ihtiyaç duyulmadan tek başına yorumlanması ATM'ye göre üstünlük sağlamaktadır. Diğer bir araştırma sonucu güvenirlik değerlerinin iki model için benzer sonuç vermesidir. MHM için hesaplanan parametrelerin, küçük örneklem ve kısa test koşullarında standart hataları ATM kestirimlerine göre oldukça düşüktür ve MHM için hesaplanan parametrelerin standart hataları tüm koşullarda birbirine yakın değer almıştır.

Anahtar Kelimeler: Parametrik olmayan madde tepki kuramı, aşamalı tepki modeli, monoton homojenlik modeli

Abstract

This research aimed to identify the effects of independent variables as sample size, sample distribution, the number of items in the test, and the number of response categories of items in the test on the estimations of Graded Response Model (GRM) under Parametric Item Response Theory (PIRT) and by Monotone Homogeneity Model (MHM) under Non-Parametric Item Response Theory (NIRT) for polytomously scored items. To achieve this aim, the research was performed as a fundamental study in which 192 simulation conditions were designed by the combination of sample size, sample distribution, the number of items, and the number of categories of items. Estimates by GRM and MHM were examined under different levels of sample size (N= 100, 250, 500, 1000), sample distribution (normal, skewed), the number of items (10, 20, 40, 80), and the number of categories of items (3, 5, 7) conditions, by respectively calculating model-data fit, reliability values, standart errors of parameters. As a result of the research, it was found that since the values used to evaluate model-data fit were influenced by the increase of variable while calculating model-data fit and since they can not be interpreted alone, it is difficult to compare and generalize the results. The practical calculation of model data fit, which can be interpreted without the need for another value, in MHM provides superiority over GRM. Another research result is that the reliability values give similar results for both models. The

*Bu çalışma, ilk yazarın, ikinci yazar danışmanlığında tamamladığı "Parametrik ve Parametrik Olmayan Madde Tepki Kuramı Modellerinin Farklı Örneklemler ve Test Uzunluğunda Karşılaştırılması" isimli doktora tezinden üretilmiştir.

** Arş. Gör. Dr., Adnan Menderes Üniversitesi, Eğitim Fakültesi, Aydın-Türkiye, ozgebikmaz@adu.edu.tr, ORCID ID: orcid.org/0000-0003-2219-2026

***Doç. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, e-posta:, nurid@hacettepe.edu.tr, ORCID ID: orcid.org/0000-0001-6274-2016

DOI: 10.21031/epod.346650
Geliş Tarihi: 13.10.2017
Kabul Tarihi: 06.11.2017

standard errors of the MHM parameter estimates is lower than the GRM estimates under small sample and few items conditions and the standard errors of the MHM parameter estimates are close to each other in all conditions.

Keywords: Nonparametric item response theory, graded response model, monotone homogeneity model

GİRİŞ

Eğitim alanında bireyler hakkında bilgi toplamak için başarı, ilgi, tutum vb. psikolojik yapılar ölçülmektedir. Psikolojik yapıların doğrudan gözlenememesi ve değişkenlik göstermesi nedeniyle ölçme alanında test kuramları geliştirilmiştir (Hambleton & Jones, 1993). Bu kuramlardan Madde Tepki Kuramı (MTK) günümüzde, araştırmalarda sıklıkla kullanılan bir test kuramıdır (De Ayala, 2009; Ostini & Neing, 2006).

Madde Tepki Kuramı (MTK); Klasik Test Kuramı (KTK)'nın bir uzantısı olmakla birlikte genellikle KTK'ya kıyasla daha modern ve üstün bir alternatif olarak sunulmaktadır (Embretson & Reise, 2000; Nering & Ostini, 2010). KTK'dan farklı olarak MTK örneklemden bağımsız madde ölçeklemesini ve maddelerden bağımsız yetenek kestirimini matematiksel modellerle olanaklı hâle getirme iddiasındadır (De Ayala, 2009).

MTK'da bireylerin testteki maddelere verdikleri yanıtın altında o test ile ölçülmeye çalışılan yetenekleri (θ) olduğu kabul edilmektedir. Yetenek gözlenemeyen bir özelliktir. MTK'da gözlenemeyen özelliğin ancak gözlenebilir olan teste yanıt verme davranışı ile ölçülebildiği düşünülmektedir. Yani bireyin yetenek düzeyi, bireyin maddelere verdiği yanıtlar ile kestirilmeye çalışılır. Bu kestirimde matematiksel fonksiyonlardan yararlanılır. Bu doğrultuda bireylerin yeteneği ile testteki maddelere verdikleri yanıtlar arasındaki ilişki incelenmektedir. Bu ilişki, bireyin belli bir maddeye tepki verme olasılığı ile bireylerin test maddelerinin altında yatan yetenek üzerindeki konumu arasında matematiksel bir bağlantı olan matematiksel fonksiyonlarla kurulmaktadır (Ostini & Neing, 2006).

MTK'da bireyin gözlenemeyen özelliği ile gözlenen özelliği arasındaki matematiksel fonksiyonu veren bir madde karakteristik fonksiyonu (MKF) ve fonksiyona dayalı olarak elde edilen madde karakteristik eğrisi (MKE) bulunmaktadır. MKE madde puanının yetenek üzerindeki regresyonudur ve MKE'nin şekli, yetenek düzeyindeki değişim ile doğru cevaplama olasılığı arasındaki ilişkiyi ortaya koymaktadır. MKE madde parametrelerine bağlıdır ki elde edilen parametreler eğrinin şeklini belirlemektedir (Doğan, 2002). MKE eğrisinin şekline dayalı olarak MTK modelleri parametrik ve parametrik olmayan modeller olarak iki ana kategoride incelenmektedir (Sijtsma ve Molenaar, 2002). Parametrik MTK modellerinde MKE'nin normal veya lojistik olması beklenirken, parametrik olmayan MTK modellerinde eğrinin şekli belirli bir formla sınırlı değildir (Sijtsma, 2005).

Alan yazın incelendiğinde MTK çerçevesinde yapılan çalışmalarda daha çok parametrik modeller (PMTK) üzerinde odaklanıldığı görülmektedir (Molenaar, 2001). Bu durumun gerekçeleri arasında parametrik modellerin daha önce önerilmesi, kestirimlerde kullanılan istatistiksel yazılımlarının daha fazla olması sayılabilir (Sjistma, Emon, Bouwmeester, Nyklicek ve Roorda, 2008; Junker & Sijtsma, 2001).

MTK temel birtakım varsayımlara sahiptir. Temel varsayımlardan biri maddenin tek bir değişkeni ölçmesi (Hambleton ve Swaminathan, 1985), bir testteki tüm maddelerin tek bir özellikle açıklanması (Sijtsma, 2005) anlamına gelen tek boyutuluktur. MTK'nın diğer bir varsayımı yerel bağımsızlıktır. Varsayım bireylerin maddeyi yanıtlarken verdikleri tepkilerin birbirinden bağımsız olduğunu temel almaktadır. Kuramlardan elde edilen sonuçların nitelikli olması kuramların varsayımlarının karşılanmasına bağlıdır. MTK çalışmalarında tekboyutluluk, yerel bağımsızlık varsayımlarının karşılanmasının yanısıra parametrik modellerle çalışılıyorsa PMTK ile çalışma sonuçlarının nitelikli olması için birtakım gerekliliklerin olduğu belirtilmektedir (Molenaar, 2001).

Bu gerekliliklerden biri PMTK ile çalışmada geniş örnekleme ihtiyaç duyulmasıdır (Demars, 2010). PMTK ile geniş örneklem ile çalışılmasının önerilmesi MTK'nın erken dönem çalışmalarına dek

uzanmaktadır (Hulin, Lissak & Drasgow, 1982; Thissen & Wainer, 1982). Geniş birey kitlesiyle çalışan araştırma sonuçlarının beklenen düzeyde olması (Zenisk, Hambleton & Sireci, 2002) diğer yandan küçük örneklem üzerinden elde edilen sonuçların istenen nitelikte olmaması (De Ayala, 2009) araştırmacıların önerilerini büyük örneklemin gerekliliği noktasında birleştirmektedir. Parametrik modellerle MTK çalışmasında iyi sonuçlar elde edilmesi için önerilen bir başka koşul geniş madde havuzu kullanımıdır (Demars, 2010). Çalışmanın güvenirliğinin ve geçerliğinin istenen nitelikte olmasında madde sayısı artırılmasının olumlu etkiye sahip olduğu bilinmektedir. Madde sayısı değişiminin sonuçların doğruluğu üzerindeki etkisinin incelendiği çalışmalarda daha çok maddeli testlerle çalışmanın sonuçların doğruluğunu arttırdığına (ölçmenin standart hatasının düştüğüne) yönelik bulgular elde edilmiştir (Ankenmann & Stone, 1992; Hulin, Lissak & Drasgow, 1982). Parametrik modellerle MTK çalışmalarında karşılanması beklenen diğer bir nokta, madde yapısının kullanılan modelle uyum sağlaması gerekliliğidir. Parametrik MTK modelleri genel olarak iki ve çok kategorili modeller olarak iki ana başlık altında incelenmektedir (Thissen & Steinberg, 1986). Maddenin yanıt kategori sayısının değişimi farklı model kullanımını gerektirmektedir. Çok kategorili yanıtlanan maddelerin iki kategorili yanıtlananlara göre daha kapsamlı bilgi verdiği düşünülerek (Ostini & Neing, 2006) çalışmalarda sıklıkla beş kategorili maddeler kullanılmaktadır. Yapılan çalışmalar model veri uyumu, güvenirlik gibi niteliklerin sağlanmasında kategori sayısının etkili olduğuna işaret etmektedir (Zenisky, Hambleton & Sireci, 2002). Belirtilen PMTK'ya ait güçlü varsayımların ve zahmetli gerekliliklerin olmasının yarattığı problemlerin çözümünde PoMTK modelleri, parametrik modellere alternatif olarak sunulmuştur (Mokken, 1971; Molenaar, 2001; Sijtsma, 1998; Stout, 1987).

PoMTK modellerinde tek boyutluluk parametrik modellerde olduğu gibi temel varsayımıdır. Bunlara ek olarak PoMTK modelleri çalışmalarında önemli bir yeri olan diğer bir varsayım monotonluktur. Monotonluk PMTK için de geçerli olan MKE veya çok kategorili yanıtlanan maddeler için madde adımı yanıt eğrisinin monoton azalmayan bir yapı sergilemesi anlamına gelmektedir. PoMTK modellerini, PMTK'dan ayıran nokta MKE'lerin monoton artışının (azalmayışının) belli şekle dayanmamasıdır. Yani PMTK'nın lojistik ya da normal eğri şeklinde yapı göstermesi beklenirken bu modelde monoton azalmayan yapının belli bir şekle bağlı olması beklenmemektedir (Sijtsma & Molenaar, 2002). Hatta eğrinin şekli oldukça düzensiz, kesikli olabilmektedir. Yani $P(\theta)$ ve θ arasındaki ilişki monoton olarak azalmayan ise herhangi bir fonksiyonel formda olması kabul edilebilir.

Junker ve Sijtsma (2001) PoMTK modellerinin PMTK modellerinin daha iyi anlaşılmasını sağladığı, parametrik modellerin zayıf model-veri uyumu sağladığı durumlarda model veri uyumu sağlamada daha esnek bir çerçeve sunduğu, daha az sayıda madde ve birey ile çalışıldığı durumlarda model veri uyumu, güvenirlik, geçerlik konularında daha iyi sonuç sağladığı gerekçeleriyle daha kullanışlı olduğunu öne sürmüştür (Emons, 2008). PoMTK modelleri daha esnek varsayımlara sahiptir (Molenaar, 2001; Emons, 2008). Araştırmacılar bu modellerin en önemli avantajının belli bir madde yanıt fonksiyonuna (lojistik ya da normal gibi) gerek duyulmaması olduğu noktasında birleşmektedir (Sijtsma, 2005). Bu yönüyle modeller madde yanıt fonksiyonun maddeye özgü olmasını sağlayarak madde yapısının özgün haliyle incelenmesine fırsat tanımaktadır. Oysa PMTK'da maddenin modelle birlikte belirlenen madde karakteristik eğrisine (örneğin lojistik) uyup uymadığına bakılarak kestirim yapılmaktadır. Bu yönüyle PoMTK'nın keşfedici; PMTK'nın ise doğrulayıcı bir yapı sergilediği belirtilmektedir (Junker ve Sjitsma, 2001). Parametrik ve parametrik olmayan madde tepki kuramı modelleri sırasıyla açıklanmıştır.

Çok Kategorili Maddeler İçin Parametrik Madde Tepki Kuramı Modelleri

Parametrik modeller kendi içinde iki ve çok kategorili maddeler için geliştirilen modeller olarak ikiye ayrılmaktadır. İki kategorili modellerde, içerdikleri parametre sayısına bağlı olarak adlandırılma yoluna gidilmektedir. Alan yazında 1 parametreli model, 2 parametreli model, 3 parametreli ve 4 parametreli modellerin olduğu görülmektedir (Embretson & Reise; 2000). Çok kategorili yanıtlanan maddelerin iki kategorili yanıtlananlara göre daha kapsamlı bilgi verdiği düşünülerek (Ostini & Neing, 2006) çalışmalarda sıklıkla ikiden fazla kategorili maddeler

kullanılmaktadır. Böyle maddelerin analizinde iki kategorili modellerin uzantısı olarak geliştirilen modellerden yararlanılmaktadır. Kısmi puan modeli, dereceleme ölçeği, nominal tepki ve aşamalı tepki modelleri çok kategorili maddeler için geliştirilen modellerdir (De ayala, 2009). Çok kategorili modeller arasında sıklıkla kullanılan Likert ölçeği gibi sıralı kategorilere sahip çok kategorili maddeler için uygun parametrik bir model Aşamalı Tepki Modeli (ATM)'dir (Hemker, Sijtsma, Molenaar ve Junker, 1997). ATM, Samejima (1969) tarafından önerilen, iki kategorili maddeler için uygulanabilen 2PLM'nin Likert tipi ölçek gibi çok kategorili verilere genişletilmiş halidir (De Ayala, 2009).

Çok kategorili maddelerde bireylerin yetenek düzeyi ile maddelere verdikleri yanıtlar arasındaki ilişki Madde Adımı Yanıt Fonksiyonu (MAYF) ile incelenmektedir. MAYF, iki kategorili maddelerdeki madde karakteristik fonksiyonu yerine kullanılmaktadır. Şans parametresinin sıfır olduğu varsayımına dayanan ATM ile ayırıcılık ve eşik olarak iki tür parameter kestirilmektedir. Kestirimde, bütün maddeler için bir ayırt edicilik ve kategori sayısının bir eksiği kadar eşik parametresi elde edilmektedir. "*a*" ile sembolize edilen maddelerin ayırıcılık parametresidir. Ayırt edicilik MAYF'ın doğru yanıtlama olasılığının 0,5 olduğu noktadaki eğimine karşılık gelmektedir. Bu bağlamda *a* parametresinin "eğim parametresi" olarak da geçtiği kaynaklar mevcuttur. Eğri ne kadar dikse, madde o kadar ayırıcıdır. Teorik olarak a parametresi ± sonsuz arasında değer almakta, uygulamada ise sıklıkla -2,8 ile 2,8 arasında kestirilmektedir. ATM ile kestirilen ikinci parametre "*b*" ile sembolize edilen eşik parametresidir. MTK'da eşik, madde adımı yanıt fonksiyonunda maksimum eğime sahip olduğu nokta olarak tanımlanmaktadır. Yani eğrinin büküm noktasının izdüşümündeki yetenek değeridir. Bu parametre maddenin en iyi ölçtüğü yetenek düzeyini göstermektedir. Eşik değerleri -3 ile +3 arasında değişmektedir, -3'e yakın değerler kolay; +3'e yakın değerler ise zor maddeleri temsil etmektedir (De Ayala, 2009; Demars, 2010).

Çok Kategorili Maddeler İçin Parametrik Olmayan Madde Tepki Kuramı Modelleri

Parametrik olmayan modeller de parametrik modeler gibi kendi içinde iki ve çok kategorili maddeler için geliştirilen modeller olarak ikiye ayrılmaktadır. İki kategorili modeller Monoton Homojenlik Modeli ve İkili Monotonluk Modeli'dir. Çok kategorili maddeler için Oranlı Eğri Modeli, Kernel Smooting Modeli ve Monoton Homojenlik Modeli modelleri önerilmiştir. Bu modellerden Monoton Homojenlik Modeli parametrik ATM'nin parametrik olmayan uzantısıdır (Sijtsma ve Molenaar, 2002). MHM, eğitim alanında sıklıkla kullanılan Likert türü ölçeklere uygundur.

MHM'de bütün maddeler için madde ayırt ediciliğini veren bir ölçeklenebilirlik (H) katsayısı ve kategori sayısının bir eksiği kadar güçlük değeri kestirilmektedir. H katsayısı teorik olarak negatif değer de almaktadır; ancak pozitif H'ye sahip olan tüm maddeler kabul edilebilirdir. Uygulamada düşük düzeyde pozitif H'ye sahip olanlar örneğin ile 0,0 ve 0,3 gibi pozitif ancak düşük ayırıcılığa sahiptir ki bu durum testin çok geçerli olmadığını düşündürebilir (Sijtsma ve Molenaar, 2002). Diğer yandan bireylerin yetenek üzerinde güvenilir şekilde sıralanmasında çok küçük etkiye sahip olduğu söylenebilir. Mokken (1971) H için değer aralığı önermiştir: maddenin H katsayısı 0,30 ile 0,40 arasında ise zayıf; 0,40 ile 0,50 arasında ise orta; 0,50 üzerindeyse güçlü ayırıcı olduğu söylenebilir. Madde güçlüğü bireylerin maddelere doğru yanıt verme olasılıklarını ifade etmektedir (Sijtsma vd, 2008). Testin geneli için hesaplanan H katsayısı ise model veri uyumunun incelenmesinde kullanılmakta, maddeler için hesaplanan H katsayısıyla benzer şekilde yorumlanmaktadır. PoMTK'daki güçlük (*P*), PMTK'daki *b* parametresinden farklı yorumlanmaktadır. Daha düşük değerli maddeler daha zor, yüksek değerli maddeler ise daha kolaydır. 1'e yaklaşan maddeler kolayken, 0'a doğru gidildikçe maddeler zorlaşmaktadır. Madde kategorilerinin olasılıklarının kategori sırası arttıkça giderek. azalması beklenmektedir. Örneğin üç kategorili bir maddede birinci *P* değerinin ikinciden daha yüksek olması istenen bir durumdur.

PoMTK daha esnek varsayıma sahip olması ve keşfedici yapısıyla PMTK'nın güçlü varsayımlarının karşılanamadığı durumlar için alternatif olduğu belirtilmektedir. Aynı koşullarda PMTK ve PoMTK modellerinden elde edilen sonuçların incelendiği çalışmalara ihtiyaç duyulmaktadır. Parametrik

olmayan MHM, parametrik ATM'nin genel bir hali ve uzantısı olmasıyla iki modelden elde edilen sonuçlar kıyaslanabilir.

Araştırmanın Amacı

Bu araştırmanın bir amacı parametrik ve parametrik olmayan MTK modeliyle yapılan kestirimlere örneklem büyüklüğü, örneklem dağılımı, testteki madde sayısı, testte yer alan maddelerin yanıt kategori sayıları değişkenlerinin etkilerini incelemektir. Diğer bir amaç iki modelden elde edilen bulguların model-veri uyumu hata, güvenirlik değerlerini karşılaştırarak ileride yapılacak araştırmalar için model seçimi konusunda bir öneri sunmaktır.

Buradan hareketle araştırmada "Aşamalı Tepki Modeli (PMTK) ve Monoton Homojenlik Modeliyle (PoMTK) yapılan kestirimlerde model veri uyumları, güvenirlik değerleri, madde parametrelerinin hataları farklı örneklem dağılımlarında (normal, -1,0 düzeyinde çarpık), farklı örneklem büyüklüklerinde (N=100, 250, 500, 1000), farklı test uzunluklarında (k=10, 20, 40, 80), ve kategori sayısı farklı olan maddeler söz konusu olduğunda (x=3, 5, 7) nasıldır?" sorusuna yanıt aranmıştır. İlgili problem çerçevesinde iki alt problem oluşturulmuştur.

1) Örneklem dağılımı normal olduğunda Aşamalı Tepki Modeli ve Monoton Homojenlik Modeliyle yapılan kestirimlerde model veri uyumları, güvenirlik değerleri, madde parametrelerinin hataları farklı örneklem büyüklükleri, farklı test uzunlukları ve kategori sayısı farklı olan maddeler söz konusu olduğunda nasıldır?
2) Örneklem dağılımı -1,0 düzeyinde çarpık olduğunda Aşamalı Tepki Modeli ve Monoton Homojenlik Modeliyle yapılan kestirimlerde model veri uyumları, güvenirlik değerleri, madde parametrelerinin hataları farklı örneklem büyüklükleri farklı test uzunlukları ve kategori sayısı farklı olan maddeler söz konusu olduğunda nasıldır?

YÖNTEM

Bu araştırma parametrik ve parametrik olmayan madde tepki kuramı kapsamında modellere ilişkin model veri uyumu, güvenirlik, madde parametrelerinin hatalarıyla benzerlik ve farklılıklarının belirlenmesine, sayıltı ve sınırlılıklarının incelenmesine, hangisinin daha fazla bilgi sağladığının saptanmasına dayanmaktadır. Çalışma farklı modellere ilişkin olarak değişen koşullarda model uyumunu, güvenirlik değerlerini, standart hataları belirlemeyi ve karşılaştırmayı amaçlaması, kuramları test etmeye dayanması bakımından temel bir araştırma niteliğindendir.

Veri Üretimi

Araştırmanın verileri simulasyon tekniği ile üretilmiştir. İstatistiksel problemlere ilişkin olarak simülasyon verisi, eğitim alanındaki problemlerin çözümü için yürütülen çalışmalarda giderek artan bir önem kazanmaktadır (Davey, Nering & Thompson, 1997). Bu araştırmada kullanılan simülasyon verileri Wingen'3 bilgisayar programları aracılığıyla üretilmiştir. Wingen'3 yazılımı hem pratik olması hem de farklı tür dağılımlar (normal, uniform, çarpık/beta) ve modeller (iki kategorili, çok kategorili, parametrik olmayan gibi) için simülasyon verisi üretmeye imkan tanımasıyla (Han, 2007; Han ve Hambleton, 2007) tercih edilmiştir.

Araştırmada, eğitim alanındaki çalışmalarda sıklıkla kullanılan veri türü çok kategorilir. Çok kategorili yanıtlanan maddelerin iki kategorili yanıtlananlara göre daha kapsamlı bilgi verdiği düşünülmektedir (Ostini & Neing, 2006). Eğitim alanında sıklıkla kullanımından yola çıkarak bu araştırma kapsamında çok kategorili veriler tercih edilmiştir. Verilerin analizinde parametrik ATM ve parametrik olmayan MHM kullanılmıştır. MHM, ATM'nin parametrik olmayan karşılığıdır (Sjitsma ve Molenaar, 2002). Dolayısıyla üretilen veri seti iki modelin de kökeni ortak olduğu için iki model de analize de uygundur.

Parametrelerin üretilmesinde ayırt edicilik ve eşik parametreleri için ayrı dağılım özellikleri dikkate alınmıştır. Ayırt edicilik parametrelerinin üretilmesinde Log-N~(0,5; 0,4) dağılımdan, güçlük parametrelerinin üretilmesinde ise N~(0; 1) standart normal dağılımdan yararlanılmıştır. Ayırt edicilik parametrelerinin ortalaması (1,6403) ve güçlük parametrelerinin ortalaması (-0,117) eşit ve tüm koşullarda sabit olarak belirlenmiştir. Yetenek (θ) parametreleri N~(0; 1) standart normal dağılım; -1,0 düzeyinde çarpık dağılım referans alınarak üretilmiştir.

Araştırmada simülasyon çalışmasının aşamaları uygulanmıştır. Simülasyon çalışmasında sırasıyla araştırmanın amacı belirlenmiş, amaca uygun problemler tanımlanmış, araştırma koşulları düzenlenmiş (bağımsız değişkenler), veri türetilmiş (3, 5, 7 yanıt kategorili madde), madde yanıtlarıyla madde parametreleri, hataları, güvenirlik ve model-uyum değerleri elde edilmiştir. Tablo 1'de verilen araştırma desenine uygun olacak şekilde veri üretilmiştir. Araştırma sonuçları her bir koşul için 25 tekrarlı verinin analiziyle elde edilmiştir.

Araştırmanın Deseni

Çalışmada örneklem dağılımının özelliği, örneklemin büyüklüğü, madde sayısı ve maddelerin yanıt kategori sayısı araştırmanın bağımsız değişkenleri olarak ele alınmıştır. Bağımsız değişkenlere ilişkin açıklamalar sırasıyla verilmiştir.

Örneklem büyüklüğü: Bağımsız değişkenlerden biri örneklem büyüklüğüdür. Alan yazında PMTK için büyük örneklemlerin önerildiği; örneklem sayısı az olduğu durumlarda alternatif olarak PoMTK modellerinin kullanılabileceği önerilmiştir. Hem küçük hem de büyük örneklemlerde her iki modelden kestirimlerin nasıl etkilendiğini belirlemek adına 100, 250, 500, 1000 kişi olmak üzere dört farklı örneklem büyüklüğü koşulu ele alınmıştır. PMTK ile PoMTK modelleriyle elde edilen bulgulara örneklem büyüklüğünün nasıl etkilediğini sınamak ve araştırmanın sonucunda gerekli minimum örneklem büyüklüğünün ne kadar olması gerektiğine ilişkin sonuç elde etmek amacıyla bu örneklem büyüklükleri çalışmaya dahil edilmiştir.

Örneklem dağılımı: Araştırmada diğer bir bağımsız değişken örneklem dağılımının şeklidir. Normallik PMTK'nın bir varsayımı olmasına karşın araştırmalarda dağılımın normalden saptığı durumlarla sıklıkla karşılaşılmakta ve bu tür durumlarda analizlere devam edildiği görülmektedir (Doğan, 2002). Bu normalden uzaklaşmanın elde edilen kestirimlerin üzerine etkisini test etmek amacıyla normal (Çarpıklık Katsayısı=0,0),) çarpık (Çarpıklık Katsayısı=-1,0), olmak üzere iki farklı örneklem dağılımları ile çalışılmıştır.

Madde Sayısı: Araştırmanın diğer bir bağımsız değişkeni madde sayısı değişimidir. Bu değişkenin parametre iyiliği üzerine etkilerini sınamak amacıyla kısa, orta, uzun testler olarak koşullar oluşturulmuştur. 10, 20, 40, 80 maddelik testler oluşturulmuştur. Maddelerin sayısı artırılırken daha geniş madde setinin kendinden önce gelen madde setlerini içermesi sağlanmıştır. Örneğin 20 maddelik testteki ilk 10 madde, 10 maddelik testteki değerlerle aynıdır. Benzer şekilde 20 madde 40 maddeli testin ilk 20 maddesini, 40 maddeli test 80 soruluk testin ilk 40 maddesini içermektedir. Farklı uzunluktaki testlerin ortak maddeleri içermesinin karşılaştırmaları daha anlamlı kılacağı düşünülmüştür.

Kategori Sayısı: Son bağımsız değişken maddelerin sahip olduğu kategori sayısıdır. Sosyal bilim çalışmalarında genellikle beş kategorili maddeler sıklıkla kullanılmaktadır. Bu sayının artması ya da azalmasının sonuçların güvenirliği ve hatasızlığı üzerindeki etkisinin incelenmesi amaçlanarak 3, 5, 7 olmak üzere her kategori için madde üretilmiştir. 3, 5, ve 7 kategorili maddelerin için a parametreleri her koşulda birebir aynı, b parametrelerinin maddedeki kategori artmasıyla artmasına rağmen ortalamaları aynı olacak şekilde simülasyon yapılmıştır.

Üretilen gerçek yetenek ve madde parametreleri üzerinde araştırılacak olan bağımsız değişkenler ile araştırmanın deseni belirlenmiştir. Araştırma deseni Tablo 1'de özetlenmiştir.

Tablo 1. Araştırmanın Deseni

Modeller	Dağılım biçimi	Kategori sayısı	Soru sayısı	Örneklem büyüklüğü	Tekrar sayısı	Toplam
PMTK (Aşamalı Tepki Modeli)	2	3	4	4	25	2400
PoMTK(Monoton Homojenlik Modeli)	2	3	4	4	25	2400

Çalışmada parametrik ve parametrik olmayan model 2, örneklem dağılımı 2, örneklem büyüklüğü 4 ve test uzunluğu 4 koşul, maddelerin kategori sayısı 3 koşul olmak üzere (2x2x4x4x3 = 192) veri elde edilmiştir. Diğer yandan her bir koşul için 25 tekrar yapıldığından araştırma toplamda (192x25) 4800 veri seti üzerinden yürütülmüştür.

Verilerin Analizi

Verilerin analizi parametrik ve parametrik olmayan modeller için ayrı ayrı yürütülmüş ve elde edilen sonuçlar birleştirilerek karşılaştırmalar yapılmıştır.

Parametrik ATM'de model veri uyumunun belirlenmesi için -2log benzerlik değerleri, güvenirlik katsayısı, madde parametreleri ve parametrelerin standart hataları kestirilmiştir. ATM'de her bir madde için bir ayırıcılık ve kategori sayısının bir eksiği kadar eşik parametresi belirlenmektedir. İlgili parametrelere ilişkin standart hata değerleri hesaplanmıştır. İlgili hesaplamalar için MULTILOG paket programı ve Microsoft Excelden yararlanılmıştır.

Parametrik olmayan MHM'de model veri uyumunun belirlenmesi için ölçeklenebilirlik katsayısı, güvenirlik katsayısı, madde parametreleri ve parametrelerin hataları kestirilmiştir. MHM'de her bir madde için H katsayısı (ayırıcılık) ve kategori sayısının bir eksiği kadar *P* (güçlük) değeri belirlenmektedir. İlgili hesaplamalar için MSP (Mokken Scale Analysis Program) ve Microsoft Excelden yararlanılmıştır.

BULGULAR

Birinci alt problem kapsamında örneklem dağılımının normal olduğu koşulda çeşitli örneklem büyüklüğü, test uzunluğu ve farklı yanıt kategorili maddeler olduğunda Aşamalı Tepki Modeli ve Monoton Homojenlik yapılan kestirimlerde sırasıyla model veri uyumları, güvenirlik değerleri, madde parametrelerinin hataları incelenmiştir.

Model veri uyumuna yönelik bulgular

Model veri uyumunun değerlendirilmesinde kestirimlerin yapıldığı parametrik model için MULTILOG, parametrik olmayan model için MSP'den elde edilen değerler baz alınmıştır. MULTILOG program model veri uyumunun değerlendirilmesine ilişkin 2*log-benzerlik değerlerini vermektedir. Bu değerin örneklem ve parametre sayısına dayalı hesaplanması ve hesaplamaya dahil edilen parameter sayısından etkilenmesi sonuçların tek başına yorumlanmasını güçleştirmektedir (Pampel, 2000). Tablo 2'de verilen parametrik ATM model veri uyumu için -2*log-benzerlik değerleri incelendiğinde madde sayısı aynı iken örneklem büyüklüğü arttıkça -2*log-benzerlik değeri de beklendiği üzere artmıştır. Örneklemden bağımsız olarak yani aynı örneklem büyüklüğünde madde sayısı arttığında değer artış göstermiştir. Maddelerin kategori sayısı arttığında -2*log-benzerlik değerlerinin arttığı tespit edilmiştir. Üç bağımsız değişkene ilişkin bulgu, değerlerin tek başına yorumlanmasının güçlüğüne işaret etmektedir. ATM ile model veri uyumunu incelemek için programdan elde edilen sonuçlar tek başına yorumlanamamaktadır. Bu bağlamda araştırmada ele alınan bağımsız değişkenlerin parametrik ATM model uyumuna etkisine ilişkin genelleme yoluna gidilmemiştir.

Tablo 2. Örneklem Dağılımı Normal Olduğunda Model Veri Uyumuna Ait Bulgular

		Aşamalı Tepki Modeli (PMTK)			*Monoton Homojenlik Model (PoMTK)*		
		-2log benzerlik			H Katsayısı		
Madde	Birey	3 kategori	5 kategori	7 kategori	3 kategori	5 kategori	7 kategori
10	100	699,2	1713,8	2225,9	0,42	0,41	0,41
	250	1448,0	3839,9	5371,8	0,42	0,41	0,40
	500	2176,7	6980,0	10433,2	0,41	0,41	0,40
	1000	3341,0	12833,0	19494,6	0,41	0,40	0,40
20	100	2187,0	4239,4	5371,0	0,42	0,41	0,40
	250	5300,9	10169,8	13078,3	0,42	0,41	0,40
	500	10053,9	19943,1	25690,3	0,40	0,40	0,40
	1000	19226,0	38917,5	50572,0	0,40	0,40	0,40
40	100	5382,7	8872,4	9800,5	0,42	0,41	0,40
	250	12983,5	21592,6	25200,2	0,42	0,41	0,40
	500	26325,0	43591,6	51800,1	0,40	0,40	0,40
	1000	51108,0	86639,7	106400,5	0,39	0,39	0,39
80	100	8472,7	14534,5	19600,0	0,42	0,41	0,40
	250	18562,7	36006,6	50462,8	0,41	0,40	0,40
	500	37855,1	74087,5	130661,0	0,40	0,40	0,40
	1000	74311,4	15282,9	212896,6	0,39	0,39	0,39

Parametrik olmayan model veri uyumunun incelenmesi ise tek başına yorumlanabilen H ölçeklenebilirlik katsayısına dayalı olarak yapılmaktadır (Sijtsma & Molenaar, 2002). Tablo 2'deki bulgular parametrik olmayan MHM için incelendiğinde, model uyumu için kullanılan ölçeklenebilirlik (H) katsayısı madde sayısı örneklem artışı ve maddenin yanıt kategorisi artışıyla istatistiksel olarak önemli ölçüde değişmemiştir. Tüm koşullarda birbirine ve 0,4'e yakın değerler elde edilmiştir. Katsayının 0,3 ve üzerinde olması kabul edilebilir model uyumu olduğunu göstermektedir (Sijtsma & Molenaar, 2002). Bu bilgiden yola çıkarak ele alınan tüm durumlarda model uyumu düşük düzeyde ve kabul edilebilirdir yorumu yapılabilir.

Model veri uyumunun değerlendirilmesinde analiz için kullanılan ATM'de MULTILOG, MHM'de MSP çıktılarında yer alan model veri uyum indeksleri temel alınmıştır. ATM için kullanılan model veri uyum katsayılarının örneklem büyüklüğünden ve analize dahil edilen parametre sayısından etkilenmesi doğrudan yorumlanmasını güçleştirmektedir. (Pampel, 2000; De Ayala, 2009). MHM sonuçları ise farklı koşullarda kararlı bir yapı sergilemiştir. H katsayısının tek başına yorumlanabilmesi farklı koşulların karşılaştırılabilmesine imkân tanımaktadır (Sijtsma & Molenaar, 2002). Bu bağlamda MHM'nin model veri uyumu kestiriminde avantajlı olduğu düşünülebilir.

Güvenirlik değerlerine ilişkin bulgular

Normal dağılım koşulları için güvenirlik katsayılarının değerlendirilmesine yönelik bulgular Tablo 3'te sunulmuştur.

Tablo 3 incelendiğinde, parametrik model için kestirilen marjinal güvenirliğin elde edilen en düşük değeri en küçük örneklem, en az sayıda ve üç kategorili maddelerin olduğu koşulda elde edilmiştir. Madde sayısı aynı iken örneklem büyüklüğü arttıkça değerler önemli bir değişim göstermemiştir. Aynı madde sayısında ve örneklem büyüklüğünde maddenin yanıt kategori sayısı arttığında güvenirlik değerleri artmıştır. Ancak bu fark Kruskal Wallis testine göre anlamlı değildir [χ^2=2,33 (2), p>.05]. Madde sayısı değişimi diğer faktörlere göre daha belirgin değişime işaret etmiştir. Aynı örneklem büyüklüğü ve kategori sayısında madde sayısının artışı güvenirliği artırmıştır.

Tablo 3'e göre MHM'de rho güvenirlik değerleri virgülden sonra yüzde bir basamağı olacak şekilde hesaplanmakta dolayısıyla parametrik hesaplama sonuçlarına göre 0,01'i geçemeyecek düzeyde küçük değişiklikler içermektedir. Bu modelden kestirilen güvenirlik değerleri madde sayısı artışından etkilenmekte, madde sayısı arttıkça güvenirlik artmaktadır. Maddenin yanıt kategori sayısı artışı madde sayısı ve örneklem büyüklüğü aynıyken değerlerde anlamlı olmayan artışa işaret etmiş

[χ^2=2,25 (2), p>.05], 40 madde ve üzerinde 5 ve 7 kategorili koşullar için aynı değerler elde edilmiştir.

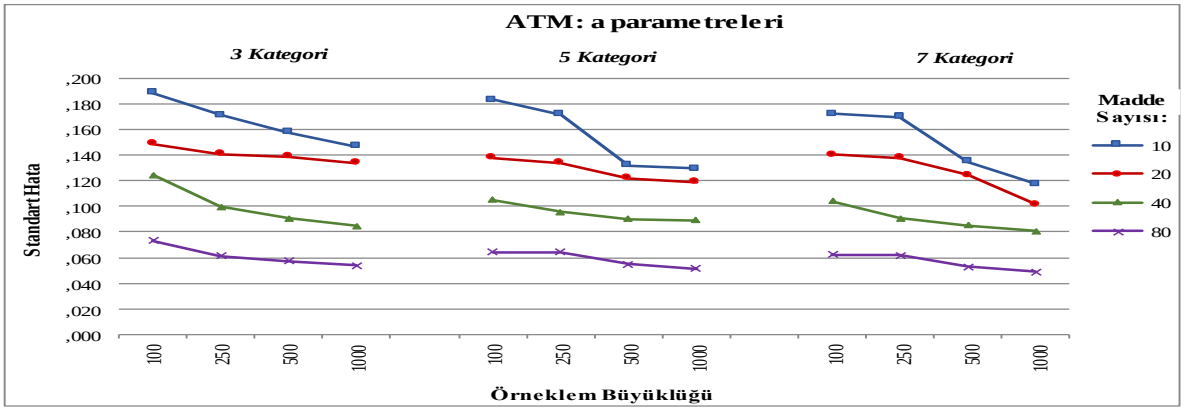
Tablo 3: Örneklem Dağılımı Normal Olduğunda Güvenirlik Değerlerine Ait Bulgular

		Aşamalı Tepki Modeli (PMTK)			*Monoton Homojenlik Model (PoMTK)*		
		r			Rho Katsayısı		
Madde	Birey	3 kategori	5 kategori	7 kategori	3 kategori	5 kategori	7 kategori
10	100	0,847	0,874	0,887	0,850	0,860	0,860
	250	0,845	0,873	0,885	0,850	0,860	0,860
	500	0,845	0,871	0,885	0,850	0,860	0,860
	1000	0,846	0,875	0,886	0,850	0,860	0,860
20	100	0,902	0,927	0,934	0,910	0,930	0,940
	250	0,901	0,930	0,936	0,910	0,930	0,940
	500	0,904	0,927	0,934	0,910	0,930	0,940
	1000	0,902	0,930	0,938	0,910	0,930	0,940
40	100	0,952	0,964	0,966	0,960	0,970	0,970
	250	0,955	0,964	0,966	0,960	0,970	0,970
	500	0,954	0,962	0,965	0,960	0,970	0,970
	1000	0,955	0,964	0,965	0,960	0,970	0,970
80	100	0,977	0,979	0,979	0,980	0,980	0,980
	250	0,973	0,979	0,978	0,980	0,980	0,980
	500	0,972	0,978	0,979	0,980	0,980	0,980
	1000	0,976	0,979	0,979	0,980	0,980	0,980

Parametrik MTK'da güvenirlik madde ve test bilgi fonksiyonları açıklanmakta, test bilgi fonksiyonlarının ortalamasına dayalı hesaplanan marjinal güvenirlik katsayısı ile ifade edilmektedir (Thissen, 1991). Ölçme sonuçlarının hatasız olması beklendiği için güvenilirliğin yüksek olması istenen durumdur. Çalışılan kuramdaki modeller güvenirlik bağlamında paralel sonuçlar vermiştir. Bu durum araştırma koşullarının MTK varsayımlarını karşıladığı durumlarda ATM ve MHM'nin özellikle güvenirlik değeri bakımından benzer sonuçlar üretmelerinin beklendiği bulgusuyla örtüşmektedir (Dyehouse, 2009). Güvenirliğe ilişkin elde edilen sonuçlar MHM'nin ATM'nin daha genel bir formu olması bilgisini doğrularken (Sjitsma vd., 2008), örneklem ve test koşulları parametrik bir kestirime uygun olmadığı durumlar için MHM'nin kullanılmasının önerilebileceğini göstermektedir.

Madde Parametrelerinin Hatalarına İlişkin Bulgular

Örneklem normal dağılım gösterdiği koşullarda ATM'den elde edilen ayırıcılık (*a*) parametrelerinin standart hatalarının araştırmanın bağımsız değişken koşullarına göre değişimini veren grafik Şekil 1'de sunulmuştur.

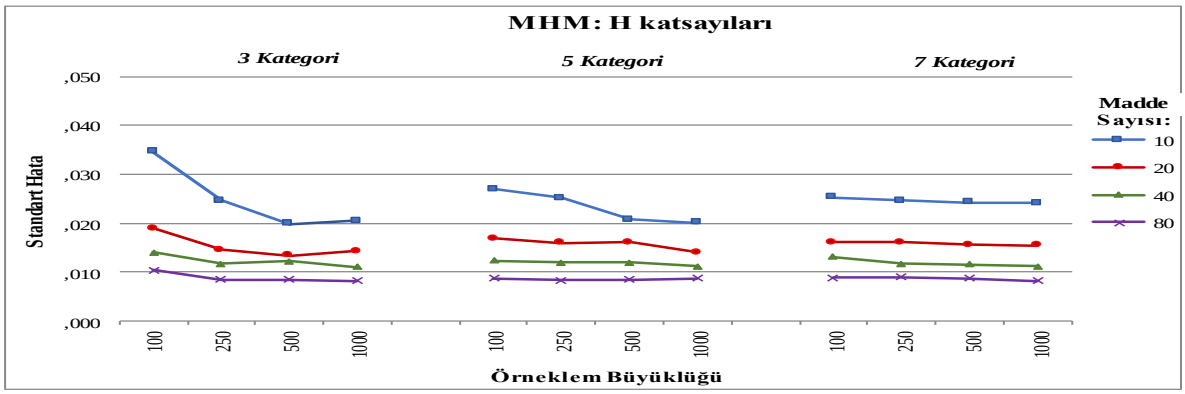


Şekil 1. Örneklem Dağılımı Normal Olduğunda Ayırıcılık Parametrelerinin Standart Hata Değerleri

Şekil 1'de verilen üç kategorili maddelerde ayırıcılık (a) parametrelerinin standart hata değerlerine ilişkin bulgular incelendiğinde, değerlerin 0,19 ile 0,05 arasında olduğu görülmektedir. Genel olarak örneklem büyüklüğü arttığında beklentiye uygun olarak hata azalmıştır. Bu azalma 10 maddeli koşullarda daha belirgindir. Madde sayısı etkisi incelendiğinde, en hatalı değerler 10 maddeli koşula aittir. Hatadaki azalma en belirgin olarak 10 maddeden 20 maddeye çıkıldığında gözlenmiştir. 40 madde ve üzerinde hem hatalar birbirine yakın hem de değerleri düşüktür. Bu bulgudan hareketle madde sayısı artışı hatada örneklem büyüklüğü artışından daha fazla etkili olmuştur yorumu yapılabilir. *a* parametrelerinin standart hatasının örneklem artışı ve madde sayısı artışıyla birlikte azalması diğer çalışmalarla paralellik göstermektedir (Koğar, 2015).

Beş ve yedi kategorili maddeler için standart hata bulguları üç kategorili durumla paraleldir. Ancak beş ve yedi kategorili maddelerde üç kategorili yanıtlanan maddelerden farklı olark görece daha düşük hatalı kestirimler yapıldığı görülmektedir. Bu durum madde sayısı değişiminin mümkün olmadığı durumlarda kategori sayısı artırılmasının hatayı azaltma yönünde etkisi olacağı şeklinde yorumlanabilir.

MHM'den elde edilen ayırıcılık H katsayılarının standart hatalarının araştırmanın bağımsız değişken koşullarına göre değişimini veren grafik Şekil 2'de sunulmuştur.



Şekil 2. Örneklem Dağılımı Normal Olduğunda Ayırıcılık Parametrelerinin Standart Hata Değerleri

Şekil 2'de verilen MHM'den kestirilen H katsayılarının hataları ise 0,035 ile 0,01 arasındadır. En yüksek hata değerleri az maddeli koşullarda görülmüş, madde sayısı arttığında hata sayısı azalmıştır. Örneklem arttığında beklentiye uygun olarak hata azalmıştır. Hatadaki azalma en belirgin olarak 10 maddeden 20 maddeye çıkıldığında görülmektedir. Örneklem büyüklüğünün artışı hata değerinde

belirgin bir değişime işaret etmemiştir. Bu bulgu MHM'nin küçük örneklemlerde kullanılabilir olması bilgisiyle tutarlıdır (Sijitsma ve Molenaar, 2002). Bulgular maddenin yanıt kategorisi açısından yorumlandığında beş ve yedi kategorili maddelerin üç kategorili maddelere göre görece daha düşük hatalı kestirildiği ve paralel sonuçlar verdikleri görülmektedir. İki modele ait bulgular karşılaştırıldığında, MHM'nin en yüksek hata değerinin ATM'nin en düşük hata değerinin altında olması MHM kestirimlerinin daha az hatalı olduğu şeklinde yorumlanabilir. ATM'den kestirilen *b* parametreleri ile MHM'den kestirilen *P* değerlerinin standart hatalarına ilişkin bulgular Tablo 4'te özetlenmiştir.

Tablo 4. Örneklem Dağılımı Normal Olduğunda Madde Parametrelerinin Standart Hataları

		Parametrik Model						***Parametrik Olmayan Model***					
*	**	SH_{b1}	SH_{b2}	SH_{b3}	SH_{b4}	SH_{b5}	SH_{b6}	SH_{P1}	SH_{P2}	SH_{p3}	SH_{P4}	SH_{P5}	SH_{P6}
10	100	0,19	0,18					0,06	0,06				
	250	0,17	0,17					0,06	0,05				
	500	0,16	017					0,06	0,05				
	1000	0,16	0,17					0,06	0,05				
20	100	0,18	0,17					0,04	0,04				
	250	0,16	0,16					0,04	0,04				
	500	0,15	0,16					0,04	0,04				
	1000	0,15	0,16					0,04	0,04				
40	100	0,12	0,15					0,03	0,03				
	250	0,11	0,13					0,03	0,03				
	500	0,11	0,11					0,03	0,03				
	1000	0,10	0,11					0,03	0,03				
80	100	0,10	0,10					0,02	0,02				
	250	0,09	0,09					0,02	0,02				
	500	0,09	0,09					0,02	0,02				
	1000	0,08	0,09					0,02	0,02				
10	100	0,16	0,15	0,15	0,14			0,02	0,03	0,03	0,02		
	250	0,15	0,12	0,14	0,12			0,02	0,02	0,03	0,02		
	500	0,12	0,11	0,11	0,10			0,02	0,02	0,02	0,02		
	1000	0,11	0,10	0,09	0,09			0,02	0,02	0,02	0,02		
20	100	0,13	0,12	0,12	0,11			0,02	0,03	0,02	0,02		
	250	0,12	0,11	0,09	0,10			0,02	0,02	0,02	0,01		
	500	0,11	0,09	0,08	0,09			0,02	0,02	0,02	0,01		
	1000	0,10	0,07	0,07	0,08			0,02	0,02	0,02	0,01		
40	100	0,12	0,11	0,10	0,10			0,02	0,02	0,02	0,01		
	250	0,11	0,09	0,07	0,09			0,02	0,02	0,02	0,01		
	500	0,10	0,08	0,06	0,07			0,02	0,02	0,02	0,01		
	1000	0,08	0,06	0,05	0,06			0,01	0,02	0,02	0,01		
80	100	0,08	0,06	0,05	0,07			0,01	0,01	0,01	0,01		
	250	0,07	0,05	0,05	0,07			0,01	0,01	0,01	0,01		
	500	0,06	0,05	0,05	0,06			0,01	0,01	0,01	0,01		
	1000	0,07	0,05	0,05	0,06			0,01	0,01	0,01	0,01		
10	100	0,13	0,13	0,11	0,11	0,11	0,11	0,02	0,03	0,03	0,03	0,02	0,02
	250	0,11	0,11	0,09	0,10	0,10	0,11	0,02	0,03	0,02	0,02	0,02	0,02
	500	0,09	0,10	0,09	0,09	0,09	0,10	0,02	0,02	0,02	0,02	0,02	0,02
	1000	0,08	0,08	0,08	0,08	0,08	0,09	0,02	0,02	0,02	0,02	0,02	0,02
20	100	0,10	0,10	0,09	0,11	0,10	0,10	0,02	0,03	0,03	0,03	0,02	0,02
	250	0,09	0,09	0,09	0,10	0,09	0,09	0,02	0,02	0,02	0,02	0,02	0,02
	500	0,09	0,08	0,08	0,08	0,09	0,09	0,02	0,02	0,02	0,02	0,02	0,02
	1000	0,08	0,08	0,08	0,07	0,08	0,08	0,02	0,02	0,02	0,02	0,02	0,02
40	100	0,08	0,08	0,08	0,07	0,06	0,07	0,02	0,02	0,02	0,02	0,02	0,01
	250	0,07	0,07	0,08	0,06	0,05	0,06	0,02	0,02	0,02	0,02	0,01	0,01
	500	0,07	0,07	0,07	0,05	0,05	0,06	0,02	0,02	0,02	0,02	0,02	0,02
	1000	0,07	0,07	0,07	0,05	0,05	0,06	0,02	0,02	0,02	0,02	0,01	0,01
80	100	0,06	0,05	0,06	0,05	0,05	0,06	0,01	0,01	0,01	0,01	0,01	0,01
	250	0,05	0,05	0,05	0,05	0,05	0,05	0,01	0,01	0,01	0,01	0,01	0,01
	500	0,06	0,06	0,05	0,05	0,05	0,05	0,01	0,01	0,01	0,01	0,01	0,01
	1000	0,05	0,05	0,05	0,05	0,04	0,05	0,01	0,01	0,01	0,01	0,01	0,01

**madde sayısı **örneklem büyüklüğü*

Tablo 4'te verilen üç kategorili maddeler için ATM ile kestirilen b_1 parametresinin standart hataları 0,19 ile 0,09; b_2'nin ise 0,18 ile 0,09 arasındadır. MHM'de kestirilen P_1 ve P_2 değerlerinin hatası 0,06 ile 0,02 arasında değişkenlik göstermektedir. En yüksek hatalar 10 maddeli koşulda kestirilmiş; madde sayısı arttıkça hatada düşüş tespit edilmiştir. Madde sayısı artışı hata üzerinde etkili olmuştur. Örneklem sayısı artışı da hatada küçük bir azalmaya işaret etse de bu değişim madde sayısındaki kadar belirgin değildir. İki modele ait bulgular karşılaştırıldığında, MHM'nin en yüksek hata değerinin ATM'nin en düşük hata değerinin altında olması MHM kestirimlerinin daha az hatalı kestirim yapıldığına işaret edebilir.

Tablo 4 beş kategorili maddeler için incelendiğinde, ATM ile kestirilen b_1 parametrelerinin hatalarının 0,16 ile 0,06 arasında olduğu; diğer üç *b* parametresiyle hata düzeylerinin farklı koşullar için birbirine yakın olduğu görülmektedir. MHM'de hata değerleri 0,02 ile 0,01 arasında değer almıştır. MHM ile kestirilen parametrelerin hata düzeyleri örneklem bütüklüğü, madde sayısı ve maddenin yanıt kategorisi sayısının artışında ATM ile paralellik göstermekle birlikte görece daha düşüktür. Buradan yola çıkarak MHM kestirimlerinin daha az hatalı olduğuna işaret edilebilir.

Yedi yanıt kategorili koşullarda hata üç ve beş kategorili maddelere göre hata görece azalmıştır. En yüksek hata 10 madde ve 100 birey olduğunda kestirilmiştir. Hem birey sayısı hem de madde sayısı artışı hatada azalmaya işaret etmiştir. Yani hata değişimi büyük örneklemde ve uzun testlerde daha az gözlenmektedir. ATM'de kestirilen hatayla karşılaştırıldığında MHM kestirimlerindeki hatanın daha düşük olduğu görülmektedir.

İkinci alt problem kapsamında örneklem dağılımının -1,0 düzeyinde çarpık olduğu koşulda çeşitli örneklem büyüklüğü, test uzunluğu ve farklı yanıt kategorili maddeler olduğunda Aşamalı Tepki Modeli (PMTK) ve Monoton Homojenlik Modeliyle (PoMTK) yapılan kestirimlerde sırasıyla model veri uyumları, güvenirlik değerleri, madde parametrelerinin hataları incelenmiştir.

Model veri uyumuna yönelik bulgular

Örneklem dağılımının -1,0 düzeyinde çarpık olduğu koşullar için model veri uyumlarının değerlendirilmesine yönelik bulgular Tablo 5'te sunulmuştur.

Tablo 5. Örneklem Dağılımı -1,0 düzeyinde Çarpık Olduğunda Model Veri Uyumuna Ait Bulgular

		Aşamalı Tepki Modeli (PMTK)			*Monoton Homojenlik Modeli (PoMTK)*		
		*-2*log benzerlik*			*H Katsayısı*		
Madde	Birey	3 kategori	5 kategori	7 kategori	3 kategori	5 kategori	7 kategori
10	100	293,4	295,14	665,54	0,17	0,18	0,19
	250	1171,4	291,09	704,32	0,18	0,19	0,18
	500	2641,6	201,52	2087,54	0,18	0,16	0,16
	1000	7037,7	1521,97	2090,21	0,18	0,16	0,16
20	100	854,4	1415,9	2232,6	0,21	0,21	0,21
	250	1817,6	2966,4	4902,8	0,19	0,17	0,19
	500	2640,0	5766,2	9771,3	0,19	0,17	0,14
	1000	58308,2	10272,5	18140,7	0,18	0,16	0,15
40	100	1762,0	3632,9	5075,0	0,23	0,19	0,16
	250	3716,5	8455,5	12222,9	0,18	0,16	0,16
	500	7649,6	17090,1	24923,3	0,17	0,15	0,15
	1000	14548	32905,7	48317,7	0,17	0,16	0,16
80	100	5519,1	7666,7	9964,7	0,22	0,17	0,21
	250	15129,5	19534,8	24397,7	0,18	0,14	0,16
	500	29933,1	37260,6	46093,8	0,18	0,16	0,16
	1000	60248,3	72733,9	89957,3	0,18	0,17	0,16

Tablo 5'te verilen parametrik modeller için model-veri uyumu için incelenen -2*log-benzerlik değerlerinin örneklem büyüklüğü, madde ve maddenin yanıt kategori sayısı artıkça arttığı tespit edilmiştir. Model veri uyumunun azalmasında değişkenlerin etkisine dayalı bir düşüş gerçekleşmiş olabilir; ancak -2*log-benzerlik değerlerinin parametre değişiminden etkilenmesi tek başına yorumlanmasını güçleştirmektedir (Liu & Maydeu- Oliveres, 2014).

Parametrik olmayan MHM ile toplam puan (X_+) üzerinden kestirilen H katsayılarının tamamı birbirine ve 0,2'ye yakın değer almıştır. Molenaar & Sijtsma (2002) kabul edilebilir uyum için alt ölçüt sınırı olarak H'nin 0,3 olmasını önermiştir. Bu bilgi doğrultusunda model uyumunun düşük olduğu görülmektedir. Böyle durumlar için MSP'nin otomatik madde seçimi ile maddelerin incelenmesine dayalı yöntemlerle model veri uyumu iyileştirilebilir (Molenaar & Sijtsma, 2002). Ancak bu çalışmada amaç parametrik modelle kıyaslama yapmak olduğundan doğrudan Mokken ölçekleme sonuçları verilmiştir. H uyum katsayısında örneklem, madde ve maddenin kategorili sayısı değişimi ATM'deki kadar etkili olmamıştır.

Bir diğer bulgu -1,0 düzeyinde çarpık dağılımdan elde edilen sonuçların normal dağılıma göre daha düşük olmasıdır. Diğer bağımsız değişkenlere göre model uyumu üzerinde dağılım şekli daha fazla etkili olmuştur. Bu bulgu, Mokken ölçeklemesinde H katsayısı hesaplanırken toplam puan (X_+) kullanılması yani birey özelliklerinin sonuçları etkilemesine dayandırılabilir (Mokken, 1971).

Güvenirlik değerlerine yönelik bulgular

Örneklem dağılımı -1.0 düzeyinde çarpık olduğu koşullar için ATM ve MHM'den hesaplanan güvenirliklerin değerlendirilmesine yönelik bulgular Tablo 6'da sunulmuştur.

Tablo 6.Örneklem Dağılımı -1,0 Düzeyinde Çarpık Olduğunda Güvenirlik Değerlerine Ait Bulgular

		ATM (PMTK)			MHM (PoMTK)		
		r			Rho Katsayısı		
Madde	Birey	3 kategori	5 kategori	7 kategori	3 kategori	5 kategori	7 kategori
10	100	0,553	0,599	0,596	0,57	0,75	0,77
	250	0,556	0,595	0,600	0,55	0,75	0,76
	500	0,559	0,594	0,593	0,56	0,74	0,77
	1000	0,555	0,598	0,599	0,56	0,75	0,77
20	100	0,640	0,761	0,755	0,75	0,81	0,81
	250	0,643	0,727	0,701	0,75	0,77	0,78
	500	0,640	0,694	0,679	0,75	0,77	0,78
	1000	0,643	0,682	0,686	0,75	0,77	0,78
40	100	0,766	0,784	0,827	0,88	0,87	0,86
	250	0,766	0,798	0,827	0,84	0,85	0,86
	500	0,765	0,793	0,824	0,85	0,86	0,86
	1000	0,764	0,793	0,825	0,85	0,86	0,86
80	100	0,858	0,870	0,898	0,93	0,92	0,92
	250	0,857	0,871	0,898	0,91	0,92	0,92
	500	0,857	0,874	0,899	0,92	0,92	0,92
	1000	0,858	0,877	0,893	0,92	0,92	0,92

Tablo 6'da görülüğü gibi ATM'den kestirilen güvenirlik değerleri üç, beş ve yedi kategorili maddeler için 0,6 ile 0,9 arasında değişkenlik göstermiştir, ancak değerlerin ondalık kesir kısımlarında farklılık mevcuttur. Beş ve yedi kategorili maddelerin güvenirlikleri birbirine daha yakındır. Yapılan Kruskal Wallis analizi sonucuna göre üç ayrı yanıt kategorisi için değerler birbirinden anlamlı düzeyde farklı değildir [χ^2 (2)=2.138, p>.05]. ATM'de güvenirlik madde sayısı artışıyla yükselmiştir. Örneklem büyüklüğünün artışı katsayıda belirgin bir artış ya da azalışa işaret etmemiş; değerler birbirine yakın olarak kestirilmiştir. Normal ve çarpık dağılıma kıyasla

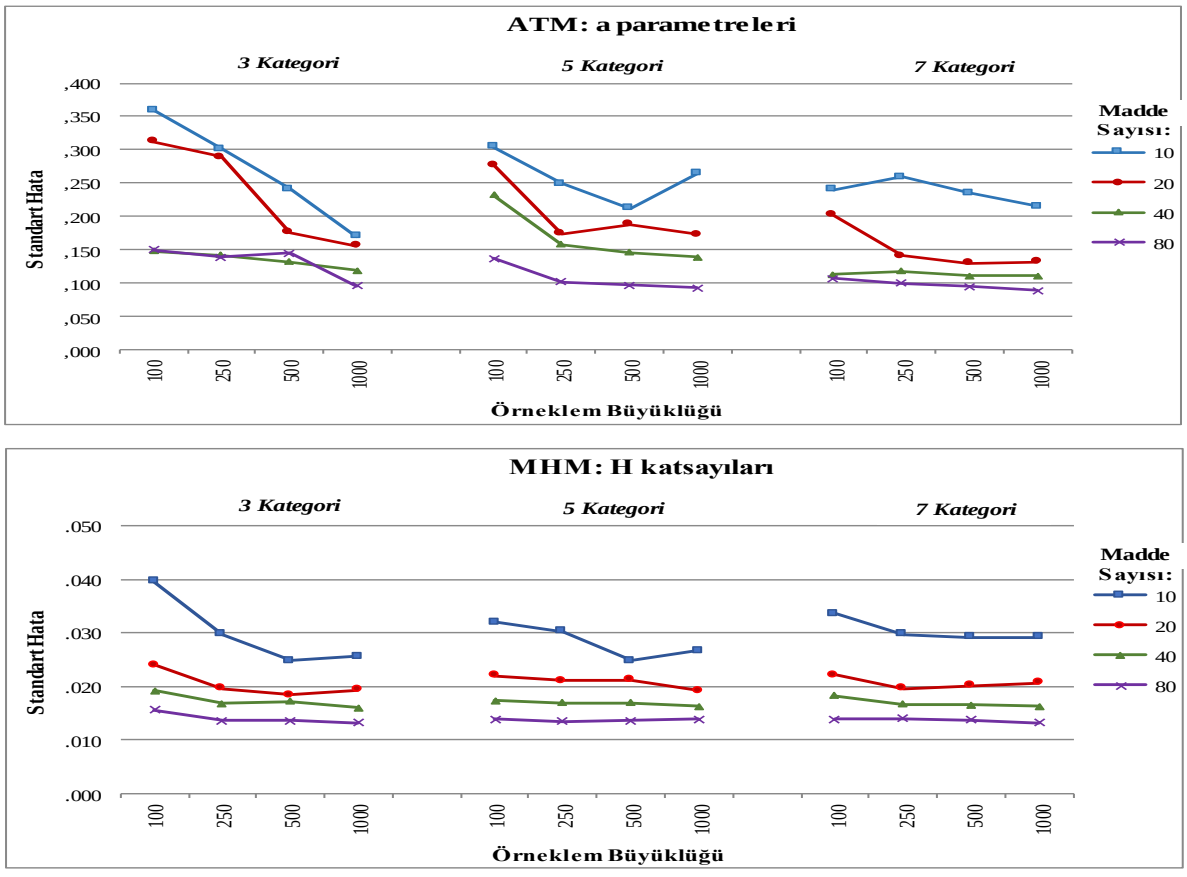
güvenirliklerin düşük olması kestirimler üzerinde dağılım şeklinin etkili olduğu şeklinde yorumlanabilir.

Tablo 6 MHM açısından incelendiğinde üç kategorili maddeler için güvenirlik değerleri 0,6 ile 0,9; beş ve yedi kategorili maddeler için 0,7 ile 0,9 arasında elde edilmiştir. ATM'de olduğu gibi beş ve yedi kategorili maddelerin güvenirlik bulguları birbirine daha yakındır. Yapılan Kruskal Wallis analizi sonucuna göre üç ayrı kategori değerleri birbirinden anlamlı düzeyde farklı değildir [χ^2 (2)=.902, p>.05]. Buradan yola çıkarak güvenirlik değerleri üzerine kategori artışının büyük bir etkiye sahip olmadığı söylenebilir. Madde sayısı değişiminin ise değerler üzerinde etkili olduğu görülmektedir. Örneklem artışı açısından değerlendirildiğinde güvenirlik üzerinde belirgin bir etki tespit edilmemiştir. Değerlerin normal dağılıma nazaran az olması çarpık dağılımda daha hatalı kestirim yapıldığı şeklinde ifade edilebilir.

İki model bulguları birlikte incelendiğinde MHM güvenirlik değerlerinin ATM'ye göre daha yüksek kestirildiği görülmektedir. Ancak yapılan Mann Whitney U testi sonuçlarına göre fark anlamlı değildir (üç kategorili maddeler için (U=108; p>.05); beş kategorili maddeler için (U=79,5; p>.05); yedi kategorili maddeler için (U=80; p>.05). İki model arasında istatiksel olarak anlamlı olmasa da fark olmasının nedeni ATM'nin daha sıkı kurallarla kestirim yapması; MHM'nin daha esnek olması nedeniyle hataları daha fazla tolere etmesine dayalı olabilir (Sijitsma vd, 2008).

Madde hatalarına ilişkin bulgular

ATM'den elde edilen *a* parametrelerinin ve MHM'den kestirilen H değerlerinin standart hatalarının araştırmanın bağımsız değişken koşullarına göre değişimini veren grafik Şekil 3'te sunulmuştur.



Şekil 3. Örneklem Dağılımı -1,0 Düzeyinde Çarpık Olduğunda Ayırıcılık Parametrelerinin Standart Hata Değerleri

Şekil 3'te görselleştirilen üç kategorili maddelerde *a* parametrelerinin standart hata değerlerine ilişkin bulgular incelendiğinde, değerlerin 0,36 ile 0,10 arasında olduğu görülmektedir. MHM'den kestirilen H katsayılarının hataları ise 0,035 ile 0,009 arasındadır. Genel olarak örneklem büyüklüğü arttığında beklentiye uygun olarak hata azalmıştır. Bu azalma düşük sayıda maddeli koşullarda daha belirgindir. Madde sayısı değişiminin etkisi incelendiğinde en yüksek değerler 10 maddeli koşula aittir. Hatadaki azalma en belirgin olarak ATM'de 20 maddeden 40 maddeye; MHM'de ise 10 maddeden 20 maddeye çıkıldığında gözlenmiştir. 40 madde ve üzerinde hem hatalar birbirine yakın hem de değerleri daha düşüktür. Bu bulgudan hareketle madde sayısı artışı hata üzerinde örneklem büyüklüğü artışından daha fazla etkili olmuştur yorumu yapılabilir. Beş ve yedi kategorili maddelerin bulguları üç maddeli durumla paraleldir. ATM ve MHM'deki üç kategorili maddelerin *b* parametrelerinin standart hata bulguları Tablo 7'de sunulmuştur.

Tablo 7.Örneklem Dağılımı -1,0 Düzeyinde Çarpık Olduğunda Standart Hata Değerlerine Ait Bulgular

		Parametrik Model				***Parametrik Olmayan Model***			
*	**	SH_{b1}	SH_{b2}			SH_{P1}	SH_{P2}		
10	100	0,90	0,50			0,06	0,06		
	250	0,55	0,54			0,06	0,05		
	500	0,61	0,46			0,06	0,05		
	1000	0,36	0,29			0,06	0,05		
20	100	0,82	0,34			0,05	0,04		
	250	0,40	0,37			0,04	0,04		
	500	0,40	0,30			0,04	0,04		
	1000	0,34	0,26			0,04	0,04		
40	100	0,60	0,25			0,03	0,03		
	250	0,37	0,29			0,03	0,03		
	500	0,32	0,22			0,03	0,03		
	1000	0,21	0,18			0,03	0,03		
80	100	0,35	0,24			0,02	0,02		
	250	0,33	020			0,02	0,02		
	500	0,32	0,25			0,02	0,02		
	1000	0,14	0,12			0,02	0,02		
		SH_{b1}	SH_{b2}	SH_{b3}	SH_{b4}	SH_{P1}	SH_{P2}	SH_{p3}	SH_{P4}
10	100	0,24	0,20	0,12	0,24	0,03	0,03	0,03	0,03
	250	0,19	0,16	0,10	0,18	0,02	0,02	0,03	0,03
	500	0,11	0,13	0,10	0,15	0,03	0,03	0,03	0,03
	1000	0,08	0,10	0,10	0,10	0,03	0,03	0,03	0,03
20	100	0,23	0,12	0,10	0,23	0,02	0,03	0,02	0,02
	250	0,13	0,12	0,06	0,17	0,02	0,03	0,02	0,02
	500	0,08	0,11	0,05	0,09	0,02	0,02	0,02	0,02
	1000	008	0,08	0,06	0,08	0,02	0,03	0,02	0,02
40	100	0,12	0,14	0,12	0,17	0,02	0,03	0,02	0,02
	250	0,11	0,11	0,05	0,15	0,02	0,02	0,02	0,02
	500	0,10	0,09	0,06	0,07	0,02	0,02	0,02	0,02
	1000	0,08	0,08	0,06	0,08	0,02	0,02	0,02	0,02
80	100	0,08	0,11	0,10	0,15	0,01	0,02	0,02	0,02
	250	0,07	0,10	0,10	0,12	0,01	0,02	0,02	0,02
	500	0,06	0,10	0,09	0,09	0,01	0,02	0,02	0,02
	1000	0,07	0,08	0,10	0,07	0,01	0,02	0,02	0,02

**madde sayısı* ***örneklem büyüklüğü*

Tablo 7'de verilen üç kategorili maddeler için ATM ile kestirilen b_1 parametresinin standart hataları 0,90 ile 0,14; b_2'nin 0,50 ile 0,12 arasındadır. MHM'de hem P_1 hatası 0,06 ile 0,02; hem de P_2 hatası 0,06 ile 0,02 arasında değişkenlik göstermektedir. Madde sayısı artışı hata üzerinde etkili olmuştur. En yüksek hatalar 10 maddeli koşulda kestirilmiş; madde sayısı arttıkça hatada düşüş tespit edilmiştir. Örneklem büyüklüğü artışı hatada küçük bir azalmaya işaret etse de bu değişim madde sayısındaki kadar belirgin değildir. İki modele ait bulgular karşılaştırıldığında, MHM'nin en yüksek

hata değerinin ATM'nin en düşük hata değerinin altında olması MHM kestirimlerinin daha az hatalı kestirim yaptığına işaret edebilir.

Tablo 7'de beş kategorili maddeler için ATM ile kestirilen b_1 parametrelerine hatalarının 0,24 ile 0,06 arasında olduğu; diğer üç *b* parametresiyle hata düzeylerinin farklı koşullar için birbirine yakın olduğu görülmektedir. MHM'de hata değerleri 0,03 ile 0,01 arasında değer almıştır. MHM ile kestirilen parametrelerin hata düzeyleri örneklem bütüklüğü, madde sayısı ve maddenin yanıt kategorisi sayısının artışında ATM ile paralellik göstermekle birlikte görece daha düşüktür. Buradan yol çıkarak MHM kestirimlerinin daha az hatalı olduğu söylenebilir.

Tablo 8. Örneklem Dağılımı -1,0 Düzeyinde Çarpık Olduğunda Yedi Kategorili Maddelerin Standart Hata Değerlerine Ait Bulgular

		Parametrik Model						***Parametrik Olmayan Model***					
*	**	SH_{b1}	SH_{b2}	SH_{b3}	SH_{b4}	SH_{b5}	SH_{b6}	SH_{P1}	SH_{P2}	SH_{p3}	SH_{P4}	SH_{P5}	SH_{P6}
10	100	0,20	0,14	0,17	0,18	0,12	0,19	0,03	0,03	0,03	0,03	0,03	0,03
	250	0,18	0,13	0,16	0,16	0,10	0,18	0,02	0,03	0,02	0,03	0,03	0,03
	500	0,15	0,11	0,13	0,11	0,09	0,15	0,02	0,02	0,02	0,03	0,03	0,02
	1000	0,10	0,08	0,10	0,08	0,09	0,12	0,02	0,02	0,02	0,03	0,03	0,02
20	100	0,20	0,09	0,12	0,15	0,10	0,19	0,02	0,03	0,03	0,03	0,03	0,02
	250	0,17	0,08	0,12	0,13	0,06	0,15	0,02	0,02	0,03	0,03	0,02	0,02
	500	0,09	0,08	0,11	0,08	0,05	0,12	0,02	0,02	0,02	0,03	0,02	0,02
	1000	008	0,08	0,08	0,08	0,06	0,11	0,02	0,02	0,02	0,02	0,02	0,02
40	100	0,17	0,08	0,14	0,13	0,12	0,18	0,02	0,02	0,02	0,02	0,02	0,01
	250	0,16	0,07	0,11	0,08	0,05	0,13	0,02	0,02	0,02	0,02	0,02	0,01
	500	0,07	0,08	0,09	0,07	0,06	0,06	0,02	0,02	0,02	0,02	0,02	0,02
	1000	0,08	0,08	0,08	0,08	0,06	0,05	0,02	0,02	0,02	0,02	0,02	0,02
80	100	0,15	0,06	0,08	0,10	0,10	0,07	0,01	0,01	0,02	0,02	0,02	0,01
	250	0,12	0,05	0,11	0,10	0,10	0,07	0,01	0,01	0,02	0,02	0,02	0,01
	500	0,09	0,03	0,09	0,09	0,09	0,06	0,01	0,01	0,02	0,02	0,02	0,01
	1000	0,08	0,02	0,08	0,10	0,10	0,06	0,01	0,01	0,02	0,02	0,02	0,01

**madde sayısı **örneklem büyüklüğü*

Tablo 8'de görüldüğü gibi yedi yanıt kategorili koşullarda standart hata üç ve beş kategorili maddelere göre görece azalmıştır. En yüksek hata 10 madde ve 100 birey olduğunda kestirilmiştir. Hem birey sayısı hem de madde sayısı artışı hatada azalmaya işaret etmiştir. Yani hata değişimi büyük örneklemde ve uzun testlerde daha az gözlenmektedir.

SONUÇLAR ve TARTIŞMA

Parametrik ve parametrik olmayan MTK modellerinin karşılaştırılması kapsamında parametrik olmayan modellerin PMTK modellerinin daha iyi anlaşılmasını sağladığı, parametrik modellerin zayıf model-veri uyumu sağladığı durumlarda daha esnek bir çerçeve sunduğu, daha az sayıda madde ve birey ile çalışıldığı durumlarda model veri uyumu, güvenirlik, geçerlik konularında daha iyi sonuç sağladığı gerekçeleriyle daha kullanışlı olduğunu öne sürülmüştür (Junker ve Sijtsma, 2001). Bu araştırmada alt problemler parametrik (ATM) ve parametrik olmayan (MHM) MTK modellerinden kestirilen ait model veri uyumları, güvenirlik değerleri, madde parametrelerinin hataları birinci alt problem örneklem dağılımı normal, ikinci alt problem örneklem dağılımı -1,0 düzeyinde çarpık olduğunda bulguların verilmesi şeklinde düzenlenmiştir. Araştırmada dağılımın çarpıklık özelliği de bir bağımsız değişken olarak ele alındığı için bu bölümde alt problemlere ilişkin sonuçlar birlikte değerlendirilerek verilmiştir.

ATM'de normal ve çarpık dağılım kendi içinde değerlendirildiğinde örneklem büyüklüğü, testin uzunluğu ve testteki maddelerin kategori sayısı artışı değişkenlerinin, teste ait model veri uyumu değerlerini artırdığı görülmektedir. Bu artış normal dağılım koşullarında çarpık dağılımlara göre daha düzenlidir. Parametrik model veri uyumu için hesaplanan -2*log-benzerlik değerinin örneklem

ve parametre sayısına dayalı hesaplanması sonuçların tek başına yorumlanmasını güçleştirmektedir (Pampel, 2000). Bu bağlamda araştırmada ele alınan bağımsız değişkenlerin parametrik model uyumuna etkisine ilişkin genellemeye gidilmemiştir.

Teste ait model veri uyumunun değerlendirilmesine ilişkin sonuçlar incelendiğinde PoMTK'da model veri uyumu için parametrik modelden farklı olarak her kestirim için tek başına değerlendirilebilecek H katsayısının hesaplanması bir üstünlük olarak düşünülebilir (Sjitsma & Molenaar, 2002). Bu yönüyle PoMTK, bağımsız değişkenlerin etkisinin model veri uyumu üzerinde doğrudan gözlenmesine fırsat tanıyarak farklı koşullar için karşılaştırma yapma imkanı sağlamaktadır. PoMTK'da her üç dağılım koşulunda da örneklem büyüklüğü arttığında, testin uzunluğu ve testteki maddelerin kategori sayısı artışıyla birlikte, model uyumu için kullanılan ölçeklenebilirlik (H) katsayısı önemli değişim göstermemiş, tüm koşullarda bu katsayının görece birbirine yakın olduğu tespit edilmiştir. Bu durum, belirtilen değişkenlerin H katsayısının kestirim üzerinde etkili olmadığını işaret etmektedir. Dağılım normal olduğunda elde edilen H katsayılarının birbirine yakın olması ve dağılım çarpıklığı arttığında katsayıda görülen ani düşüşe dayanarak MHM'de model veri uyumunu en çok etkileyen değişkenin dağılımın çarpıklığı olduğu sonucuna ulaşılmıştır.

Testlere ait güvenirliklerin değerlendirilmesine ilişkin sonuçlar incelendiğinde MHM ile kestirimde genel olarak madde sayısı ve maddenin yanıt kategori sayısı arttıkça güvenirlik artmaktadır. Normal dağılım koşullarında ATM ile MHM'den paralel bulgular elde edilirken, dağılım çarpıklaştıkça MHM'nin ATM'den daha yüksek ya da daha düşük değerler verdiği görülmekte, bu bağlamda güvenirliğin dağılım şeklinden diğer bağımsız değişkenlere göre daha fazla etkilendiği söylenebilir. Örneklem büyüklüğü değişiminin ise etkili olmadığı görülmektedir.

PMTK ile yapılan kestirimlerde parametrelerin standart hatalarının örneklem büyüklüğü, madde sayısı ve maddelerin kategori sayısı arttıkça azaldığı görülmektedir. Dağılım çarpıklaştıkça hata düzeyi de artmaktadır. Standart hata değeri ayırıcılık parametrelerinde, eşik parametrelerine göre görece daha düşüktür. Eşik parametresi kestirimlerinin a parametrelerine göre daha hatalı olduğu görülmektedir. Bu durum *b* parametresinin her madde için ayırıcılık değerinden daha fazla sayıda olmasından kaynaklı olabilir. PoMTK ile yapılan kestirimlerde parametrelerin standart hatalarının parametrik kestirimlere oranla görece düşük olduğu; örneklem büyüklüğü, madde sayısı ve maddelerin kategori sayısı arttıkça azaldığı görülmektedir. Dağılım çarpıklaştıkça hatalarda parametrik kestirimdeki kadar büyük artış gözlenmemektedir.

Sonuç olarak, ATM ile parametre kestirimlerinin daha güvenilir ve daha az hatalı olması dağılımın normallik özelliği göstermesi ve en az 500 örneklem büyüklüğünün sağlanmasıyla gerçekleşmektedir. Madde sayısının 20, kategori sayısının en az 5 olmasının kestirimlerde parametre değişmezliğinin sağlanmasında etkili olan faktörler olduğu sonucuna ulaşılmıştır. Dolayısıyla araştırma koşulları örneklem büyüklüğü, madde sayısı ya da madde kategori sayısı değişimine imkân vermediği durumlarda tüm koşullardan daha az hatalı ve daha kararlı yapıda kestirim sunan MHM tercih edilmesi önerilebilir.

KAYNAKÇA

Ankenmann, R. D., & Stone, C. A. (1992, April). *A monte carlo study of marginal maximum likelihood parameter estimates fort he graded model*. Paper presented at the Annual Meeting of the Council on Measurement in Education, SanFrancisco, CA.

Davey, T., Nering, M. L., & Thompson, T. (1997). *Realistic simulation of item response data* (ACT Research Report Series 97-4, July). Retrived from: http://files.eric.ed.gov/fulltext/ED414297.pdf

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

DeMars, C. (2010). *Item response theory*. New York: Oxford University.

Dyehouse, M. A. (2009). *A comparison of model-data fit for parametric and nonparametric item response theory models using ordinal level ratings*. (Doctoral Dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3379330).

Doğan, N. (2002). *Klasik test kuramı ve örtük özellikler kuramının örneklemler bağlamında karşılaştırılması*. (Doktora tezi, Hacettepe Üniversitesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, Ankara).

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New Jersey, NJ: Lawrence Erlbaum Associates.

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement,* 32(3), 224–247. http://dx.doi.org/10.1177/0146621607302479.

Junker, B. W., & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement*, *25*, 211-220.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Academic Puslishers Group.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement, 12*, 38-47. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.1993.tb00542.x/pdf

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

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

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psikometrika, 62*, 331-347.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two and three parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement*, *6*, 249-260.

Kogar, H. (2015). Madde tepki kuramına ait parametrelerin ve model uyumlarının karşılaştırılması: Bir Monte Carlo çalışması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 6*(1), 142–157.

Liu, Y., & Maydeu-Olivares, A. (2014). Identifying the source of misfit in item response theory models. *Multivariate Behavioral Research, 49*, 354-371.

Maydeu-Olivares, A., & Joe, H. (2005). Further emprical results on parametric vs. nonparametric IRT modeling of Likert type personality data. *Multivariate Behavioral Research, 40*, 275-293.

Mokken, R. J. (1971). *A theory and procedure of scale analysis with applications in political research.* Berlin: Walter de Gruyter, Mouton.

Molenaar, I. W. (2001). Thirty years of nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 295-299. Retrieved from http://dx.doi.org/10.1177/01466210122032091

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

Pampel, F. C. (2000). *Logistic Regressio*n. Aprimer. Thousand Oaks, CA: Sage Publications.

Sijitsma, K., Emons, W. H. M., Bouwmeester, S., Nyklicek, I., Roorda, L. D. (2008). Nonparametric IRT analysis of Quality-of-Life Scales and its application to the World Health Organization Quality-of-Life Scale (WHOQOL-Bref). *Quality of Life Research, 17*(2), 275-290.

Sijitsma, K. (2005). Nonparametric item response theory models. Encyclopedia of Social Measurement, Volume 2.

Sijtsma, K., & Molenaar, I. W. (2002). Nonparametric item response theory and related topics. London: Sage.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52, 589-617. Retrieved from http://dx.doi.org/10.1016/j.paid.2010.02.011

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567-577.

Zenisky, R. K., Hambleton, S. G. Sireci. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement, 39*(4), 291-309. Retrieved from http://dx.doi.org/10.1111/j.1745-3984.2002.tb01144.x

EXTENDED ABSTRACT

Introduction

Item response theory (IRT) models are becoming more popular. Parametric item response theory (PIRT) models have been used more than nonparametric IRT models, especially in educational measurement. A reason for this may be that nonparametric IRT models were developed later than parametric IRT models. Nonparametric IRT models have several advantages over more-restrictive ***parametric*** IRT models. NIRT models are based on less-restrictive assumptions, thus they allow more items into the measurement tool. NIRT models offer diverse tools for item analysis that give information about the dimensionality of the data and the properties of the items; and provide item location and discrimination values, which have an easy interpretation for users. Several families of PIRT models for polytomous item scores have been proposed. The family of Graded Response Models (GRM) is suitable for analyzing ordered item scores collected by means of polytomous

response scales. And Monotone Homogeneity Model (MHM) is nonparametric extension of GRM. All known parametric IRT models for polytomous items are special cases of the nonparametric MHM. This means that any item set satisfying the requirements of a parametric IRT model for polytomous items also satisfies the requirements of the nonparametric MHM. In literature related to educational measurement, there is a need for comparison of both models in various conditions. To achieve this purpose, item response theory applications are developed.

The purpose of this study was to compare the standart errrors of item parameter estimates, reliability values, and model-data fit for test under different sample size, sample distribution, number of items, and number of categories of items conditions.

Method

The purpose of this study was to identify the effects of sample size, sample distribution, test length, the number of categories in items on the estimates of Graded Reponse Model and Monotone Homogenity Model for polytomously scored items. To achieve this aim, 192 simulation conditions which is composed of sample size, sample distribution, number of items, number of category items variables were designed. Estimates for GRM and Monotone Homogeneity Model were investigated through sample size (N = 100, 250, 500, 1000), sample distribution (normal, skewed), item number (10, 20, 40, 80), item category number (3, 5, 7) by calculating model-data fit, reliability values, item parameters, standard error of parameters values. The findings had been obtained through 25 replications.

With the purpose of estimating item parameters, a and b values for GRM; H and P values for MHM had been estimated. In order to evaluate the model-data fit for the test, -2log-likelihood value for GRM; scalability coefficients for MHM had been obtained. Reliability estimates of tests according to GRM and MHM were estimated. In addition, the standart erros of item parameter estimates had been calculated.

Results and Discussion

As a result of the research, it was found model-data fit values were affected by increase of variable number in Graded Response Models and since these values cannot be interpreted alone, it is difficult to make comparisons and generalization of those values. The practical calculation of model-data fit, and interpretation without need for another value in MHM provided superiority over GRM. Due to similar results in small samples and fewer items conditions and in larger samples and multiple items conditions, MHM had wider range of implementation.

It was also found that the reliability values gave similar results for both models. The increase in sample size had little effect on reliability estimates. In both models, values increased with the number of items and item response category. Reliability estimates decreased as the skewness of distribution increased.

Standard errors of item parameter estimates decreased as the number of items, number of categories of items, and sample size increased in Graded Response Model. Standart error of parameters were estimated higher at small samples and at conditions with fewer items. In Monotone Homogeneity Model, the standart error of parameters were lower than GRM at small sample and short test conditions and took close values to each other at all the conditions.

In conclusion, standart error of parameter estimates were smaller with normal distribution and larger sample sizes (with at least 500 examinees). And it was concluded that the number of items which was at least 20 and the number of categories which was at least 5 were effective factors in providing the parameter goodness in the estimates. Thus, in case the research conditions do not allow changes in sample size, the number of items, or number of categories, it can be suggested that MHM can be preferred since it provides less faulty and more stable estimations across all the conditions.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(4);373-390

Öğrenci Özelliklerinin Cinsiyete Dayalı Değişen Madde Fonksiyonuna Etkisi*

The Effect of Background Variables on Gender Related Differential Item Functioning

Nermin KIBRISLIOĞLU UYSAL** Kübra ATALAY KABASAKAL***

Öz

Araştırmada, sosyoekonomik düzey ve okuma becerisinin cinsiyete dayalı DMF'nin ortaya çıkmasına etkisi amaçlanmıştır. Bu kapsamda PISA 2015 fen uygulamasında yer alan maddelerin ele alınan dokuz ülkede cinsiyete göre değişen madde fonksiyonu (DMF) gösterip göstermediğinin belirlenmiştir. Araştırma kapsamında bilgisayar tabanlı uygulamada yer alan bir fen madde kümesi ele alınmıştır. Araştırmaya dahil edilen ülkeler bilgisayar tabanlı uygulamaya katılan ülkeler arasında başarı sıralamalarına göre seçilmiştir. Analizlerde çoklu göstergeler çoklu nedenler modeli (MIMIC) yöntemi kullanılmıştır. MIMIC ile DMF analizleri belirli gruplarda maddelerin örtük özelliği eşit şekilde ölçüp ölçmediğini belirlemek amacıyla, tam ve sınırlandırılmış modellerin uyumlarının karşılaştırılmasını içerir. Analizler iki aşamada gerçekleştirilmiştir. İlk aşamada maddelerin cinsiyet grupları arasında DMF gösterip göstermediği incelenmiştir. Daha sonra sosyoekonomik düzey ve okuma becerisi değişkenleri sırasıyla modele eklenerek söz konusu değişkenlerin cinsiyetten kaynaklı DMF'ye etkisi incelenmiştir. Araştırmanın bulgularına göre seçilen ülkelerin tamamında cinsiyetle ilişkili DMF'li maddeler yer almakta ve DMF'li madde sayıları 2 ile 6 arasında değişmektedir. Ülkelerin dördünde modele eklenen değişkenler cinsiyete ilişkin DMF'li madde sayısını manidar şekilde etkilememiştir. Ancak diğer dört ülkede söz konusu değişkenlerin modele eklenmesi DMF'li madde sayısını azaltmıştır. Cinsiyete dayalı DMF'li madde sayısını azaltan değişkenler her bir ülke kapsamında ayrı ayrı tartışılmıştır.

Anahtar Kelimeler: Sosyoekonomik düzey, okuma başarısı, değişen madde fonksiyonu, MIMIC

Abstract

In this study, the effect of socioeconomic status and reading ability, on the presence of gender-related DIF were examined. For this purpose, presence of differential item functioning (DIF) between gender groups in PISA 2015 science items in nine selected countries were detected. One cluster of science items from computer-based assessment (CBA) was taken into consideration. The countries were selected among the ones that implemented CBA, on the basis of their rank in science achievement. Multiple Indicator Multiple Causes method (MIMIC) was used for DIF analyses. DIF analysis in the MIMIC involves fit comparisons of both full and reduced models to determine if the items can measure the latent trait equally among the specified groups. The MIMIC analysis was conducted in two steps. First, the items were tested for exhibiting DIF between gender groups. Then the socioeconomic status and the reading ability were added to the model to test gender-related DIF items and their effects, respectively. According to the results of the study, gender-related DIF appeared in all of the selected countries with between two and six items. In four of the countries, none of the selected variables significantly affected the presence of gender-related DIF. Instead, in the remaining countries, the number of gender-related

* Bu çalışma "12th International Conference on Social Sciences" kongresinde sözlü bildiri olarak sunulmuştur.

** Arş. Gör., Hacettepe Üniversitesi, Eğitim Bilimleri Bölümü, Ankara-Türkiye, e-posta: nkibrislioglu@hacettepe.edu.tr, ORCID ID: orcid.org/0000-0002-9592-469X

*** Yrd. Doç. Dr., Hacettepe Üniversitesi, Eğitim Bilimleri Bölümü, Ankara-Türkiye, e-posta: katalay@hacettepe.edu.tr, ORCID ID: orcid.org/0000-0002-3580-5568

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, Cilt 8, Sayı 4, Kış 2017, 373-390.
Journal of Measurement and Evaluation in Education and Psychology, Vol. 8, Issue 4, Winter 2017, 373-390.

DOI: 10.21031/epod.333451

Geliş Tarihi: 08.08.2017
Kabul Tarihi: 18.10.2017

DIF items was decreased by adding selected variables to the model. The effects of variables which reduced the number of gender-related DIF items were discussed within each country.

Keywords: Socioeconomic status, reading ability, differential item functioning, MIMIC

GİRİŞ

Uluslararası Öğrenci Değerlendirme Programı (PISA), Ekonomik İşbirliği ve Kalkınma Örgütü (OECD) tarafından 2000 yılında uygulanmaya başlayan ve üç yıl aralıklarla tekrarlanan okuma, matematik ve fen bilimleri temel alanlarındaki yeterliği ölçmeyi amaçlayan büyük ölçekli bir araştırmadır. PISA, 15 yaş grubu öğrencilerinin bu üç temel alanda okul programlarından ziyade, gerçek yaşamda karşılarına çıkabilecek sorunlardaki bilgi ve becerilerini kullanabilme kapasitelerini değerlendirir. PISA sonuçları ülkelerin eğitim politikalarını önemli düzeyde etkilemesi, ülkelerin kendi başarılarını diğer ülkelerle karşılaştırmalı olarak değerlendirmesine olanak sağlaması bakımından oldukça önemlidir. Bu karşılaştırmalar yapılırken farklı kültürlerden ve dillerden gelen bireyler, bilişsel ya da duyuşsal özellikleri bakımından sıralanır. Bu nedenlerle PISA sonuçlarının farklı ülkeler arasında karşılaştırılabilirliği ile ilgi endişeler son yıllarda artmaya başlamıştır.

Ölçme araçlarında istenilen, aracın psikometrik özelliklerinin tüm cevaplayıcılar için aynı olmasıdır (Woods, 2009). Bir ölçme aracındaki maddelerin psikometrik özelliklerinin aynı yeteneğe sahip farklı gruplar için aynı olmaması durumu yanlılık olarak tanımlanır. Psikometride madde düzeyinde yanlılığın test edildiği istatistiksel analizlerden biri değişen madde fonksiyonudur (DMF). DMF, aynı yetenek düzeyine sahip farklı gruptaki bireylerin maddeyi doğru cevaplama olasılıklarının farklılaşmasıdır. Madde Tepki Kuramı (MTK) çerçevesinde, DMF gösteren bir madde farklı gruplar için farklı kategori cevap fonksiyonlarına sahip olacaktır. Diğer bir ifadeyle, örtük değişken (θ) üzerinde eşleşen iki farklı grubun bireylerinin, aynı maddeye aynı yanıtı verme olasılıkları aynı olmayacaktır. Gruplar cinsiyet, etnik köken ya da deneysel duruma göre tanımlanabilir. Genellikle çoğunluğu oluşturan ya da testin orijinalinin geliştirildiği grup referans grup, diğer grup odak grup olarak isimlendirilir (Zumbo, 1999). Referans ve odak gruba dayalı olarak uygulanan DMF belirleme yöntemleri ile grup farklılıkları tespit edilir. DMF tek biçimli ve tek biçimli olmayan olmak üzere iki farklı şekilde görülebilir (Camilli ve Shepard, 1994; Mellenbergh, 1989). Tek biçimli DMF, madde yanıt fonksiyonlarının iki grup arasında aynı olmadığı ve herhangi bir çakışma olmadığı durumda gözlenir, yani tüm yetenek düzeylerinde (θ) bir grubun daha yüksek tepki vermesinin daha olası olması durumudur. Eğer madde yanıt fonksiyonları iki grup için çakışırsa tek biçimli olmayan DMF görünür (Camilli ve Shephard, 1994). Bir testte DMF gösteren maddelerin bulunması yanlılığa neden olacağından testin grup karşılaştırmalarına ilişkin geçerliğini düşürebilir (Dorans ve Holland, 1993). Yanlılık bir sistematik hata nedeni olduğundan, sistematik hatanın artması yalnızca gruplar arası farkların belirlenebilirliğini azaltmaz, aynı zamanda farkların yönü ve büyüklüğünün yanlış yorumlanmasına da neden olabilir (OECD, 2015). Dolayısıyla, PISA gibi uluslararası değerlendirme araştırmalarının temelde ülkeler arası yeterlikleri karşılaştırdığı düşünüldüğünde, dilsel ve kültürel olarak DMF gösteren maddeleri içermemesinin sağlanması gerekmektedir.

Alan yazında pek çok çalışmada farklı kültürlerle ve dillerle yapılan değerlendirmelerde DMF'nin ortaya çıktığı bulunmuştur (Allalouf, Hambelton, ve Sireci, 1999; Budgell, Raju, ve Quartetti, 1995; Ercikan ve Koh, 2005; Grisaya ve Monseur, 2007; Huang, Wilwon ve Wang, 2014; Sireci ve Swaminathan, 1996; Wu ve Ercikan, 2006). Örneğin, Grisaya ve Monseur (2007) çalışmalarında PISA 2000 uygulaması kapsamında yer alan okuma maddelerinin DMF gösterip göstermediğini 47 ülke arasında incelemişlerdir. Araştırmanın bulguları, çevirinin DMF'nin ortaya çıkmasında çok etkili olduğunu ve farklı diller arasında ortaya çıkan DMF'li madde sayısının aynı dili konuşan farklı kültürler arasındaki DMF'li madde sayısından çok daha fazla olduğunu göstermiştir. Benzer şekilde Asil ve Gelbal (2012), PISA 2006 kapsamında yer alan öğrenci anketlerinin kültürler ve diller arası eşdeğerliğini Avustralya, Yeni Zelanda, Amerika Birleşik Devletleri ve Türkiye örneklemleri üzerinde incelemişlerdir. Araştırmanın bulguları ülkeler arasında dil değiştikçe ve kültürel özellikler farklılaştıkça DMF'li madde sayılarının değiştiğini göstermiştir. Huang, Wilson ve Wang (2014) PISA

2006 fen uygulamasının kültürler ve diller arasında DMF gösterip göstermediğini Birleşik Devletler ve Kanada, Çin-Hong Kong ve Çin, Birleşik Devletler ve Çin ülkeleri arasında incelemişlerdir. Araştırmanın sonuçları en çok DMF gösteren maddenin Birleşik Devletler ve Çin arasında ortaya çıktığını, İngilizce konuşan Birleşik Devletler ve Kanada arasında ise DMF'nin göz ardı edilebilir düzeyde olduğunu göstermiştir. Bu çalışmalar dil farklılığının DMF'nin ortaya çıkmasında oldukça etkili bir faktör olduğunu göstermektedir.

Bu çalışmalara ek olarak alan yazında kültürler ve diller arasında ortaya çıkan DMF'yi etkileyebileceği düşünülen öğrenci özelliklerinden kaynaklanabilecek etmenler de ele alınmıştır. Wu ve Ercikan (2006) TIMSS 1999 uygulaması kapsamında Birleşik Devletler ve Tayvan arasında maddelerin DMF gösterip göstermediğini lojistik regresyon yöntemi ile incelemişlerdir. Çalışmalarında öğrencilerin okul dışı fazladan çalışma saatlerinin bir DMF kaynağı olup olmamasının yanı sıra, bu durumun ülkeler arasında ortaya çıkan DMF'li madde sayılarına bir etkisinin olup olmadığını incelemişlerdir. Araştırmanın bulguları okul dışı fazladan çalışma saatinin bir DMF kaynağı olduğunu ve modele eklendiğinde ülkeler arasında ortaya çıkan DMF'li madde sayısını %29 oranında azaldığını göstermiştir. Husin (2014) PISA 2012 matematik maddelerinin Malezya örnekleminde ana dilde ve İngilizce uygulamaları arasında DMF gösterip göstermediğini, ayrıca DMF'nin varlığının öğrencilerin okuma becerilerinden ve sosyoekonomik düzeylerinden (SED) etkilenip etkilemediğini lojistik regresyon yöntemi ile incelemiştir. Araştırmanın bulguları testin uygulandığı dilin yanı sıra madde özelliklerinin, SED'in ve okuma becerisinin de genel test performansını etkilediğini ve modele bu değişkenlerin eklenmesinin DMF gösteren madde sayısında azalmaya neden olduğunu göstermiştir.

DMF'nin önemli düzeyde ortaya çıktığı bir diğer grup ise cinsiyet gruplarıdır. Alan yazında pek çok çalışmada (Atalay Kabasakal ve Kelecioğlu, 2012; Lan, 2014; Le, 2009; Lyons-Thomas, Sandilands ve Ercikan, 2014) uluslararası uygulanan sınavlarda cinsiyet grupları arasında DMF ortaya çıktığı raporlanmıştır. Bunlara ek olarak cinsiyete ilişkin DMF analizlerinde madde özelliklerinin DMF'nin yönüne etkisine ilişkin araştırmalar da yer almaktadır. Lyons-Thomas, Sandilands ve Ercikan (2014) çalışmalarında PISA 2009 uygulamasında yer alan matematik maddelerinin Kanada, Çin-Şangay, Finlandiya ve Türkiye örneklemlerinde cinsiyete göre DMF gösterip göstermediğini madde türleri bazında incelemiştir. Araştırmanın bulgularına göre genel olarak çoktan seçmeli maddeler erkekler lehine, açık uçlu maddeler ise kadınlar lehine DMF göstermiştir. Benzer şekilde Le (2009) PISA 2006 deneme uygulamasında yer alan fen maddelerinin cinsiyetler arasında DMF gösterip göstermediğini ülkelere ve madde özelliklerine göre incelemiştir. Araştırmada madde özellikleri olarak PISA çerçevesinde belirlenen odak, içerik, yeterlik ve bilimsel bilgi olmak üzere dört madde özelliği ve madde türleri temel alınmıştır. Araştırmanın bulgularına göre çoktan seçmeli ve kapalı uçlu maddeler, evrensel odaklı maddeler, çevre ile ilgili maddeler ve bilimsel bir olguyu açıklamayı gerektiren maddeler çoğunlukla erkekler lehine DMF göstermektedir. Diğer yandan doğal kaynaklarla ilgili maddeler, bilimsel bilgiyi tanımlama, bilimsel kanıt ve veri kullanma gerektiren maddeler ise kadınlar lehine DMF göstermiştir.

Alan yazın genel olarak incelendiğinde uluslararası uygulamalarda kültüre, dile ve cinsiyete dayalı DMF çalışmalarının çok sayıda olduğu görülmektedir. Özellikle PISA uygulamalarının uluslararası geçerliğini test etmeye yönelik ülkeler arasında DMF ve ölçme değişmezliği çalışmaları oldukça yaygındır. Öte yandan PISA 2015 teknik raporunda ülkeler arası DMF'nin nasıl belirlendiği ve DMF'li maddelerin nasıl kontrol edildiğine ilişkin bilgilere yer verilmiştir (OECD, 2015). Bu bağlamda OECD'nin PISA 2015 uygulamasında ülkelere göre olası yanlılıkları göz önünde bulundurduğu sonucuna varılabilir. Diğer yandan raporda cinsiyete yönelik herhangi bir DMF çalışmasından bahsedilmemektedir. Ayrıca alan yazında cinsiyete yönelik DMF çalışmalarında çoğunlukla madde türünün cinsiyet DMF'sinin yönü ile etkileşiminden bahsedilmiş; ancak kültüre yönelik DMF çalışmalarında olduğu gibi öğrenci özelliklerinin (SED, okuma becerisi gibi) cinsiyete göre DMF'ye etkisine ilişkin bir çalışmaya rastlanmamıştır. Bu nedenlerle bu araştırmada cinsiyete dayalı DMF ve DMF'yi etkileyebilecek olası öğrenci özellikleri farklı ülkeler bağlamında incelenmiştir. Araştırma kapsamında cinsiyete dayalı DMF'ye SED ve okuma becerisinin etkisi incelenmiştir.

PISA 2015'te SED; aile eğitimi, aile mesleği ve evdeki olanaklar değişkenlerinden temel bileşenler analiziyle oluşturulan bir indekstir (OECD, 2015). SED'in öğrencilerin eğitimsel kazanımları elde etmesinde ve akademik başarıları üzerinde önemli bir belirleyici faktör olduğu düşünülmektedir (Barr, 2015; Coleman ve arkadaşları, 1966; Hecht, Burgess, Torgesen, Wagner ve Rashotte, 2000; White, 1982). Yüksek SED ve düşük SED'in karşılaştırıldığı bazı çalışmalarda eğitimsel çıktıların zayıflığının nedenleri arasında aile eğitim düzeyi düşüklüğü, okul kaynaklarının azlığı, aile katılımın azlığı bulunmuştur (Schmidt, Cogan, & McKnight, 2011). PISA uygulamalarında SED bakımdan avantajlı olan öğrencilerin ve okulların dezavantajlı akranlarına göre çok daha yüksek puanlar aldığı gözlenmiştir (OECD, 2015). Her ne kadar düşük SED doğrudan düşük başarıya neden olur gibi bir çıkarım yapmak doğru olmasa da, sosyoekonomik düzeyin başarı üzerindeki etkisi yadsınamaz. Bu bağlamda cinsiyete bağlı DMF analizlerinde sosyoekonomik düzey değişkeninin modele dahil edilmesiyle sosyoekonomik düzeyinde bir DMF kaynağı olup olmadığının ve cinsiyete dayalı DMF'nin varlığını etkileyip etkilemediğinin değerlendirilmesine olanak sağlamaktadır.

PISA, ölçtüğü temel alanın yanı sıra öğrencilerin günlük hayatta karşılarına çıkabilecek çeşitli yazılı materyalleri anlama ve yorumlama düzeylerini de değerlendirir. PISA'da okuma becerilerinde ele alınan bilişsel yeterlikler; bir metni basit olarak çözümlemenin ötesinde metni çözümlerken uygun stratejileri kullanma becerisi ve bu becerinin farkında olma gibi üst düzey bilişsel becerileri de içine alır. Günümüz bilgi çağında bireyin okuma, okuduğunu anlama, yorumlama, muhakeme etme gibi yeterliklere sahip olması beklenmektedir. Öğrencilerin okuma becerilerinin fen ve matematik başarılarını etkilediği bilinmektedir (Demps ve Onwuegbuzie, 2001; Nolen, 2003; O'Reilly & McNamara, 2007). Cromley (2009) 32 ülke arasında okuma ve fen başarısı arasında ortalama korelasyon katsayılarını PISA 2000 verisinde 0,84, PISA 2003 verisinde 0,805 ve PISA 2006 veri setinde 0,819 bulmuştur. Her üç veri setinde de okuma-fen arasındaki korelasyon katsayıları ülkelere göre değişmekle birlikte, düşük okuma puanı olan ülkelerde en düşük katsayılar elde edilmiştir. Çoğu öğretmen ve program geliştirme uzmanı bazı matematik maddelerinin öğrencilerin doğru olarak cevaplayabilmeleri için yüksek düzeyde okuma becerisine sahip olması gerektiği belirtmektedir (NCES, 2003). Büyük ölçekli sınavlarda genellikle fen ve matematik maddelerinin cevaplanabilmesi için ciddi bir okuma yükü gerekir. Eğer öğrencinin okuma becerisi düşük bir düzeydeyse öğrenci kendi gerçek başarısını gösteremeyebilir. Diğer bir ifadeyle öğrenciler verilen problem durumunu doğru bir şekilde matematiksel formata dönüştürmede başarısız olursa kendi performansını gösteremeyecektir. Walker, Zhang ve Surber (2008) yaptıkları çalışmada çok düzeyli DMF çerçevesinde öğrencilerin matematik maddelerindeki performansının okuma başarıları ile ilişkili olduğunu, düşük okuma becerisine sahip öğrencilerin matematik maddelerini doğru cevaplama olasılıklarının düşük olduğunu bulmuşlardır. Alanyazına dayanarak cinsiyete bağlı DMF analizlerinde okuma başarısı değişkeninin modele dahil edilmesiyle bu değişkenin bir DMF kaynağı olup olmadığı ve cinsiyete dayalı DMF'nin varlığını etkileyip etkilemediği incelenmiştir.

PISA 2015 uygulamasında fen bilimleri temel alan olarak alınmıştır. Bu bağlamda uygulamada yer alan okuma ve matematik testlerinde yer alan maddeleri önceki yıllarda uygulanan maddelerden oluşurken, fen testlerinde hem önceki yıllarda kullanılan maddeler hem de ilk defa uygulanan yeni maddeler yer almaktadır. PISA 2015 uygulamasında diğer yıllardan farklı olarak madde kümeleri yer almaktadır ve bu kümeler 96 farklı forma dağıtılmıştır. Ayrıca ülkelerin bir kısmına kağıt kalem formları uygulanırken diğer kısmında uygulama bilgisayar tabanlı yapılmıştır. Formların ilk 30'u kağıt kalem testi şeklinde geri kalanı bilgisayar tabanlı uygulamalarda kullanılmıştır. Uygulamada 18 fen bilimleri madde kümesi (S1-S18) yer almaktadır ve bu kümelerin 6 tanesi (S1-S6) daha önce uygulanmış maddelerden, 12 tanesi ise (S7-S18) yeni uygulanan maddelerden oluşmaktadır (OECD, 2015). Araştırma kapsamında bilgisayar tabanlı uygulamada yer alan ve ilk defa uygulanan maddelerden oluşan S12 fen bilimleri madde kümesi ele alınmıştır. Söz konusu maddelerin cinsiyete bağlı DMF gösterip göstermediği dokuz ülke kapsamında incelenmiştir. Ülkeler seçiminde 2015 uygulaması fen bilimleri başarı sıraları bir ölçüt olarak kullanılmış ve OECD ortalaması, ortalama üzeri ve ortalama altı olmak üzere her bir düzeyden üçer ülke araştırmaya dahil edilmiştir. Bu bağlamda araştırma OECD ortalaması üzerinde yer alan Japonya, Finlandiya ve Hong Kong; ortalamada yer alan Birleşik Devletler, Fransa ve İsveç; ortalama altında ise Türkiye, Brezilya ve İsrail ülkeleri ile gerçekleştirilmiştir.

Araştırmanın Amacı

Araştırmanın amacı PISA 2015 uygulaması fen bilimleri testinde yer alan S12 madde kümesinin farklı kültürler içerisinde cinsiyete bağlı DMF gösterip göstermediğinin incelenmesi; maddelerin bilgi ve içerik alanları ile cinsiyet DMF'si arasındaki ilişkinin incelenmesi ve bununla beraber öğrencilerin sosyoekonomik düzeylerinin ve okuma becerilerinin cinsiyete dayalı DMF'yi etkileyebilecek birer DMF kaynağı olup olmadığının belirlenmesidir. Bu kapsamda araştırmada aşağıdaki araştırma sorularına cevap aranmaktadır.

1. S12 madde kümesinde yer alan fen bilimleri maddeleri Japonya, Finlandiya, Hong Kong, Birleşik Devletler, Fransa, İsveç, Türkiye, Brezilya ve İsrail verilerinde cinsiyete göre DMF göstermekte midir?
2. Maddelerin bilgi ve içerik alanları ile cinsiyet DMF'si arasında bir ilişki var mıdır?
3. Her bir ülke içinde öğrencilerin sosyoekonomik düzeylerinin grup üyeliğinin ötesinde DMF'ye bir etkisi var mıdır?
4. Her bir ülke içinde öğrencilerin okuma becerilerinin grup üyeliği ve sosyoekonomik düzeyin ötesinde DMF'ye bir etkisi var mıdır?

DMF belirleme yöntemleri

DMF belirlemede çok sayıda yöntem geliştirilmiştir. Bu yöntemler arasında Mantel-Haenszel yöntemi (Holland ve Thayer, 1988), standartlaştırma yöntemi (Dorans ve Kulick, 1986), lojistik regresyon yöntemi (Swaminathan ve Rogers, 1990), SIBTEST yöntemi (Shealy ve Stout, 1993), Lord'un ki-kare testi (Lord, 1980; Wright ve Stone, 1979), olabilirlik oranı testi (MTK-OO; Thissen, Steinberg, ve Wainer, 1988; Wang ve Yeh, 2003), çoklu göstergeler çoklu nedenler modeli (MIMIC; Finch, 2005; Oort, 1998) sayılabilir. MH, standartlaştırma yöntemi ve lojistik regresyon yöntemleri gözlenen puana dayalı yöntemlerdir. Yaygın olarak kullanılmalarına karşın örtük özellik yerine toplam puan kullanılmasının doğruluğu tartışmalıdır. Araştırmalar cevaplar iki ya da üç parametreli MTK gibi daha karmaşık bir modelle üretildiğinde bu yöntemlerin DMF'yi yanlış belirleyebileceğini göstermiştir (Meredith & Millsap, 1992; Millsap & Meredith, 1992).

MIMIC yöntemi diğer yöntemlere göre daha yeni bir yöntemdir. MIMIC, DMF belirlemede kullanılan örtük özelliğe dayalı bir yapısal eşitlik modeli türüdür. Son yıllarda farklı koşullarda MIMIC yönteminin etkililiği ile ilgili çok sayıda çalışma yapılmış ve bu yöntem DMF belirleme çalışmalarında sıklıkla kullanılmaya başlanmıştır (Finch, 2005; Fleishman, Spector, & Altman, 2002; Gallo, Anthony, & Muthe´n, 1994; Glöckner-Rist & Hoitjink, 2003; Levine et al., 2003; MacIntosh & Hashim, 2003; Muthe´n, 1985; Muthe´n, Kao, & Burstein, 1991; Shih ve Wang, 2009; Oort, 1998).

MIMIC modelinin avantajlarından biri nedensel gösterge olarak tanımlanan en az bir gözlenen değişkenin örtük değişkeni yordamasıdır (Joreskog ve Goldberger, 1975). Ayrıca MIMIC dışsal değişkenlerle MTK modellerine genişletilebilen bir doğrulayıcı faktör analizi yöntemi olduğundan madde parametreleri kestirimi, DMF belirleme, MTK'nın tek boyutluluk ve koşullu bağımsızlığı gibi varsayımlarını esnetme gibi önemli madde analizi konularına da eş zamanlı olarak değinebilmektedir. Bunlara ek olarak MIMIC yönteminde gruplama değişkeni sürekli ya da süreksiz bir değer olabilir. Dolayısıyla, sadece tek bir süreksiz gruplama değişkeniyle DMF analizi yapmaya izin veren geleneksel DMF belirleme yöntemlerine (MH, SIBTEST ya da MTK-OO gibi) göre daha esnek olduğu söylenebilir (Wang, Shih ve Yang, 2009). MIMIC yönteminde çok boyutlu maddeler veya faktörler kolayca modellenebilir ve birden fazla grup arasında DMF incelenirken yorumlaması daha kolaydır. Ayrıca MIMIC modeli DMF test edilirken fazladan kovaryant değişkenlerin modele eklenmesine ve kovaryant değişkenlerin sürekli olmasına da olanak sağlar. Diğer yandan MIMIC yöntemi tek biçimli DMF'yi belirlemeye hassastır (Woods, 2009). Ancak son yıllarda tek biçimli olmayan DMF belirlemek için de MIMIC yönteminin kullanıldığı çalışmalar yapılmaktadır (Chun, 2014; Woods ve Grimm, 2011).

Bu çalışmada DMF analizleri MIMIC yöntemi ile gerçekleştirildiğinden yalnızca bu yöntemin hesaplanmasına detaylı olarak yer verilmiştir.

MIMIC ile DMF Analizi

MIMIC modeli hem ölçme modelinden hem de yapısal modelden oluşur. MIMIC yönteminin ölçme modeline ilişkin denklem aşağıdaki gibidir:

$$y_i^* = \lambda_i \eta + \beta_i z_k + \varepsilon_i, \quad (1)$$

Denklem (1)'de y_i^*, i maddesi için testin ölçmeyi amaçladığı θ için gizil cevap değişkeni, z cinsiyet, etnik köken gibi faktör analitik modelde yer alan DMF ile ilişkili olduğu düşünülen gruplama değişkeni, λ_i ise faktör yüküdür ve MTK kapsamında i maddesi için eğim parametresidir. ε_i, tesadüfi hatadır ve normal dağılıma sahiptir.β_i ise grup değişkeni ve z'nin y_i^* üzerindeki etkisidir. $\beta_i = 0$ise i maddesi z gruplama değişkenleri üzerinde homojendir yani i maddesinde DMF yoktur, eğer $\beta_i \neq 0$ ise z'nin y_i^* üzerinde doğrudan etkisi vardır. Yani i maddesi z gruplama değişkenine göre DMF göstermektedir. y_i doğrudan gözlenemediği için test maddeleri $y_i^{*\prime}$yi ölçebilmek amacıyla $y_i^{*\prime}$yi iki kategorili madde cevaplarına (y_i) dönüştürür:

$$y_i = \begin{cases} 1, & y_i^* > \tau_{i,} \\ 0, & diğer \end{cases} \quad (2)$$

Denklem (2)' de τ_i, madde güçlüğüne ilişkin sınır parametresini göstermektedir.

Yapısal modelde, θ ile z gruplama değişkeni θ = $\gamma' z$ + ξ denkleminde olduğu gibi doğrudan ilişkilidir. γ', θ üzerindeki grup farklılıklarını tanımlayan regresyon katsayılarının bir vektörüdür ve genellikle DMF analizlerinde etki olarak kullanılır (Ackerman, 1992; Camilli, 1993); ξ ortalaması sıfır olan z'den bağımsız olarak normal dağıldığı varsayılan artıktır.

MTK'da üç parametreli model için (Hambelton, Swaminathan ve Rogers, 1991), n bireyinin i maddesini doğru yanıtlama olasılığı aşağıdaki gibi tanımlanmıştır:

$$P(U_{ni} = 1|\theta_n) = c_i + (1 - c_i)\frac{\exp(a_i(\theta_n + b_i))}{1 + \exp(a_i(\theta_n + b_i))} \quad (3)$$

Denklem (3)'de U_{ni}, n bireyinin i maddesine verdiği cevap değişkeni, θ_n, n bireyinin örtük özelliğinin düzeyi a_i, b_i ve c_i ise sırayla i maddesinin madde ayırıcılık, madde güçlük ve şans parametrelerini göstermektedir. Her bir i maddesi için $c_i = 0$ olduğunda üç parametreli model iki parametreli modele dönüşmektedir. Her bir i maddesi için $c_i = 0$ ve $a_i = 0$ olduğunda üç parametreli model bir parametreli model halini almaktadır.

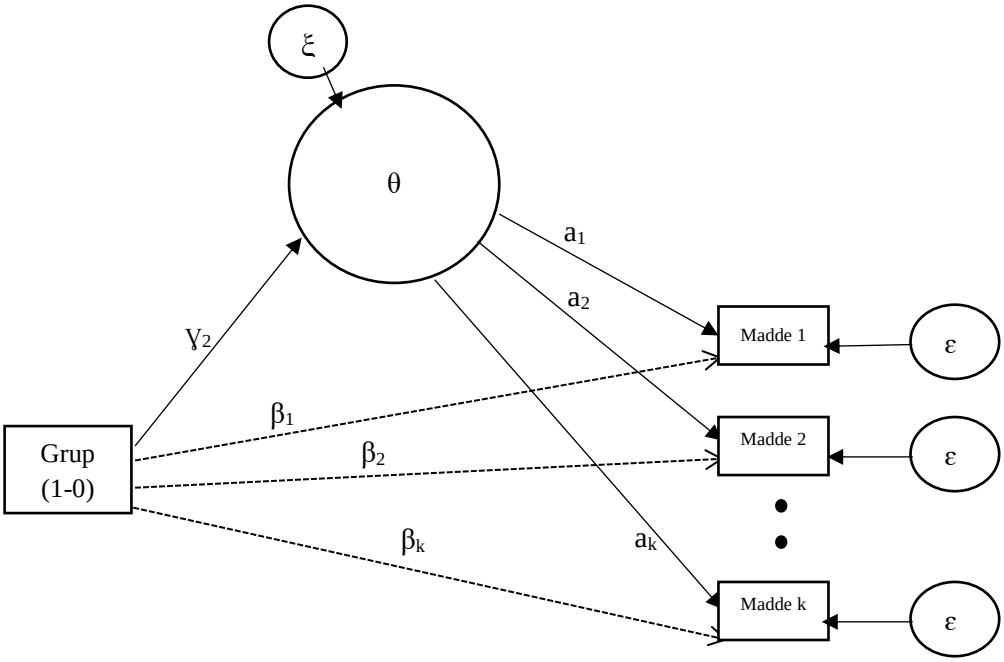
$$a_i^* = c(1 - \lambda_i^2 \psi)^{-1/2} \sigma_{00}^{1/2} \quad (4)$$

$$b_i^* = [(\tau_i - \beta_i' z)\lambda_i^{-1} - \mu_0]\sigma_{00}^{1/2} \quad (5)$$

Ψ, ξ'nın varyansı; μ_0 ve σ_{00} ise θ'nın ortalaması ve varyansıdır.

Referans grup için $z = 0$, odak grup için $z = 1$ olmak üzere tek bir gruplama değişkeni olduğunda Denklem (5)'de görüleceği üzere gruplar arası madde güçlüğü farkı $\beta_i' \lambda_i^{-1} \sigma_\theta^{-1}$'e eşit olacaktır. σ_θ, θ'nın standart sapmasıdır ve birim olarak sınırlandırıldığında, $\beta_i' \lambda_i^{-1}$tek biçimli DMF'nin etki büyüklüğüdür (Shih ve Wang, 2009). DMF testi için temel MIMIC modeli Şekil 1'de gösterilmiştir.

Şekil 1, madde cevaplarının DMF testi için gruplama değişkeni üzerinde regresyonunu gösteren standart bir tek boyutlu MTK modelidir. Bu modelde θ üzerindeki farklılıklar kontrol altına alındığında eğer gruplama değişkeninin madde cevaplarını yordaması anlamlı ise DMF gözlenir. Bu model tek biçimli DMF modeli olduğu için madde ayırt edicilikleri dolaylı olarak değişmezdir (Woods, 2009).



Şekil 1. DMF testi için temel MIMIC modeli

γ: örtük değişkende ortalamalar arası farkı gösteren regresyon katsayısı, β_i: i maddesinin eşik parametresinin grup değişkeni arasında farkını gösteren regresyon katsayısı (i=1,2..k), a_i: ayırıcılık parametersi, ε_i: i maddesi için ölçme hatası, ξ: θ için artık. (Woods, 2009, s.5)

DMF belirleme çalışmalarında referans ve odak grubu aynı ölçeğe yerleştirmek amacıyla eşleştirme değişkeni kullanılmaktadır. Bu sayede maddelerin DMF gösterip göstermediği değerlendirilir (Wainer, Sireci ve Thissen, 1991). Eşleşme değişkeni içsel ve ya dışsal olarak iki türlü seçilebilir. (Welch ve Miller, 1995). Eğer DMF analizi yapılacak maddeler aynı zamanda eşleşme değişkeni olarak kullanılıyorsa içsel değişken, eğer eşleşme değişkeni test dışından maddelerden oluşan bir set ise dışsal değişkendir. Pratikte, DMF analizi yapılan test ile aynı örtük özelliği ölçen DMF içermeyen madde setleri bulmak çoğunlukla zor olduğu için eşleşme değişkeni genellikle içsel olarak belirlenir (Shih ve Wang, 2009). Gerçek koşullarda testler kusursuz değildir ve DMF'li maddeler içerirler. Bu nedenle eşleştirme değişkeninin DMF içerme ihtimali vardır. Ancak, DMF değerlendirmesinin doğru olarak yapılabilmesi için eşleşme değişkenin DMF içermemesi gereklidir (Holland ve Thayer, 1988; Lord, 1980). Eğer DMF analizlerinde DMF içeren bir eşleşme değişkeni kullanılırsa, referans ve odak grupların performansı yanlı bir ölçüte dayalı olarak karşılaştırılacaktır ve sonuçlar yanıltıcı olacaktır. Diğer yandan eğer eşleşme değişkenleri DMF içermezse, bu durumda da DMF analizi gerekli olmayacaktır. Dolayısıyla bu durum bir kısır döngü oluşturmaktadır (Shih ve Wang, 2009).

Finch (2005) MIMIC yöntemini iki parametreli ve üç parametreli MTK modellerinde tek biçimli DMF'yi belirlemede Mantel-Haenszel, MTK-OO ve SIBTEST yöntemleri ile karşılaştırmıştır. Yapılan çalışmada test uzunluğu, örneklem büyüklüğü, gruplar arası ortalama yetenek farkı, DMF etki büyüklüğü ve eşleşme değişkenindeki DMF düzeyi değişkenleri manipüle edilmiştir. Araştırma sonuçları MIMIC yönteminin diğer yöntemler kadar uygulanabilir olduğunu göstermiştir. Ayrıca eşleşme değişkenindeki DMF düzeyinden diğer yöntemler büyük ölçüde etkilenirken, MIMIC yönteminde sadece I.Tip hatada az miktarda azalma, güçte ise küçük bir düşme gözlenmiştir. Eşleşme değişkeninde DMF gösteren madde olması ihtimali olduğunda, MIMIC yöntemin DMF içeren

maddelerden diğer yöntemlere göre daha az etkilenmesi nedeniyle kullanılmasının uygun olduğunu belirtmiştir.

Bir testte bir ya da daha fazla sayıda DMF içeren maddenin varlığı durumunda testteki diğer maddeler için DMF araştırmalarının sonuçları doğru olmayabilir. Örneğin; DMF içermeyen maddeler yanlışlıkla DMF'li olarak saptanabilir. Bu durum testin 1. tip hatasını istenmeyen şekilde arttırır (Clauser, Mazor ve Hambleton, 1993). İçsel eşleşme değişkenindeki DMF etkisini sınırlamanın bir yolu ölçek arıtma yöntemi kullanmaktır. Ölçek arıtma yöntemleri DMF belirleme yöntemlerine uyarlanmıştır. Örnek olarak iki-aşamalı ya da iteratif Mantel-Haenszel (Holland ve Thayer, 1988), iteratif Mantel yöntemi, iteratif genelleştirilmiş Mantel yöntemi (Wang ve Su, 2004a, 2004b), iteratif lojistik regresyon (French ve Maller, 2007), ve iteratif MTK yöntemi (Candell ve Drasgow,1988) verilebilir. Aynı prensipteki ölçek arındırma yöntemi MIMIC için de aşağıdaki şekilde uyarlanmıştır:

1. MIMIC ile 1. madde DIF açısından değerlendirilir ve testteki diğer tüm maddeler (DMF analizlerinin yürütüldüğü maddeler) çalışan madde olarak kullanılır. İlk adım testteki tüm maddeler için tekrarlanır.
2. Önceki adımlarda DMF içermeyen maddeler çalışan madde olarak tanımlanır ve tüm maddeler için DMF analizi yapılır.
3. İki ardışık iterasyonda aynı DMF'li madde kümesi belirleninceye kadar 3. adım tekrar edilir (Shih ve Wang, 2009).

YÖNTEM

Araştırma kapsamında PISA 2015 uygulaması S12 madde kümesinde yer alan fen maddelerinin cinsiyet grupları arasında DMF gösterip göstermediği ve cinsiyete dayalı DMF'yi SED ve okuma başarısı değişkenlerinin etkileyip etkilemediği incelenmiştir. Bu bağlamda araştırma PISA 2015 uygulaması fen maddelerinin geçerliğini test etmeye yönelik betimsel bir araştırmadır.

PISA 2015 uygulamasında yer alan fen maddeleri yeterlik, bilgi ve içerik olmak üzere üç boyutta sınıflandırılmıştır. Yeterlilik bağlamında maddeler bilimsel araştırma desenleme ve değerlendirme (Evaluate and design scientific enquiry), olguları bilimsel yollarla açıklama, veri ve kanıtları bilimsel olarak yorumlama olmak üzere üç grupta sınıflandırılmıştır. Benzer şekilde bilgi bağlamında alan (content), epistemik ve süreç bilgisi olmak üzere üç grupta; içerik bağlamında dünya ve uzay, yaşam ve fiziksel olmak üzere üç grupta sınıflandırılmıştır (OECD, 2015). Araştırma kapsamına dahil edilen S12 madde kümesinde beş başlıkta 17 madde yer almaktadır. Maddelerin yeterlilik, bilgi ve içerik alanına göre dağılımları Tablo 1'de verilmiştir.

Tablo 1. S12 Fen Maddelerinin Özellikleri

Konu	Madde kodları*	Yeterlik	Bilgi	İçerik
Göktaşları ve Kraterler	CS641Q01S-M1	Olguları bilimsel yollarla açıklama	Alan	Fiziksel
	CS641Q02S-M2	Olguları bilimsel yollarla açıklama	Alan	Dünya ve Uzay
	CS641Q03S-M3	Veri ve kanıtları bilimsel olarak yorumlama	Alan	Dünya ve Uzay
	CS641Q04S-M4	Veri ve kanıtları bilimsel olarak yorumlama	Alan	Dünya ve Uzay
Deniz ortamında sesler	CS626Q01S-M5	Olguları bilimsel yollarla açıklama	Alan	Fiziksel
	CS626Q02S-M6	Bilimsel araştırma desenleme ve değerlendirme	Süreç	Fiziksel
	CS626Q03S-M7	Veri ve kanıtları bilimsel olarak yorumlama	Süreç	Fiziksel
	DS626Q04C-M8	Veri ve kanıtları bilimsel olarak yorumlama	Süreç	Yaşam
Eğim-Yüzey İncelemesi	DS637Q01C-M9	Olguları bilimsel yollarla açıklama	Epistemik	Dünya ve Uzay
	CS637Q02S-M10	Olguları bilimsel yollarla açıklama	Epistemik	Dünya ve Uzay
	DS637Q05C-M11	Veri ve kanıtları bilimsel olarak yorumlama	Epistemik	Dünya ve Uzay
Beyin Kontrollü Robotik	DS610Q01C-M12	Olguları bilimsel yollarla açıklama	Alan	Yaşam
	CS610Q02S-M13	Olguları bilimsel yollarla açıklama	Alan	Yaşam
	CS610Q04S-M14	Bilimsel araştırma desenleme ve değerlendirme	Alan	Yaşam
Sürdürülebilir Balık Çiftliği	CS601Q01S-M15	Olguları bilimsel yollarla açıklama	Alan	Yaşam
	CS601Q02S-M16	Veri ve kanıtları bilimsel olarak yorumlama	Alan	Yaşam
	CS601Q04S-M17	Olguları bilimsel yollarla açıklama	Alan	Fiziksel

*Madde kodları araştırmacılardan tarafından M1-M17 şeklinde tekrar kodlanmıştır.

Örneklem

Araştırma kapsamında seçilen fen bilimleri uygulaması S12 madde kümesinde yer alan maddelerin cinsiyete göre DMF gösterip göstermediği Japonya, Finlandiya, Hong Kong, Birleşik Devletler, Fransa, İsveç, Türkiye, Brezilya ve İsrail örneklemlerinde incelenmiştir. Dolayısıyla araştırmanın evrenini söz konusu dokuz ülkedeki 15 yaş grubu öğrencileri oluşturmaktadır.

Araştırmanın örneklemini dokuz ülkede PISA uygulamasına katılan ve S12 madde kümesinin dahil olduğu kitapçıkları cevaplayan araştırmanın kapsamı dahilinde ele alınan değişkenlerden eksik verisi bulunmayan öğrenciler oluşturmaktadır. Öğrencilerin ülkelere göre dağılımları Tablo 2'de özetlenmiştir. PISA uygulamalarında örnekleme yöntemi olarak tabakalı örnekleme kullanılmaktadır ve uygulama formları öğrencilere rastgele olarak dağıtılmaktadır. Tüm ülkeler için madde kümesinin uygulandığı örneklem rastgele olduğundan belirli bir madde kümesinin uygulandığı bireyleri örnekleme dahil etmek herhangi bir örnekleme yanlılığı oluşturmayacaktır.

Tablo 2. Ülkelere Ve Cinsiyetlere Göre Öğrenci Frekansları

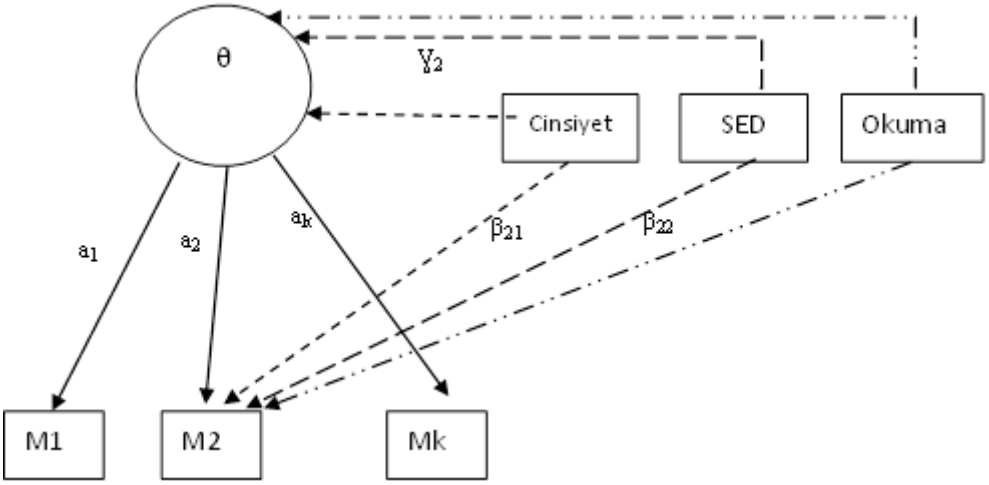
	Kadın		Erkek		Toplam	
Ülke	f	%	f	%	f	%
Japonya	600	48,7	632	51,3	1232	100
Finlandiya	599	48,4	639	51,6	1238	100
Hong Kong	495	50,3	490	49,7	985	100
Birleşik Devletler	505	50,2	500	49,8	1005	100
Fransa	557	50,5	546	49,5	1103	100
İsveç	469	50,3	464	49,7	993	100
Türkiye	511	49,9	513	50,1	1024	100
Brezilya	1281	51,0	1232	49,0	2513	100
İsrail	654	57,1	492	42,9	1146	100
Toplam	5671	0,51	5508	0,49	11239	100

Tablo 2 incelendiğine S12 madde kümesinin uygulandığı örneklem büyüklüğünün ülkeler arasında genel olarak benzer dağıldığı görülmektedir. Yalnızca Brezilya örneklemi diğer ülkelerden daha büyük görünmektedir. Genel olarak incelendiğinde İsrail hariç ülkelerde kadın ve erkek grupları arasında fark en fazla %2 kadardır. Dolayısıyla örneklem büyüklükleri cinsiyet grupları arasında benzerdir.

Verilerin Analizi

Verilerin analizinde cinsiyet grupları arasında olası DMF içeren maddeleri belirlemeye yönelik MIMIC yöntemi kullanılmıştır. Analizler iki aşamada gerçekleştirilmiştir. İlk aşamada maddelerin her bir ülke içinde cinsiyet grupları arasında DMF gösterip göstermediği incelenmiştir. Analizlerde odak grup olarak kızlar alınmıştır. Daha sonra modele sırasıyla SED ve okuma becerisi değişkenleri eklenmiş ve söz konusu değişkenlerin cinsiyete bağlı DMF'yi nasıl etkilediği incelenmiştir. Araştırma kapsamında SED değişkeni olarak PISA kapsamında hesaplanan indeks değişkeni kullanılmıştır. Öğrencilerin okuma becerilerine yönelik ise uygulama kapsamında okuma testi sonucunda hesaplanan okuma becerisi puanları kullanılmıştır.

MIMIC yöntemi ile DMF analizleri, maddelerin örtük özelliği gruplar arasında eşit düzeyde ölçüp ölçmediğini belirlemek amacıyla tam model ile sınırlandırılmış model arasındaki uyumun karşılaştırılmasını içermektedir. Model karşılaştırmalarında sınırlandırılmış temel model kullanılmıştır. Araştırma kapsamında ele alınan tam model Şekil 2'de verilmiştir.



Şekil 2. Sınırlandırılmış temel model yaklaşımına yönelik tek biçimli DMF için tam model

Verilerin analizinde öncelikle maddelerin cinsiyet gruplarına göre DMF gösterip göstermediği incelenmiştir. Sonrasında DMF gösteren maddeler için modele öncelikle SED, ardından okuma becerisi değişkenleri eklenmiştir. DMF'li maddeler belirlenirken ilk olarak modelde cinsiyet değişkeninin manidarlığına bakılmıştır. Sonrasında ise tam model ile sınırlandırılmış model uyumları arasında anlamlı bir farklılık bulunup bulunmadığı incelenmiştir. Veriler 1-0 şeklinde puanlandığı için parametre kestirimlerinde ağırlıklandırılmış en küçük kareler (WLS) yöntemi kullanılmıştır. WLS kestirimlerinde model karşılaştırmalarında ki-kare değerleri arasındaki farklar, ki-kare dağılımı göstermediğinden ki-kare fark testi yapılmaz. Bu nedenle model uyumları arasındaki fark değerlendirilirken program tarafından sağlanan DIFFTEST fark testi kullanılmıştır (Muthen ve Muthen, 2010).

BULGULAR

Araştırmanın birinci alt problemine yönelik PISA 2015 fen bilimleri uygulamasında yer alan S12 madde kümesinde yer alan maddelerin Japonya, Finlandiya, Hong Kong, Birleşik Devletler, Fransa, İsveç, Türkiye, Brezilya ve İsrail örneklemlerinde cinsiyete göre DMF gösterip göstermediği incelenmiştir. DMF analizi bulguları Tablo 3'te özetlenmiştir.

Tablo 3. S12 Madde Kümesinde Yer Alan Soruların Cinsiyete Göre DMF Analizi Sonuçları

		Göktaşları ve Kraterler				Deniz Ortamında Sesler				Eğim-Yüzey İncelemesi			Beyin Kontrollü Robotik			Sürdürülebilir Balık Çiftliği		
		M01	M02	M03	M04	M05	M06	M07	M08	M09	M10	M11	M12	M13	M14	M15	M16	M17
Japonya	β	0,149	0,5	-0,346	-0,309	0,226	0,158	-0,09	-0,381	-0,06	0,074	-0,033	0,201	0,036	-0,088	0,104	-0,188	0,095
	p	0,154	**0,000**	0,069	0,232	**0,004**	0,071	0,4	**0,000**	0,445	0,327	0,786	0,517	0,781	0,356	0,478	0,126	0,266
Hong Kong	β	0,031	0,474	-0,315	0,036	0,367	-0,012	-0,021	-0,243	-0,038	-0,011	0,051	0,084	0,075	-0,154	0,485	-0,165	-0,069
	p	0,684	**0,000**	**0,013**	0,851	**0,000**	0,919	0,83	**0,035**	0,678	0,892	0,618	0,356	0,696	0,171	**0,008**	0,365	0,501
Finlandiya	β	0,192	0,113	-0,13	-0,774	0,259	0,219	0,108	-0,473	-0,247	-0,068	0,251	0,142	-0,042	-0,006	0,276	-0,258	-0,108
	p	0,036	0,275	0,363	0,425	**0,030**	0,067	0,451	**0,006**	0,12	0,848	**0,003**	0,689	0,798	0,954	**0,015**	0,051	0,649
Birleşik Devletler	β	0,256	-0,577	-0,159	0,007	0,303	-0,138	0,148	-0,487	-0,205	0,089	0,132	-0,057	-0,157	0,187	0,492	-0,095	-0,259
	p	**0,005**	0,24	0,327	0,965	**0,001**	0,184	0,386	0,075	0,111	0,27	0,23	0,596	0,296	0,099	**0,010**	0,351	0,249
Fransa	β.	0,204	0,068	-0,331	-0,109	0,135	0,062	-0,163	-0,064	-0,416	-0,233	0,442	0,226	0,146	-0,171	0,142	-0,156	0,198
	p	0,196	0,674	0,141	0,694	0,093	0,472	0,4	0,571	0,462	**0,003**	**0,013**	0,250	0,295	0,286	0,39	0,265	0,414
İsveç	β	0,168	0,102	0,035	-0,465	0,354	-0,055	0,064	-0,469	-0,246	-0,096	0,298	3,412	0,408	0,245	0,081	0,08	-0,323
	p	0,061	0,387	0,812	0,105	**0,000**	0,646	0,587	**0,009**	0,087	0,342	**0,008**	0,413	**0,01**	**0,008**	0,57	0,745	0,498
Türkiye	β	0,217	0,183	-0,195	-0,235	0,302	0,007	0,029	-0,35	0,04	0,019	0,015	0,05	-0,014	-0,004	0,276	-0,046	-0,212
	p	**0,005**	**0,027**	**0,033**	**0,006**	**0,000**	0,926	0,704	**0,005**	0,798	0,851	0,872	0,874	0,896	0,972	0,245	0,59	0,726
Brezilya	β	0,067	0,172	-0,049	-0,595	0,163	0,09	-0,024	-0,296	0,193	-0,016	-0,02	0,095	-0,021	-0,077	0,181	-0,01	-0,069
	p	0,153	**0,000**	0,674	0,322	**0,002**	0,119	0,703	**0,003**	0,461	0,766	0,75	0,232	0,833	0,373	0,176	0,899	0,369
İsrail	β	0,29	-0,031	-0,244	-0,168	0,149	-0,192	0,026	0,061	-0,023	-0,05	0,157	0,281	-0,36	0,033	0,444	-0,247	0,016
	p	**0,002**	0,803	0,118	0,28	0,063	0,125	0,816	0,547	0,827	0,554	0,223	**0,004**	0,035	0,793	**0,008**	**0,012**	0,87

*Konu başlıkları araştırmacılar tarafından çevrilmiştir.

Tablo 3 incelendiğinde Türkiye verisinde 6, Hong Kong, İsveç ve İsrail verilerinde 5'er, Finlandiya verisinde 4, Japonya, Birleşik Devletler ve Brezilya verilerinde 3'er, Fransa verisinde 2 maddede cinsiyete dayalı DMF ortaya çıktığı görülmektedir. Genel olarak incelendiğinde 17 maddeden 13'ü farklı ülkelerde cinsiyete göre DMF göstermiştir. Madde özellikleri incelendiğinde olguları bilimsel yollarla açıklama yeterliği gerektiren 1., 2., 5., 12., ve 15. maddeler ve bilimsel araştırma desenleme ve değerlendirme yeterliği gerektiren 14. madde erkekler lehine, veri ve kanıtları bilimsel olarak yorumlama yeterliği gerektiren 3., 4., 8. ve 16. maddeler kadınlar lehine DMF göstermiştir. Bu bulgular daha önce Le (2009) tarafından yapılan çalışma ile benzerlik göstermektedir. Diğer yandan olguları bilimsel yollarla açıklama yeterliği gerektiren 10. madde yalnızca Fransa örnekleminde kadınlar lehine DMF göstermiştir. 12. madde ise İsveç örnekleminde erkekler lehine DMF gösterirken İsrail örnekleminde kadınlar lehine DMF göstermiştir. Benzer şekilde veri ve kanıtları bilimsel olarak yorumlama yeterliği gerektiren 11. madde Fransa ve İsveç örneklemlerinde erkekler lehine DMF göstermiştir. Cinsiyete göre DMF'nin anlamlı olduğu her bir madde için modele SED değişkeni eklenmiş ve SED değişkenin anlamlı etkisinin olduğu analiz sonuçları Tablo 4'te verilmiştir.

Tablo 4. Cinsiyet ve SED DMF Modeli Sonuçları

			Türkiye	İsveç	
			M04	M11	M13
Cinsiyet Modeli	Cinsiyet	β	-0,235	0,298	0,408
		p	0,006	0,008	0,010
Cinsiyet ve SED Modeli	Cinsiyet	β	-0,196	0,317	0,308
		p	-0,231	0,180	0,104
	SED	β	0,006	-0,254	-0,213
		p	0,031	0,728	0,146

Tablo 4 incelendiğinde modele SED değişkeni eklendiğinde cinsiyete göre DMF gösteren maddelerden Türkiye ve İsveç örneklemlerinde yer alan 3 maddenin etkilendiği gözlenmektedir. 4. madde başlangıçta yalnızca Türkiye örnekleminde kadınlar lehine DMF göstermektedir. Modele SED değişkeninin eklenmesiyle söz konusu maddenin cinsiyete ilişkin DMF'sinin ortadan kalktığı ve SED değişkenine dayalı bir DMF'nin ortaya çıktığı görülmektedir. 11. ve 13. maddeler ise başlangıçta İsveç örnekleminde erkekler lehine DMF göstermektedir. Modele SED değişkeninin eklenmesiyle söz konusu maddelerde cinsiyete dayalı DMF ortadan kalkmıştır. Her bir ülke için modele SED değişkeninin yanı sıra okuma değişkeni eklendiğinde anlamlı etki elde edilen analiz sonuçları ise Tablo 5'te verilmiştir.

Tablo 5. Cinsiyet, SED ve Okuma Becerisi DMF Modeli Sonuçları

			Türkiye	Finlandiya	Hong Kong	
			M03	M08	M03	M08
Cinsiyet Modeli	Cinsiyet	β	-0.195	-0.473	-0.315	-0.243
		p	0.033	0.006	0.013	0.035
Cinsiyet, SED ve Okuma Modeli	Cinsiyet	β	-0.188	-0.817	-0.581	-0.5
		p	0.075	0.262	0.083	0.191
	SED	β	-0.088	-0.07	0.19	-0.231
		p	0.036	0.715	0.007	0.105
	Okuma	β	0.000	-0.008	-0.004	-0.006
		p	0.859	0.605	0.394	0.494

Tablo 5'te görüldüğü gibi modele SED değişkeni ile birlikte okuma becerisi değişkeni eklendiğinde başlangıçta Türkiye ve Finlandiya'da cinsiyet DMF'si gösteren birer maddenin, Hong Kong'da ise iki maddenin cinsiyete dayalı DMF göstermediği gözlenmektedir. 1. maddenin Türkiye ve Hong Kong örneklemlerinde SED değişkenine göre DMF gösterdiği görülmektedir. Genel olarak incelendiğinde maddelerin başlangıçta kadınlar lehine DMF gösterdiği modele okuma değişkeninin eklenmesiyle DMF'nin ortadan kalktığı görülmektedir. Alan yazında genel olarak kadınların okuma becerisi açısından erkeklerden daha yüksek puanlar aldığı raporlanmaktadır (Hyde ve Lin, 1988; Logan ve Johnstone, 2009). Bu bağlamda söz konusu maddelerde başlangıçta ortaya çıkan cinsiyet DMF'sinin okuma becerileri arasındaki farklardan kaynaklanabileceği sonucuna varılabilir. Söz konusu maddelerde yalnızca üç ülkede okuma becerisinin cinsiyet DMF'si üzerinde etkili olması ise ülkelerde konuşulan dillerin özelliklerinden kaynaklanabilir.

Bulgular genel olarak incelendiğinde söz konusu madde kümesi içinde maddelerin 14'ünün farklı ülkelerde cinsiyete dayalı DMF gösterdiği görülmektedir. DMF'nin yönü incelendiğinde ise olguları bilimsel yollarla açıklama yeterliği gerektiren maddelerin çoğunlukla erkekler lehine; veri ve kanıtları bilimsel olarak yorumlama yeterliği gerektiren maddelerin ise çoğunlukla kadınlar lehine DMF gösterdiği görülmektedir. Ülkeler karşılaştırıldığında ise en çok cinsiyete bağlı DMF'li madde içeren ülkeler Türkiye, Hong Kong, İsveç ve İsrail'dir. En az olan ülke ise Fransa'dır. Modele SED değişkeninin eklenmesi Türkiye ve İsveç'te cinsiyete dayalı DMF gösteren madde sayısını azaltırken diğer ülkelerde herhangi bir manidar etki göstermemiştir. Modele okuma değişkeninin eklenmesi ise Hong Kong, Finlandiya ve Türkiye'de cinsiyete dayalı DMF gösteren madde sayısını azaltmıştır. Modele eklenen değişkenlerin Japonya, Birleşik Devletler, Fransa, Brezilya ve İsrail'de cinsiyet DMF'sine herhangi bir etkisi olmamıştır.

SONUÇLAR ve TARTIŞMA

Uluslararası karşılaştırmalarda temel alınan PISA gibi büyük ölçekli sınavlarda bireylerin demografik özelliklerinin test puanlarına etkisinin incelenmesi oldukça önemlidir. Test puanları bireylerin demografik özelliklerinden bağımsız olmalıdır; ancak yapılan birçok çalışmada cinsiyetin büyük ölçekli sınavlar için bir DMF kaynağı olduğu bilinmektedir. Bu bağlamda araştırmanın amaçlarından biri PISA 2015 fen uygulamasında yer alan S12 madde kümesindeki maddelerin dokuz ülkede cinsiyete bağlı DMF gösterip göstermediğinin belirlenmesidir. Araştırma kapsamına dahil edilen ülkelerin tamamında cinsiyete dayalı DMF gösteren maddeler olduğu görülmektedir. Benzer bulgular uluslararası yapılan sınavlara yönelik pek çok çalışmada da raporlanmıştır (Atalay Kabasakal ve Kelecioğlu, 2012; Lan, 2014; Le, 2009; Lyons-Thomas, Sandilands ve Ercikan, 2014).

Araştırma kapsamında ayrıca cinsiyete dayalı DMF gösteren maddelerin özellikleri de incelenmiştir. Araştırmanın sonuçlarına göre Le'nin (2009) bulgularına paralel olarak olguları bilimsel yollarla açıklama yeterliği gerektiren maddelerin çoğunlukla erkekler lehine, veri ve kanıtları bilimsel olarak yorumlama yeterliği gerektiren maddelerin ise çoğunlukla kadınlar lehine DMF göstermektedir. Yalnızca İsrail örnekleminde olguları bilimsel yollarla açıklama yeterliği gerektiren bir madde kadınlar lehine, veri ve kanıtları bilimsel olarak yorumlama yeterliği gerektiren bir madde de İsveç ve Fransa örnekleminde erkekler lehine DMF göstermiştir. Bu bağlamda madde türünün cinsiyet ile etkileşiminin olabileceği ancak söz konusu etkileşimin farklı kültürlerde farklı şekillerde ortaya çıkabileceği göz önünde bulundurulmalıdır.

Araştırmalarda yalnızca maddelerin gruplar arasında DMF gösterip göstermediğinin belirlenmesi yeterli değildir. Farklı cinsiyetteki öğrenciler için testin yapı geçerliğinin sağlanamamasının altında yatan nedenlerin araştırılması önemlidir. Bu bağlamda araştırmanın ikinci temel amacı SED ve okuma becerisi değişkenlerinin cinsiyete dayalı DMF analizlerinde dolaylı bir etkisinin olup olmadığının belirlenmesidir. SED ve okuma becerisinin başarı ile ilişkisinin yüksek olduğu bilindiğinden araştırmaya dahil edilen maddelerin cinsiyet DMF'sini daha doğru belirleyebilmek amacıyla belirtilen değişkenler modele eklenmiştir. Araştırma kapsamında söz konusu değişkenler her bir ülke için ayrı ayrı değerlendirilmiştir. Söz konusu değişkenlerin modele eklenmesi dört ülkede DMF'li madde sayısını azaltmıştır. Bu bulgular SED ve okuma becerisi değişkenlerinin

cinsiyetle ilişkili DMF'li madde sayısının azaltılması ile ilişkili olduğunu göstermektedir. SED değişkeni "Göktaşları ve Kraterler, Eğim-Yüzey İncelemesi, Beyin Kontrollü Robotik" konularına ilişkin üç maddede Türkiye ve İsveç'te cinsiyete dayalı DMF üzerinde anlamlı bir etkiye neden olmuştur. Okuma becerisi ise "Göktaşları ve Kraterler" ve "Deniz ortamında sesler" konularına ilişkin iki maddede Türkiye, Finlandiya ve Hong Kong'da cinsiyete dayalı DMF üzerinde anlamlı bir etkiye neden olmuştur. Okuma becerisinin modele eklenmesinin kadınlar lehine ortaya çıkmış olan DMF'yi ortadan kaldırmış olması önemli bir bulgu olarak görülebilir. Gelecek çalışmalarda okuma becerisinin cinsiyet DMF'sine etkisi daha kapsamlı bir şekilde değerlendirilerek ortaya çıkan cinsiyet DMF'sinin okuma becerileri arasındaki farktan kaynaklanıp kaynaklanmadığı incelenebilir. Bu çalışmanın sonuçları genel olarak test performansındaki farklılıkların bir kısmının cinsiyetle açıklanabileceğini, ayrıca sosyoekonomik düzey ve okuma becerisinin de bu farklılığa bir katkısı olabileceğini göstermiştir.

DMF, maddelerin sabit bir özelliği değildir. Karşılaştırma yapılan gruplara bağlı olarak değişebilir. Dolayısıyla farklı ülkelerde cinsiyete dayalı DMF'yi etkileyen farklı özelliklerin olması beklenen bir durumdur. Araştırma kapsamına dahil edilen SED değişkeni Türkiye ve İsveç'te; okuma becerisi ise Türkiye, Finlandiya, ve Hong Kong'da cinsiyete dayalı DMF'yi etkilerken diğer ülkelerde her iki değişkenin herhangi bir etkisi gözlenmemiştir. Gelecek çalışmalarda farklı kültürlerde cinsiyete dayalı DMF'yi etkileyebilecek farklı değişkenler incelenebilir. Ayrıca madde türlerinin cinsiyet DMF'sinin yönü ile olan ilişkisinin de kültürlere göre farklılaşabileceği gözlemlenmiştir. Dolayısıyla değişkenlerin etkisinin her bir kültür bağlamında değerlendirilmesi de oldukça önemlidir.

Çalışmada gruplar arası puan karşılaştırılabilirliğini etkileyebileceği düşünülen değişkenler DMF modeline eklenmiştir. Gelecek çalışmalarda öğrenci başarısı ile ilişkili olduğu bilinen ikinci dilde test alma, konu alanı, madde türü gibi diğer dışsal değişkenlerin de gruplar arası karşılaştırmaları nasıl etkileneceği incelenebilir. Ayrıca çalışmada SED ve okuma becerisi sürekli değişkenler olarak ele alınmıştır. Ancak alan yazında genellikle DMF analizlerinde kesikli değişkenler kullanılmaktadır. Bu bağlamda söz konusu değişkenlerin sürekli veya kesikli biçimde ele alınmasının etkisi de incelenebilir.

KAYNAKÇA

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67–91. doi:10.1111/j.1745-3984.1992.tb00368.x

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement,36(3),* 185–198.

Asil, M. ve Gelbal, S. (2012). PISA öğrenci anketinin kültürler arası eşdeğerliği. *Eğitim ve Bilim, 37*(166), 236-249.

Atalay Kabasakal, K. ve Kelecioğlu, H. (2012). PISA 2006 öğrenci anketinde yer alan maddelerin değişen madde fonksiyonu açısından incelenmesi. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 45*(2), 77-96.

Barr, A. B. (2015). Family socioeconomic status, family health, and changes in students' math achievement across high school: A mediational model. *Social Science & Medicine, 140,* 27-34.

Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement, 19*(4), 309-321.

Camilli, G. (1993). The case against item bias techniques based on internal criteria: Do item bias procedures obscure test fairness issues? The use of differential item functioning statistics: A discussion of current practice and future implications. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 397-413). Hillsdale, NJ: Lawrence Erlbaum.

Camilli, G., & Shepard, L. A. (1994*). Methods for identifying biased test items.* London: Sage.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12,* 253-260.

Chun, S. (2014). *Using MIMIC methods to detect and identify sources of DIF among multiple groups*. Unpublished master thesis. University of South Florida, USA.

Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel–Haenszel procedure. *Applied Measurement in Education, 6,* 269-279.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. (1966). *Equality of educational opportunity.* DC: US Department of Health, Education & Welfare. Office of Education (OE-38001 and supp.), 1066-5684.

Cromley, G. J. (2009). Reading achievement and science proficiency: International comparisons from the programme on international student assessment. *Reading Psychology, 30*(2), 89-118.

Demps, D. L., & Onwuegbuzie, A. J. (2001). The relationship between eighthgrade reading scores and achievement on the Georgia High School Graduation Test. *Research in the Schools, 8*(2), 1–9.

Dorans, N. J., & Holland, P.W. (1993). *DIF detection and description: Mantel-Haenszel and standardization.* In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hilsdale: Lawrence Erlbaum Associates.

Dorans, N. J., & 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,* 355-368.

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing, 5*(1), 23 - 35.

Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *Journal of Gerontology: Social Sciences,57*(5), 275-284.

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,* 278-295.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement, 67,* 373-393.

Gallo, J. J., Anthony, J. C., & Muthe´n, B. O. (1994). Age differences in the symptoms of depression: A latent trait analysis. *Journal of Gerontology: Psychological Sciences,49,* 251-264.

Glöckner-Rist, A., & Hoitjink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling, 10,* 544-565.

Grisaya, A., & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation, 33*(1), 69-86.

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

Hecht, S. A., Burgess, S. R., Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2000). Explaining social class differences in growth of reading skills from beginning kindergarten through fourth-grade: The role of phonological awareness, rate of access, and print knowledge. *Reading and Writing, 12*(1),99-128.

Holland, W. P., & Thayer, D. T. (1988). *Differential item performance and the Mantel–Haenszel procedure.* In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Huang, X., Wilson, M., & Wang, L. (2016). Exploring plausible causes of differential item functioning in the PISA science assessment: Language, curriculum or culture. *Educational Psychology,36*(2), 378-390. http://dx.doi.org/10.1080/01443410.2014.946890

Husin, M. (2014). *Assesiıng mathematical competence in second language: Exploring DIF evidences from PISA Malaysian data*. Unpublished master thesis, University of Wisconsion, Milwaukee.

Hyde, J. S., & Lin, M. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin, 104(1),* 53-69. Doi: http://dx.doi.org/10.1037/0033-2909.104.1.53

Joreskog, K., & Goldberger, A. S. (1975). Estimation of a model with a multiple indicators and multiple causes of a single latent variable. *Journal of American Statistical Association,70,* 631-639.

Lan, M. C. (2014). *Exploring gender differential item functioning (DIF) in eight grade mathematics items for the United States and Taiwan.* Unpublished doctoral disstertation. University of Washington.

Le, L. T. (2009). Investigating gender differential item functioning across countries and test languages for PISA science items. *International Journal of Testing, 9*(2), 122-133. http://dx.doi.org/10.1080/15305050902880769

Levine, D. W., Bowen, D. J., Kaplan, R. M., Kripke, D. F., Naughton, M. J., & Shumaker, S. A. (2003). Factor structure and measurement invariance of the women's health initiative ınsomnia rating scale. *Psychological Assessment, 15,* 123-136.

Logan S., & Johnstone, R. (2009). Gender differences in reading ability and attitudes: examining where these differences lie. *Journal of Research in Reading, 32*(2), 199-214. doi: 10.1111/j.1467-9817.2008.01389.x

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ:Lawrence Erlbaum.

Lyons-Thomas, J., Sandilands, D. D., & Ercikan, K. (2014). Gender differential item functioning in mathematics in four international jurisdictions. *Education and Science, 39*(172), 20-32.

MacIntosh, R., & Hashim, S. (2003). Variance estimation for converting MIMIC model parameters to IRT parameters in DIF analysis. *Applied Psychological Measurement, 27,* 372-379.

Mellenbergh, J. G. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*(2),127-143.

Meredith, W., & Millsap, R. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57*(2), 289-311.

Millsap, R., & Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias. *Applied Psychological Measurement, 16*(4), 389-402.

Muthén, B. O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics, 10,* 121-132.

Muthén, B. O., Kao, C. F., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement, 28*(1), 1-22.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus statistical analysis with latent variables user's guide* (6th Edition). Los Angeles, CA: Muthén & Muthén.

Nolen, S. B. (2003). Learning environment, motivation, and achievement in high school science. *Journal of Research in Science Teaching, 40*(4), 347–368.

OECD (2015). PISA 2015 technical report. OECD: http://www.oecd.org/pisa/data/2015-technical-report/

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling, 5,* 107-124.

O'Reilly, T., & McNamara, D. S. (2007). The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional "high-stakes" measures of high school students' science achievement. *American Educational Research Journal, 44*(1), 161–196.

NCES (2003). NAEP validity studies: An agenda for NEAP validity studies (Report No. 2003-07). Retrieved from https://nces.ed.gov/pubs2003/200307.pdf

Schmidt, W. H., Cogan, L. S., & McKnight, C. C. (2011). Equality of educational opportunity: Myth or reality in U.S. schooling?.*American Educator 34*(4), 12-19.

Shealy, R. T., & Stout, W. F. (1993). A model-biased standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58,* 159-194.

Shih, C. L., & 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.

Sireci, S. G., & Swaminathan, H. (1996). Evaluating translation equivalence: So what's the big DIF? Paper presented at the AERA, Ellenville, NY.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

Thissen, D., Steinberg, L., & Wainer, H. (1988). *Use of item response theory in the study of group differences in trace lines.* In H. Wainer, & H. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum.

Wainer, H., Sireci, S., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28,* 197-219.

Walker, C. M., Zhang, B., & Surber, J. (2008) Using a multidimensional differential item functioning framework to determine if reading ability affects student performance in mathematics. *Applied Measurement in Education, 21*(2), 162-181, doi:10.1080/08957340801926201

Wang, W. C., Shih, C. L., & Yang, C. C. (2009). The MIMIC method with scale purification procedure for detecting differential item functioning. *Educational and Psychological Measurement, 69*(5), 713-731.

Wang, W. C., & Su, Y. H. (2004a). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education, 17,* 113-144.

Wang, W. C., & Su, Y. H. (2004b). Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement, 28,* 450-480.

Wang, W. C., & Yeh, Y. L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27,* 479-498.

Welch, C. J., & Miller, T. R. (1995). Assessing differential item functioning in direct writing assessments: Problems and an example. *Journal of Educational Measurement, 32,* 163-178.

White, K. R. (1982). The relation between socioeconomic status and academic achievement, *Psychological Bulletin, 91*(3), 461-481

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: MESA.

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44*(1), 1-27.

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement, 35*(5) 339–361. http://dx.doi.org/10.1177/0146621611405984.

Wu, A. D., & Ercikan, K. (2006). Using multiple-variable matching to identify cultural sources of differential item functioning. *International Journal of Testing, 6(*3), 287-300.

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*. Ottawa on Directorate of Human Resources Research and Evaluation, Department of National Defense.

EXTENDED ABSTRACT

Introduction

Program of international student assessment (PISA) is an international survey that has been conducted every 3 year by OECD aiming to investigate 15-year-old students' mathematics, science, and reading literacy since 2000. The results of PISA have important implications as it provides governments to compare themselves within an international platform and affect countries' education policies. Hence, concerns about the comparability of PISA results becoming an important and controversial issue.

Measurement instruments are expected to have same psychometric properties for all examinees (Woods, 2009). When an instrument has differentiating psychometric properties between some groups under controlling for differences in the abilities, it is defined as bias. Differential item functioning (DIF) is one of the statistical indicator of bias. Within item response theory perspective, DIF can be defined as differentiation of the probability of correctly answering a question between groups, which have same ability level (Camilli ve Shepard, 1994). When a test has items with DIF, it may reduce the validity of test in group comparisons. Moreover, as the bias causes systematic error, increasing systematic error not only cause decrease in detection of possible differences but also may cause misleading results regarding amount and direction of the difference. Hence, it is important to eliminate DIF items in high stake assessments like PISA.

The purpose of this study was to investigate the presence of DIF between gender groups in PISA 2015 science items in nine selected countries. Moreover, the effect of socioeconomic status and reading ability, on the presence of gender-related DIF were examined, respectively. One cluster from computer-based assessment (CBA) was taken into consideration. The countries were selected among the ones that implemented CBA, on the basis of their rank in science achievement.

Method

The current study was designed as descriptive study since it aims to test validity of PISA 2015 assessment's selected science items. S12 item cluster consisting new items was taken into consideration in the study. S12 cluster consists of 17 items within 5 content area were addressing three competencies: Evaluate and design scientific enquiry, explain phenomena scientifically, interpret data and evidence scientifically, defined in PISA framework.

Population of the study consists of 15-year-old students in Japan, Finland, Hong Kong, United States, France, Sweden, Turkey, Brazil, and Israel. The sample consists of the students who participate PISA 2015 examination and take S12 science cluster in selected countries.

In data analysis, multiple indicators multiple methods (MIMIC) was used. First, items were tested whether they have DIF between gender groups. Then, SES and reading ability included in to the model and their effects on gender DIF was tested. In parameter estimation WLS method was used.

Results and Discussion

The results of the study indicated that 13 items have exhibited gender DIF within different countries. When the interaction between direction of DIF and item competencies examined, in general items requiring explain phenomena scientifically has shown DIF favoring boys and items requiring interpretation of data and evidence scientifically has shown DIF favoring girls. These results are consistent with Le (2009). However, one item in Israel sample and two items in Sweden and France samples contradict the below classification related to item competencies and DIF direction.

The second main objective of the study was to determine whether SES and reading ability have any indirect effects on gender-based DIF analysis. The specified variables were added to the model to better determine the gender DIF on the items specified since they were highly correlated with science achievement. The addition of these variables into the model reduced the number of DIF items in four countries. These findings indicate that the variables of SES and reading ability reduce the number of gender-related DIF items. The SES has a significant impact on gender-based DIF in Turkey and Sweden. The reading ability has a significant impact on gender-based DIF in Turkey, Finland and Hong Kong. In the literature, it is reported that girls generally have higher scores on reading ability than boys (Hyde and Lin, 1988; Logan and Johnstone, 2009). After addition of reading ability into the model, the disappearance of DIF, which was in favor of girls before, can be considered as an important finding.

When the findings are examined in general, it can be seen that 13 of the items in the item cluster reveal gender-based DIF in different countries. When the direction of the DIF is examined, it can be seen that the items, which require the explanation competency of the phenomenon by scientific means, mostly reveal DIF in favor of boys while those require the interpretation competency of data and evidence, mostly reveal DIF in favour of girls. When compared with the countries, the countries having the highest number of items with gender-based DIF are Turkey, Hong Kong, Sweden, and Israel. France has the least number of DIF items. The inclusion of the SES variable into the model reduced the number of gender-based DIF items in Turkey and Sweden, but had no significant effect in other countries. The inclusion of the reading variable into the model, on the other hand, reduced the number of gender-based DIF items in Hong Kong, Finland, and Turkey. The variables added into the model had no effect on gender DIF in Japan, the United States, France, Brazil, and Israel.

It is highly important to examine the effect of demographics of individuals on test scores in large-scale assesments such as PISA, which are considered as base for international comparisons. Test scores should be independent of individual demographics, yet many studies report that gender is a DIF source for large-scale exams. The results of this study, in general terms, show that some of the differences in test performance may be explained by gender In addition socioeconomic level and reading ability may also contribute to this difference. Further studies can examine how other external variables, such as testing in a foreign language, subject area, item type, which are known to be associated with student achievement, influence intergroup comparisons. In addition, SES and reading ability are considered as continuous variables in the study. In the literature, on the other hand, generally discrete variables are used in DIF analyzes. In this context, the effect of the consideration of the specified variables as continuous or discrete can also be investigated.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(4);391-403

The Impact of Q-matrix Misspecification and Model Misuse on Classification Accuracy in the Generalized DINA Model*

Miao GAO** M. David MILLER*** Ren LIU****

Abstract
This simulation study explored the impact of Q-matrix misspecification and model misuse on examinees' classification accuracy within the generalized deterministic input, noisy "and" gate (G-DINA) model framework under the different conditions. The data was generated by saturated G-DINA model. Along with the generating model, two reduced models were used to fit the data: the additive CDM (A-CDM) and DINA model. The manipulated conditions included number of respondents, attribute correlations and test length. Two types of classification accuracy were examined: the overall classification accuracy and the class-specific classification accuracy. Results showed that the Q-matrix misspecification influenced classification accuracy more ominously than model misuse. The proportion of examinees classified correctly for each latent class was related to the types of Q-matrix misspecification. More test items had greater positive impact on classification accuracy than more respondents taking the test.

Key Words: Classification, cognitive diagnostic assessment, the generalized DINA model, Q-matrix misspecification

INTRODUCTION

Researchers and educational stakeholders have increasingly demanded more formative test information (Mislevy, 2006; Robets & Gierl, 2010; Rupp & Templin, 2008). They often wish to obtain the classification of respondents with respect to their skills. Teachers, students and parents often want to know the individual's level of skill mastery to facilitate an individual's development. Cognitive diagnosis models (CDMs) are used to measure the respondents' knowledge structures and the multiple attributes for the purpose of making classification-based decisions (Rupp, Templin, & Henson, 2010).

Despite the diversity of parametric models, general DCMs have gained increasing attention in recent years because they do not have idiosyncratic hypotheses about the impact of attribute relationship among items. They subsumed many popular models that were developed earlier such as the deterministic inputs, noisy, "and" gate (DINA; Junker & Sijtsma, 2001) models, the deterministic inputs, noisy, "or" gate (DINO; Templin & Henson, 2006) models, and the reduced reparameterized unified (R-RUM; Hartz, 2002) model. The three most common general DCMs are the general diagnostic model (GDM; von Davier, 2008, 2010), the log-linear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009), and the generalized DINA model (G-DINA; de la Torre, 2011). Among the three models, the G-DINA extends the logit link function of the other two models to multiple link functions including identity and log links.

One of the most important steps before specifying CDMs is to identify the attributes measured by each items. This item by attribute specification is usually constructed by the content experts and is called the Q-matrix. In practice, if the item-attribute alignment we specified a priori is not supported by the

* This study is derived from a part of Ph.D. dissertation by Miao GAO entitled Assessing the Model Fit and Classification Accuracy in Cognitive Diagnosis Models, submitted to University of Florida Graduate School under the supervision of Professor M. David MILLER.

** Assistant professor, Nanjing Normal University, College of Education, Nanjing, China, e-mail: miaogaonj@163.com, ORCID ID: 0000-0001-6842-5468
*** Professor, University of Florida, Research and Evaluation Methodology Program, Gainesville-Florida, US, e-mail: dmiller@coe.ufl.edu, ORCID ID: 0000-0002-3506-1167.
**** Ph.D candidate, University of Florida, Research and Evaluation Methodology Program, Gainesville-Florida, US, e-mail: liurenking@ufl.edu, ORCID ID: 0000-0002-6708-4996.

DOI: 10.21031/epod.332712
Received: 04.08.2017
Accepted: 06.10.2017

data, the Q-matrix may be misspecified. Previous research has shown that parameter estimates and classification accuracy were affected by the misspecification of Q-matrix (e.g., Rupp & Templin, 2008; Kunina-Habenicht, Rupp, & Wilhelm, 2012). Specifically, Rupp and Templin (2008) used the different types of Q-matrix misspecification under the DINA model. They found the Q-matrix misspecification had caused biased parameter estimates and lower classification accuracy corresponding to the examinees' latent class. However, questions such as whether the results may be generalizable to more general contexts. The purpose of this study is to estimate the effects of specific types of Q-misspecification on examinee classification accuracy under the generalized G-DINA model.

The rest of the manuscript is structured as follows: In the theoretical framework, we first provide an overview of the Q-matrix, types of Q-matrix misspecifications and the generalized DINA model. In the method, the simulation design, the model estimation and the outcome assessment are described. Next the findings of this study are described. Lastly this manuscript is closed with a discussion of the findings.

Background

Q-matrix

A critical step in cognitive diagnostic model is to develop the Q-matrix because CDM and the Q-matrix are essential modeling process. Developing the Q-matrix defines the attribute structure measured by an assessment. An example of a JxK Q-matrix can be demonstrated as follows:

$$Q = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ & & \\ 1 & 1 & 1 \end{bmatrix} \quad (1)$$

Where j indicates "item" and k indicates "attributes." The element, qjk, is specified as "1" if the jth item requires the kth attribute to answer this item correctly; otherwise, qjk is specified as "0".

In a Q-matrix, each element qjk indicates whether the item j measures the attribute k, where qjk =1 means item j measures the attribute k and qjk =0 means item j does not measure the attribute k. The Q-matrix reflects the loading structure of the multiple attributes on the items. The Q-matrix is specified by content experts and this specification process is a subjective activity (Rupp, Templin, & Henson, 2010). Hence, the quality of the Q-matrix determines the diagnostic information obtained from the CDM analysis.

The Generalized DINA Model Framework

The generalized DINA model, like all other CDMs, requires a *J x K* Q-matrix. The G-DINA discriminates latent classes into $2^{K_j^*}$ latent groups, where $K_j^* = \sum_{k=1}^{K} q_{jk}$ represents the required attributes for item j. Each latent group is reduced to an attribute vector represented by α_{lj}^*. In this study, it would suffice to use the reduced vector $\alpha_{lj}^* = (\alpha_{lj1}^*, \dots, \alpha_{ljK_j^*}^*)$ instead of the full vector $\alpha_{lj} = (\alpha_{lj1}, \dots, \alpha_{ljk})$. Each latent group has the probability of answering correctly the item represented by $P(\alpha_{ij}^*)$. The item response function (IRF) for G-DINA could be written as:

$$P(\alpha_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} \dots + \delta_{j12\dots K_k^*} \prod_{k=1}^{K_j^*} \alpha_{lk} \quad (2)$$

where δ_{j0} is the intercept, δ_{jk} is the main effect by α_k , $\delta_{jkk'}$ is the interaction effect by α_k and $\alpha_{k'}$, and $\delta_{j12\dots K_j^*}$ is the interaction effect by α_1 ,…, α_{k*}.

The DINA model, that is the most commonly used reduced model, is a special case of the G-DINA model. By setting all the parameters, except δ_{j0} and $\delta_{j12\ldots K_j^*}$, to zero, the IRF for DINA model is as follows:

$$P(\alpha_{lj}^*) = \delta_{j0} + \delta_{j12\ldots K_k^*} \prod_{k=1}^{K_j^*} \alpha_{lk} \tag{3}$$

Another special case of the G-DINA model is the A-CDM, which contains only the intercept and the main effects. The IRF for A-CDM is defined as follows:

$$P(\alpha_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk} \alpha_{lk} \tag{4}$$

This model contains only the intercept and the main effect of each attribute.

METHOD

Simulation Study

The simulation study was aimed to examine the effects of Q-matrix misspecification and CDM misuse on classification accuracy. All data generation and estimations were conducted using the software R (R Core team, 2016). The data was generated using the saturated model G-DINA. The number of respondents, the correlation between attributes, and the number of items measured in a test were manipulated and resulted in 12 data-generating conditions with 1000 replications for each condition. For each of the generated datasets, three CDMs within the G-DINA framework were applied for the data analysis: the G-DINA, A-CDM and DINA models. Six Q-matrices, including 1 correctly specified Q-matrix and 5 misspecified Q-matrices were examined. In total, there were 216 different settings for data analysis, which included 18 diverse estimations and 12 different data-generating conditions.

Number of respondents

Three levels of number of respondents reflecting small, moderate and large samples were investigated in this study: N = 500, 1000 and 5000. Previous research has shown this is a relevant factor that influences model fit, parameter estimates, and classification (Chen, de la Torre, & Zhang, 2013; Cui, Gierl, & Chang, 2012; de la Torre, 2009; de la Torre & Douglas, 2004; Shu, Henson, & Willse, 2013). Several studies have shown that number of respondents should be at least 500 in order to have an acceptable model fit and relatively accurate parameter estimates even when using the reduced model as the generated model (Chen et al., 2013; Cui et al., 2012; Shu et al., 2013). The pilot study indicated that when the sample size increased to 500, the model fit achieved an acceptable level.

Number of attributes

This study focused on one level of the number of attributes K =4. A review of the CDM simulation studies indicates that there are usually three to eight attributes being designed in an assessment, which also reflects the number of attributes in application examples (Cheng, 2009; Chen et al., 2013; DeCarlo, 2012; de la Torre, 2009; de la Torre & Douglas, 2004; Huebner & Wang, 2011; Kunina-Habenicht et al., 2012). Considering all the other factors being manipulated in the simulation and a fairly large estimation process, the attributes' number was fixed at four in this study.

Marginal attribute difficulty

A multivariate normal distribution for latent attributes with the mean vector and correlation matrix were used to generate respondents' true attribute patterns. In this study, the mean vector of(0, 0, 0, 0) was used for the four attributes test; this led to the same marginal mastery proportions for all attributes of .50. This mean vector is also called marginal attribute difficulty.

Correlation between attributes

Two levels of attribute correlation were set to values of .4 and .8 to represent moderate and high correlation(Henson, Roussos, Douglas & He, 2008), respectively. A range of .3 to .9 of the tetrachoric correlation is typical in educational assessment and CDM research (Cui et al., 2012; Henson, Templin & Douglas, 2007; Kunina-Habenicht et al., 2012). A weakly correlated attributes level could be included as a contrast, but I chose not to do this to keep the overall simulation and estimation manageable. The correlations were set to be equal across all attribute pairs in the correlation matrix.

Q-matrix specification

The number of items in a test was set to two levels in this study: J =14 and 28. The number of items and the number of attributes measured in a test are associated. For K =4, the number of all possible attribute patterns was 2^4=16, and there are 15 attribute patterns. Considering the computational time, we set the maximum number of attributes being assessed by an item to three. The item 1-14 in Table 1 showed the Q-matrix specification for generation when J=14. This simulation design also investigated the conditions where the test length is equal to and greater than the number of possible attribute patterns. Two levels of the item number were examined in this study: J = 14 and 28 for the number of attributes K = 4. The Q-matrix for J=28 was a duplicate of Q-matrix for J-14. The correctly identified Q-matrix for J = 28 is also shown in Table 1. The Q-matrix for J = 14 was embedded as a subset of this Q-matrix.

Table 1. Correct Q-Matrix of J = 14 and 28

	Attribute					Attribute			
Item	#1	#2	#3	#4	Item	#1	#2	#3	#4
1	1	0	0	0	15	1	0	0	0
2	0	1	0	0	16	0	1	0	0
3	0	0	1	0	17	0	0	1	0
4	0	0	0	1	18	0	0	0	1
5	1	1	0	0	19	1	1	0	0
6	1	0	1	0	20	1	0	1	0
7	1	0	0	1	21	1	0	0	1
8	0	1	1	0	22	0	1	1	0
9	0	1	0	1	23	0	1	0	1
10	0	0	1	1	24	0	0	1	1
11	1	1	1	0	25	1	1	1	0
12	1	1	0	1	26	1	1	0	1
13	1	0	1	1	27	1	0	1	1
14	0	1	1	1	28	0	1	1	1

Note. Items 1-14 are used when J = 14.

Different types of the Q-matrix misspecification were investigated: under-fitting the Q-matrix (defining 1 as 0), over-fitting the Q-matrix (defining 0 as 1), and a balanced misfit (exchanging 0 and 1). As shown in Table 2, taking the test with J=14 items as an example, *qt-14* was the true Q-matrix for data generation. Two under-specified Q-matrices *qu3-14* and *qu2-14* meant that qu3-14 Q-matrix

changed all 3-attribute items into selected 2-attribute items, and this selection of the attribute deletion was random for each item; *qu2-14* Q-matrix changed all 2-attribute items into selected 1-attribute items, and this selection of the attribute deletion was random for each item. Similarly, two over-specifications *qo1-14* and *qo2-14* Q-matrices were created by randomly selecting the attribute being added. For creating the balanced misfit for the Q-matrix (*qm-14*), the items that needed to be altered were first randomly selected; then, the attributes that needed to be altered were selected randomly for each item.

Table 2. The Q-Matrix Misspecification and True Q-Matrix

K	J	Q-matrix	Alternations	Item Altered	Total # of changes (1 to 0)	Total # of changes (0 to 1)	Ave. # of attributes per item	Ave. # of items per attribute
4	14	*qt-14*	Data generating Q-matrix	0	0	0	2	7
		qu3-14	All 3-attribute items are changed into selected 2-attribute items.	I11 - I14	4	0	1.71	6
		qu2-14	All 2-attribute items are changed into selected 1-attribute items.	I5 - I10	6	0	1.57	5.5
		qo1-14	All 1-attribute items are changed into selected 2-attribute items.	I1-4	0	4	2.29	8
		qo2-14	All 2-attribute items are changed into selected 3-attribute items.	I5 - I10	0	6	2.43	8.5
		qm-14	Attributes are deleted and added to balance out the overall number of changes.	2 items randomly selected from I1-I4; 3 items randomly selected from I5-I10; 2 items randomly selected from I11-I14	7	7	2	7
4	28	*qt-28*	Data generating Q-matrix	0	0	0	2	14
		qu3-28	Half of the 3-attribute items are changed into selected 2-attribute items.	I11 - I14	4	0	1.86	13
		qu2-28	Half of the 2-attribute items are changed into selected 1-attribute items.	I5 - I10	6	0	1.79	12.5
		qo1-28	Half of the 1-attribute items are changed into selected 2-attribute items.	I1-4	0	4	2.14	15
		qo2-28	Half of the 2-attribute items are changed into selected 3-attribute items.	I5 - I10	0	6	2.21	15.5
		qm-28	Attributes are deleted and added to balance out the overall number of changes.	2 items randomly selected from I1-I4; 3 items randomly selected from I5-I10; 2 items randomly selected from I11-I14	7	7	2	14

The assessment with number of items J = 28 had doubled the items as in the assessment J =14. The misspecification of Q-matrix in J =28 only occurred in items 1 to 14, and items 15 to 28 always

remained the same as in true Q-matrix (qt-28). In this way, the number of misspecified items in J = 28 was the same as in J =14 when controlling the type of misspecification, which made the results comparable for different test length.

Item parameter specification for data generation

The parameter setting was referenced from an empirical study (Basokcu, Ogretmen, &Kelecioglus, 2013). The true item parameters (δ_{jk}) used in this simulation study were ranged from 0.12 to 0.68, and the detailed values were presented in Table 3. For simplicity, all the one-attribute items used the same parameter setting, and the same idea was followed for the two- and three-attribute items.

Table 3. Item Parameters for Data Generation (α_{jk})

	Attribute Pattern and Parameters							
1-attribute item	0	1						
	.21	.68						
2-attribute item	00	10	01	11				
	.18	.25	.15	.59				
3-attribute item	000	100	010	001	110	101	011	111
	.26	.12	.17	.18	.13	.27	.26	.51

Model selection

Each of the generated datasets was analyzed by three CDMs within the G-DINA framework. The true generating model was the G-DINA model. In addition to the true model, two misused CDMs were used to analyze the data. misusage of CDM refers to incorrect parameterization of the modeling process. As two comparison models, A-CDM contained only intercept and main effects for each item; and the DINA model contained only intercept and the highest order of interaction effect for each item.

Outcome Measures

Classification accuracy (CA) is defined as the degree to which the classification of examinees' latent classes analyzed by observed data agrees with examinees' true latent classes (Cui et al., 2012). The simulated examinee attribute patterns were used as the true examinees' latent classes; the attribute patterns estimated from the response data using MLE method were used as the estimated latent classes. The simulated and estimated latent class were then compared for each examinee. If they were consistent, a value of "1" was assigned to the examinee to represent being classified accurately; otherwise, a value of "0" was assigned for being classified inaccurately. By taking the average of 0/1 over all examinees and all replications, the overall correct classification rates were calculated for each condition, which refers to overall classification accuracy (OCA). By taking the average of 0/1 for the examinees by each latent class, the class-specific correct classification rates were calculated, which refers to class-specific classification accuracy (CCA). In order to simplify the interpretation of the findings, the CCA was calculated based on one generating condition (n = 5000, ρ = .4 and J = 14) and being fitted with the various CDMs and Q-matrices. The OCA and CCA were then compared for all the estimation settings.

RESULTS

In CDM estimations, the classification is usually of primary interest because the decisions about the examinees are made based on the classification (Rupp, Templin, & Henson, 2010). Two types of the classification accuracy were illustrated in this part: the OCA and CCA.

Overall Classification Accuracy (OCA)

Table 4 demonstrated the overall correct classification rates by the different levels of all factors. For the purpose of explaining the results more explicitly, the effects of N, J and on the OCA were the focus in Table 4. The impact of CDM misuse and Q-matrix misspecification on OCA is examined in Figure 1.

Table 4. Overall Classification Accuracy (OCA) in All Conditions

				Q-matrix Specification					
Model	ρ	*J*	*N*	*qt*	*qu3*	*qu2*	*qo1*	*qo2*	*qm*
G-DINA	0.4	14	500	0.711	0.689	0.525	0.706	0.695	0.467
			1000	0.719	0.698	0.533	0.717	0.713	0.477
			5000	0.727	0.705	0.542	0.726	0.725	0.481
		28	500	0.886	0.879	0.837	0.885	0.883	0.815
			1000	0.889	0.884	0.843	0.889	0.888	0.826
			5000	0.893	0.888	0.850	0.893	0.893	0.834
	0.8	14	500	0.720	0.731	0.647	0.714	0.709	0.629
			1000	0.723	0.734	0.651	0.720	0.720	0.639
			5000	0.726	0.733	0.650	0.725	0.724	0.644
		28	500	0.873	0.876	0.846	0.871	0.870	0.824
			1000	0.875	0.879	0.852	0.874	0.873	0.831
			5000	0.877	0.881	0.858	0.877	0.876	0.837
A-CDM	0.4	14	500	0.669	0.645	0.536	0.657	0.641	0.460
			1000	0.675	0.652	0.540	0.666	0.647	0.462
			5000	0.689	0.653	0.543	0.678	0.654	0.453
		28	500	0.838	0.833	0.810	0.833	0.828	0.776
			1000	0.850	0.846	0.816	0.847	0.845	0.792
			5000	0.859	0.851	0.820	0.860	0.859	0.800
	0.8	14	500	0.738	0.726	0.641	0.730	0.717	0.617
			1000	0.750	0.736	0.642	0.751	0.743	0.619
			5000	0.755	0.742	0.643	0.760	0.758	0.609
		28	500	0.887	0.868	0.845	0.887	0.886	0.827
			1000	0.888	0.868	0.846	0.889	0.888	0.831
			5000	0.889	0.870	0.848	0.890	0.889	0.835
DINA	0.4	14	500	0.643	0.648	0.512	0.448	0.526	0.374
			1000	0.647	0.652	0.513	0.454	0.530	0.375
			5000	0.648	0.652	0.515	0.458	0.532	0.379
		28	500	0.851	0.855	0.767	0.736	0.773	0.737
			1000	0.858	0.861	0.771	0.740	0.785	0.744
			5000	0.865	0.867	0.781	0.746	0.794	0.753
	0.8	14	500	0.673	0.664	0.652	0.505	0.613	0.485
			1000	0.678	0.666	0.653	0.509	0.613	0.488
			5000	0.682	0.668	0.653	0.511	0.613	0.490
		28	500	0.870	0.872	0.828	0.712	0.811	0.750
			1000	0.878	0.879	0.832	0.715	0.821	0.753
			5000	0.882	0.883	0.836	0.717	0.825	0.759

As shown in Table 4, when test length increased, the correct overall classification rates were much higher. For example, in G-DINA model with qt matrix, the correct overall classification rates went up from .711 to .886 as test length increased from J=14 to 28,controlling ρ= .4 and N = 500. This is expected because more pieces of information provided by the items for each dimension can be used to detect the classification. Second, as the sample size increased, the overall classification rates slightly increased for all conditions. For example, again in G-DINA model with qt matrix, the overall classification accuracy increased from .886 to .893 as sample size increased from 500 to

5000,controlling ρ= .4 and J=28. Comparing the effects of J and N on classification accuracy, we can see that more items in a test are more critical than more examinees to get a better classification accuracy. Third, the increase in attribute correlation slightly increased the overall classification accuracy with few exceptional conditions.

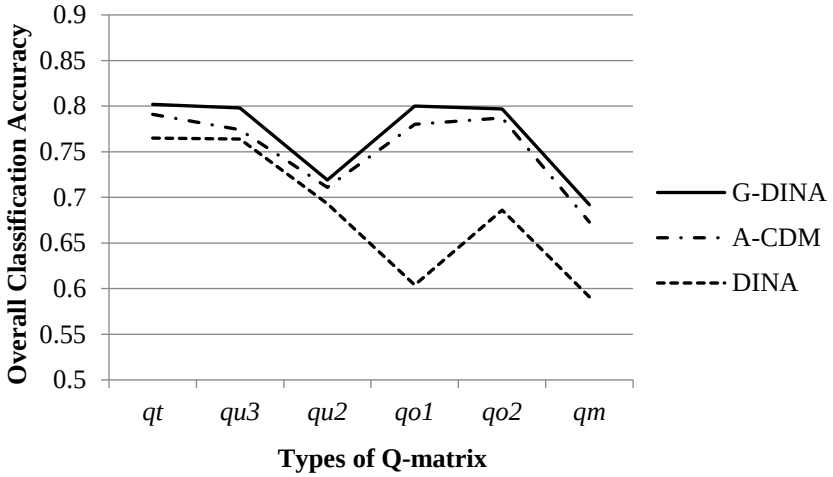


Figure 1. Overall Classification Accuracy (OCA) by CDM and Q-matrices

To investigate the effects of the misspecification of CDM and Q-matrix, the correct overall classification rates were shown in Figure 1. The classification rates used in this figure were collapsed over the other factors N, J and for the simpler illustration.

For CDM misuse, Figure 1 showed that the overall classification accuracy was highest in G-DINA no matter which specified Q-matrix was used. This makes sense because G-DINA was the generating model. Comparing the other two CDMs, A-CDM has higher classification rates than DINA. The A-CDM yielded very similar overall classification rates with the true model G-DINA where A-CDM contained only main effects of the attributes and omitted all the interactions. The DINA model showed the lowest classification rates among three CDMs where DINA contained only the highest order of interactions among attributes.

For investigating Q-matrix misspecification, the condition qt was the correct Q-matrix and could be used as baseline rates in each CDM. Figure 1 showed that the OCA in qt was higher than the other misspecified Q-matrices in three CDMs. The effects of the misspecified Q-matrices on classification accuracy were then compared with the true Q-matrix in different CDMs. The classification rates in G-DINA and A-CDM showed similar patterns for the Q-matrix misspecification. Within these two models, the OCA for the condition qu3, qo1 and qo2 was close to the rates in qt. The misspecified qu2 had lower overall classification rates, and the misspecified qm showed the lowest overall classification rates. This is not surprising because the qm included all types of misspecification. To compare the effects of different Q-matrices in the DINA model, the OCA was highest in qt; the condition qu3 yielded almost the same results with qt; while the lowest classification still occurred in qm among all the conditions. The Q-matrices qo2 and qu2 in DINA model yielded the moderate classification rates.

Table 5. Correct Overall and Class-specific Classification Rates for Misspecifications of CDM and Q-matrix

Model	Q-matrix	Overall	Attribute Classes														
			0000	1000	0100	0010	0001	1100	1010	1001	0110	0101	0011	1110	1101	1011	0111	1111
G-DINA	qt	.802	.587	.653	.666	.666	.673	.810	.816	.830	.822	.836	.847	.922	.934	.934	.940	1
	qu3	.798	.688	.637	.609	.609	.578	.756	.728	.726	.758	.756	.815	.840	.890	.890	.912	1
	qu2	.719	.792	.532	.568	.568	.305	.502	.745	.500	.532	.406	.492	.849	.807	.807	.672	.944
	qo1	.800	.586	.648	.661	.661	.668	.806	.812	.827	.818	.832	.843	.923	.933	.933	.939	1
	qo2	.797	.592	.631	.648	.648	.654	.792	.801	.816	.804	.817	.829	.921	.932	.932	.936	1
	qm	.692	.741	.290	.353	.353	.526	.438	.414	.556	.518	.471	.624	.673	.557	.557	.574	.994
A-CDM	qt	.791	.647	.628	.623	.623	.599	.731	.753	.759	.767	.771	.776	.864	.872	.872	.877	.995
	qu3	.774	.740	.612	.592	.592	.541	.636	.652	.637	.688	.680	.782	.712	.809	.809	.830	.991
	qu2	.711	.781	.545	.593	.593	.307	.488	.728	.485	.524	.375	.481	.826	.800	.800	.664	.936
	qo1	.780	.612	.608	.603	.603	.583	.723	.733	.746	.759	.764	.764	.869	.879	.879	.882	.994
	qo2	.787	.638	.608	.611	.611	.589	.726	.747	.754	.753	.762	.768	.867	.877	.877	.880	.997
	qm	.673	.735	.334	.362	.362	.481	.390	.355	.549	.434	.429	.516	.654	.553	.553	.531	.994
DINA	qt	.765	.597	.551	.593	.593	.608	.684	.690	.698	.710	.715	.733	.922	.908	.908	.925	1
	qu3	.764	.601	.505	.547	.547	.571	.715	.705	.727	.765	.749	.766	.887	.906	.906	.925	1
	qu2	.693	.735	.449	.482	.482	.253	.506	.705	.462	.463	.330	.384	.843	.726	.726	.675	1
	qo1	.604	.246	.242	.240	.240	.286	.664	.629	.668	.548	.724	.627	.892	.870	.870	.848	1
	qo2	.686	.555	.439	.455	.455	.467	.568	.557	.570	.579	.579	.573	.864	.820	.820	.610	1
	qm	.591	.396	.221	.376	.376	.419	.384	.433	.492	.462	.445	.626	.725	.517	.517	.635	1

Class-specific Classification Accuracy (CCA)

When we examined the respondents' classification at each latent class level, it was worthwhile to note that classes with more attributes tended to have generally higher classification accuracy in various CDMs and Q-matrices (Table 5). For example, in the G-DINA and qt condition, the CCA ranged from 0.587 to 1 for the class with no attribute to the class with all attributes. The attribute class in which all attributes were mastered (attribute pattern 1111) maintained very high correct classification rates no matter which CDM and Q-matrix were used. Especially in the DINA model, the misclassification of examinees in this attribute class never occurred.

Comparing the different CDMs, the G-DINA model yielded the highest CCA in almost all the latent classes with few exceptions. When using qt in G-DINA model, the correct classification rates for one-attribute mastery classes were at least 65%; and these rates reached at least 80% and 90% for two- and three-attribute mastery classes, respectively. A-CDM performs better than DINA in the classes with zero-, one- and two-attributes. The CCA by using qt and A-CDM were approximately .6 for one-attribute mastery classes, .75 for two-attribute mastery classes, .87 for three-attribute mastery classes.

However, the DINA model had higher than expected classification accuracy in three- and four-attribute mastery classes, even with misspecified Q-matrices. More specifically, focusing on the three-attribute mastery classes, the CCA of the DINA model using qt were .922, .908, .908 and .925, while the G-DINA model using qt has almost the same classification accuracy. In qu2, qo1 and qo2, the CCA of the DINA model was slightly lower than G-DINA and higher than A-CDM in three-attribute classes' estimations. In qu3, DINA even performed best among three CDMs in the classification accuracy of three-attribute latent classes (.887, .906, .906 and .925).

Considering the Q-matrix misspecification, the class-specific classification rates are related to the different types of misspecified Q-matrices (under-, over- or mixed misspecification). G-DINA and A-CDM showed a similar pattern: The over-specified Q-matrices (qo1 and qo2) did not have much impact on the class-specific classification accuracy. The under-specified Q-matrices, especially qu2, had much lower CCA in these two models. While in the DINA model, the misspecified qu2 and qo2 seemed to have a more severe impact on CCA; the qo1 mainly affected the correct classification rates on the classes with fewer attributes. The misspecified qm, for all three fitting models, showed the lowest classification rates, and the low class-specific classification rates occurred in almost all attribute classes.

Furthermore, we noticed that the low class-specific classification rates corresponded to the attribute patterns that matched the manipulated attribute classes. For example, in the misspecified qu2 where two-attribute items were changed into one-attribute items, the correct classification rates of two-attribute mastery classes (e.g. attribute class [1100]) dropped a great deal when compared with qt condition. The correct classification rates of one-attribute mastery classes (e.g. attribute class [0001]) decreased as well in all three CDMs. Unlike G-DINA and A-CDM, in the condition qo1 where the one-attribute items were changed to two-attribute items, the classification rates for having one attribute in DINA were very low which matched the manipulated items. In the condition qo2, the CCA of two-attribute mastery classes were low as well in the DINA model.

DISCUSSION and CONCLUSION

The G-DINA model offers a flexible framework to investigate the issues in examinees' diagnostic classification. The specification of Q-matrix and the choice of CDM play a critical role for achieving better classification accuracy. This study helps to understand better of the effects of CDM misuse and Q-matrix misspecification on classification accuracy under various conditions. The different factors, such as number of test items, number of examinees and attribute correlation, all have certain impacts on examinees' classification. The outcome of CDMs provides meaningful formative test information about the multiple proficiencies of the attributes measured in each examinee. Although this study is sufficiently complex, it clearly can be extended by using a broader range of design.

This simulation study contributed in the following four aspects. First, the G-DINA model was used as a framework that aligned with the trend in CDM development. The simulation was conducted in the saturated model and fit the data with two reduced models as well as the saturated model, which better aligns to the practice of real data analysis. Second, both the Q-matrix misspecification and CDM misuse were investigated separately and conjunctively. Third, the under-, over- and mixed misspecified Q-matrices allow us to detect the more specific effects of Q-matrix misspecification under various conditions in a generalized CDM framework. Fourth, the overall classification accuracy and the class-specified classification rates (often the primary interest in CDM analysis) were investigated under different conditions in this study.

Both the number of respondents and test length illustrated clear positive effects on classification accuracy. Despite the model selection and Q-matrix specification, the increase of the number of respondents and/or the test items always demonstrated the growth in the correct classification rates. One noticeable finding is that the increase in test length improved the classification accuracy more dramatically than the increase in sample size. It provides an insightful direction to the practitioners, to assist in making the decision of which factors will be manipulated, in order to effectively improve the examinees' classification accuracy.

Our results also demonstrated that model misuse does not noticeably affect the overall classification accuracy, even though the G-DINA model still maintained the highest level of classification accuracy. We simulated data in the saturated G-DINA model by mimicking the complex empirical situation. When estimating the data with various CDMs, we found the models performed differently under the consideration of examinees' latent classes. For the examinees who have fewer attributes (e.g. one- or two-attribute), G-DINA and A-CDM models yield more accurate classification rates than the DINA model. A-CDM showed a better classification accuracy in the non-attribute mastery class. This may due to the structure of G-DINA and A-CDM models that they contains the main effects. For the examinees who have more attributes (e.g. three- or four-attribute), the DINA model that contained only the highest order of interaction had higher than expected classification accuracy even with the Q-matrix misspecification. Given these, although A-CDM is easier to interpret in practice, if we have large number of attributes, it may be worth considering having higher order interaction effects.

One important finding in this study is that the misspecification of Q-matrix affected the overall classification accuracy in a more obvious way than model misuse. In practical application, the true Q-matrix is unknown and there is a possibility that Q-matrix could be misspecified in the designing process. As expected, the true Q-matrix yielded the most accurate classification. In general, the under-misspecified Q-matrices had more severe impact on CA than over-misspecified Q-matrices especially in the models with main effects. The misspecified Q-matrix qm was most problematic because the correct classification rates were low in almost all the conditions. Although the number of attributes held constant in qm, a large number of misspecification occurred. The qm contained all types of the misspecification and represented the most severe misspecification. Thus, it is not only the number of misspecified items that matters but also the types of misspecification. The attribute structure, rather than the number of attribute by item, is a much more important component in the diagnosis process. In practice, we may face a situation where there is an uncertainty in determining whether one item measures the attribute. We suggest that over-specification may be better than under-specification.

Besides the effect on overall classification rates, the different types of misspecified Q-matrices also showed the effects on the corresponding latent class. When a certain attribute combination is not represented in the Q-matrix, the respondents mastering the same attribute combination are more likely to be misspecified. A typical example in all three CDMs is the misspecified qu2, where two-attribute items were changed into one-attribute items. The classification rates decreased noticeably in the corresponding two-attribute mastery classes in all three CDMs. Thus inferences for the examinees in the associated classes should be more cautious.

Moreover, the effects of differently specified Q-matrices on classification accuracy varied in three CDMs. For example, the over-specified Q-matrix (qo1and qo2) influenced the DINA model more severely, but not in the G-DINA and A-CDM. The balanced misfit Q-matrix qm had shown more

dramatic negative effect on the classification rates in DINA than the other two models. This may be due to the different features of three CDMs. The saturated G-DINA model contains the main effects and all the ways of interactions, the A-CDM contains the main effects only, and the DINA includes only the highest order of interaction. In sum, the G-DINA model had a more stable performance in all latent classes when considering Q-matrix misspecification, although A-CDM performed well in zero- and one-attribute mastery classes and DINA showed high classification accuracy in three- and four-attribute mastery classes.

Regardless of the different types of CDMs and Q-matrices, it was noteworthy that the examinees in the latent classes with more attributes had higher classification accuracy, and the examinees in the latent classes with fewer attributes could not be classified accurately. This becomes considerable in practice when applying these CDMs to identify the mastery and non-mastery of multiple attributes, especially for the examinees at the lower end. The attribute class mastering all attributes almost never showed any misspecification rates; while the attribute class with no attributes had low correct classification rates. For addressing the possible reasons of this phenomenon, future research may examine the impact of item difficulty and the distribution of attribute patterns.

In practice, the importance of diagnostic test development framework and Q-matrix validation methods should be emphasized. After the Q-matrix is designed, we recommend validating the Q-matrix using the method proposed in de la Torre (2008) and de la Torre and Chiu (2016) to check the possibility of misspecification. Yet it is not easy to evaluate the correctness of the Q-matrix due to its subjective nature and the complexity when applied to the model. When there is an uncertainly in determining if one item measures the attribute, over-specification may be better than under-specification. In order to classify the examinees into latent groups, the selection of the CDMs may relate to which group of examinees are more concerned with. The saturated model usually yields more stable classification accuracy across all the latent classes. The model with higher order interactions should be considered when there are a number of attributes, although the model with only main effects is easier to interpret. Hopefully the findings of this study will provide some insights for practitioners and researchers in determining the Q-matrix and cognitive diagnostic models when facing various situations.

REFERENCES

Basokcu, T. O., Ogretmen, T., & Kelecioglu, H. (2013). Model data fit comparison between DINA and G-DINA in cognitive diagnostic models. *Education Journal, 2*(6), 256-262.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*(2), 123-14.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*(4), 619-632.

Cui, Y., Gierl, M. J., & Chang, H. H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement, 49*(1), 19-38.

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement, 36*(6), 447-468.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*(4), 343-362.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34,* 115-13.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179-199.

de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*(2), 253-273.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333-353.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.

Henson, R., Roussos, L., Douglas, J., & He, X. (2008). Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement, 32*(4), 275-288.

Henson, R., Templin, J., & Douglas, J. (2007). Using efficient model based sum-scores for conducting skills diagnoses. *Journal of Educational Measurement, 44*(4), 361-376.

Henson, R. A., Templin, J. L., &Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191-210. Doi: 10.1007/S11336-008-9089-5

Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement, 71*(2), 407-419.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258-272.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement 49*, 59-81.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257-305). Washington, DC: American Council on Education.

R Core Team (2016). R (Version 3.3) [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing.

Robets, M. R., & Gierl, M. J. (2010). Developing score reports for cognitive diagnostic assessments. *Educational Measurement: Issues and Practice, 29*(3), 25-38.

Rupp, A. A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*(1), 78-96.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications.* New York: Guilford Press.

Shu, Z., Henson, R., & 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.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological methods*, *11*(3), 287-305. Doi: 10.1037/1082-989X.11.3.287

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 287-307.

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychology Science Quarterly*, *52*(1), 8-28.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(4);404-420

Öğretmen Adaylarının Eğitimde Ölçme ve Değerlendirme Dersindeki Kavram Yanılgılarının İncelenmesi

Examination of Pre-Service Teachers Misconceptions in Measurement and Evaluation Concepts

Sibel AYDOĞAN ** Selahattin GELBAL ***

Öz

Bu çalışmanın amacı öğretmen adaylarının eğitimde ölçme ve değerlendirme dersindeki ortak kavram yanılgılarını belirlemektir. Bu amaç doğrultusunda 17 sorudan oluşan iki aşamalı, çoktan seçmeli bir tanı testi geliştirilmiştir. Test puanlarının geçerliği uzman görüşü alınarak, güvenirliği ise test-tekrar test yöntemi kullanılarak incelenmiştir. Araştırmanın çalışma grubunu 2014-2015 öğretim yılında Hacettepe, Kırıkkale ve Akdeniz Üniversiteleri Eğitim Fakültelerinde öğrenim gören 328 öğretmen adayı oluşturmaktadır. Verilerin analizinde frekans ve yüzde kullanılmıştır. Bulguları yorumlama aşamasında her bir soru için seçilebilecek (4*4) 16 farklı seçenek ikilisi arasından grubun %10 ve daha fazlası tarafından işaretlenen yanlış kombinasyonlar yaygın yanılgı için bir ölçüt olarak değerlendirmeye alınmıştır. Testteki 17 sorunun analizinden toplam 30 yaygın yanılgı tespit edilmiştir. Çalışmada katılımcıların en çok "güçlük indeksi" ve "normal dağılım" kavramlarında ortak kavram yanılgısına sahip oldukları görülmüştür.

Anahtar Kelimeler: eğitimde ölçme ve değerlendirme dersi, iki aşamalı test, kavram yanılgıları, yaygın kavram yanılgısı.

Abstract

The purpose of this study was to determine pre-service teachers' common misconceptions in measurement and evaluation concepts. In accordance with this objective a two-tier multiple choice test which consists of 17 questions has been developed. The validity of the test was determined by judgements of assessment experts and reliability was examined by test-retest method. The present study was conducted in 2014-2015 academic year with total number 328 pre-service teachers of Hacettepe, Kırıkkale and Akdeniz Universities, Faculty of Education. The obtained data have been analyzed with frequency and percentage. To interpret the findings, the criteria for common misconceptions for each questions was accepted as the 10 percent or more of the options which were chosen among 16 binary alternatives by the group. Thirty common misconceptions were determined among the 17 questions. The most common misconceptions were about the "difficulty index" and "normal distribution" conceptions.

Keywords: measurement and evaluation concepts, two-tier test, common misconception.

GİRİŞ

Başarılı öğrenciler yetiştirmek eğitim sisteminin birincil amaçlarındandır. Bu amaçta büyük sorumluğa sahip öğretmenlerin, ölçme ve değerlendirme etkinliklerini sürekli olarak kullanmaları gerekmektedir (Şahin ve Ersoy, 2009). Eğitimin bütün alanlarını ilgilendiren ölçme ve değerlendirme işlemlerinde önemli olan etkinlikleri doğru olarak uygulayabilmektir. Bu sürecin doğru işlemesi için ise ilk şart alanın kapsadığı kavramların doğru yapılandırılmasıdır. Kavramlar, insan zihninde anlamlanan farklı obje ve olguların değişebilen ortak özelliklerini temsil eden bilgi

* Bu araştırma "Öğretmen Adaylarının Eğitimde Ölçme ve Değerlendirme Dersindeki Kavram Yanılgılarının İncelenmesi" adlı yüksek lisans tezinden üretilmiştir.

** Arş. Gör., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, e-posta: sibeldemirbilek07@gmail.com ORCID ID: orcid.org/0000-0002-0699-6510

*** Prof. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, e-posta: sgelbal@gmail.com ORCID ID: orcid.org/0000-0001-5181-7262

DOI: 10.21031/epod.355275
Geliş Tarihi: 17.11.2017
Kabul Tarihi: 06.12.2017

formları olarak adlandırılabilir (Ülgen 2004). İnsanlar kendi düşünce sistemlerince ortaya koydukları kavramlar aracılığıyla düşünebilirler. Ayrıca kavramlar kullanılarak geniş kapsamlı bilgiler daha küçük ve kullanıma elverişli birimler haline getirilebilirler (Senemoğlu, 2003). Kavramların özellikleri sürekli olarak incelenmekte ve yeniden tanımlanmaktadır. Bu sürekli inceleme kavramların tecrübeye dayalı olarak algılanan özellikleri kadar tanımlanabilmesinden kaynaklanmaktadır (Ülgen, 2001). İnsanlar doğumundan ölümüne kadar yeni kavramlar öğrenir ya da var olan kavramlarını yenileri ile değiştirir veya geliştirirler.

Öğrenmelerin başarıya ulaşması için içinde kavramları barındıran yeni bilgiler edinilmesi gerekmektedir. Bu yeni bilgilerin ise var olan bilgiler ile ilişkilendirilmesi ve uygun düzenlemelerin yapılması gerekmektedir. Yeni öğrenilen bir kavram, öğrencilerin zihninde var olan bilgiler ile uygun bir şekilde ilişkilendirilmek yerine, yanlış bir şekilde yapılandırılabilir. Alan yazında kişilerin kavramları bilimsel çevrelerce kabul edilmiş tanımlarından farklı olarak anlamlandırmasına; "kavram yanılgısı", "ön kavrama", "alternatif çatı", "alternatif kavramlar", "çocukların bilimi" gibi çok çeşitli isimler verilebilmektedir (Treagust, 1988). Bu çalışmada yanlış yapılandırılmış kavramlar "kavram yanılgısı" olarak isimlendirilecektir.

Kavram yanılgısı bireylerin dikkatsizlikleri nedeniyle yaptıkları bir hata veya bilmediklerinden dolayı bir soruya yanlış cevap vermeleri değildir. Kavram yanılgısına sahip olma durumu kişilerin zihinlerinde bir kavramın bilimsel tanımından farklı olarak yapılanması ve değişime dirençli bir şekilde gerekçeleri ile birlikte yerleşmesidir. Kişiler hatalarının doğru olduğunu sebeplendirerek açıklıyor ve kendilerinden emin olduklarını söylüyorlarsa o zaman kavram yanılgılarına sahip oldukları söylenebilir. Sahip olunan bütün kavram yanılgıları birer hatadır ama bütün hatalar birer kavram yanılgısı değildir (Eryılmaz ve Sürmeli, 2002). Bu nedenle kavram yanılgılarını diğer hatalardan ayırma süreci zorlaşmaktadır. Bu zorlu süreç kavram yanılgılarını doğru tespit edebilmek amaçlı birçok araç ve yöntem geliştirilmesine neden olmuştur. Doğru tespit amacıyla geliştirilen araç ve yöntemler onları diğer hatlardan ayırt edebilecek nitelikte olmalıdır.

Kavram yanılgılarını belirlemek amacıyla en çok kullanılan yöntem mülakattır. Fakat bu yöntem gerek uygulama sırasında, gerekse bulguları yorumlama aşamasında araştırmacı tarafından karşılaşılabilecek birtakım zorluklar barındırmaktadır (White ve Gunstone, 1992). Bu zorluklar arasında; analizlere araştırmacı sübjektifliğinin karışması ve bu yöntemi kullanmak isteyen araştırmacıların yöntemle ilgili yeterli bilgilerinin olmaması yer almaktadır (Karataş, Köse ve Coştu, 2003). Bu nedenlerle uygulanması ve yorumlanması mülakatlara göre daha kolay olan çoktan seçmeli testler, alan yazında mülakatlardan sonra en çok kullanılan yöntem olmuştur. Fakat çoktan seçmeli test yöntemi ölçülen kavramla ilgili muhakeme konusunda yetersiz kalabilir (Karataş, Köse ve Coştu, 2003). Bu nedenle kavram yanılgılarını ölçmede yetersiz veya hatalı sonuçlar verebilirler. Bu gerekçeler göz önüne alınarak 1988 yılında temelleri Treagust tarafından atılan iki aşamalı kavram testleri günümüzde pek çok araştırmacı tarafından kavram yanılgılarını belirlemek amaçlı kullanılabilmektedir. Alanyazın incelendiğinde bu testlerin iki veya üç aşamalı formlarda kullanıldığı görülmektedir. İki aşamalı testlerin ilk aşaması çoktan seçmeli veya doğru yanlış test formatında olabilmektedir. Bu testlerdeki soruların ikinci aşamasında birinci aşamada seçilen seçeneğin nedeninin belirtilmesi istenmektedir. Eğer üçüncü aşama var ise bu da seçilen seçeneklerin doğruluğundan emin olunup olunmadığının sorulduğu bölümü içermektedir (Eryılmaz, 2010). Bu testlerin ikinci aşaması araştırmacının amacına göre açık uçlu, çoktan seçmeli veya bir seçeneği açık uçlu çoktan seçmeli bir yapıda olabilmektedir. İkinci aşama "çünkü" bağlacıyla başlayarak ilk aşamada seçilen seçeneğin gerekçelendirilmesine imkân tanımaktadır. Bu testler öğrencilerin düşünce yapılarını daha açık bir biçimde ortaya koymasından dolayı, bilinen çoktan seçmeli testlere göre öğrencilerin kavramları nasıl anladıklarını ortaya koymada daha başarılı olabilir. Buna ek olarak iki aşamalı testler, kavram yanılgılarını ölçmede en çok tercih edilen mülakat yöntemine göre de araştırmacıya daha kolay yorumlanabilir sonuçlar sunabilme avantajına sahiptir. Bu özellikleri göz önüne alınan iki aşamalı test, bu çalışmada öğrencilerin kavram yanılgılarını ortaya koyabilmek için tercih edilmiştir.

Kavram yanılgıları, öğrencilerin hem ilgili kavramları hem de bu kavramlarla ilişkisi bulunan diğer kavramları öğrenmelerini engellemekte veya geciktirmektedir (Özalp ve Kahveci, 2011). Bu nedenle kavram yanılgılarının tespit edilmesi ve giderilmesi için gerekli çalışmaların yapılması son derece önemlidir. Alanyazın incelendiğinde daha çok fen, matematik ve sağlık alanlarına konu olan kavram yanılgısı terimi, ölçme ve değerlendirme alanında çalışan uzmanları da yakından ilgilendirmektedir. Eğitimde ölçme ve değerlendirme dersi, içerisinde istatistik ve matematikle ilişkili kavramlarını da barındıran kapsamlı bir derstir. Ölçme ve değerlendirme ders içeriğinde, doğru yapılandırılamaz ise yanılgı yaşanabilecek kavramlar olabilir. Örneğin ne için kullanıldığı öğrenilmeden ezberlenmiş bir standart sapma formülü, öğrencilerin bu kavramla ilgili bilimsellikten uzak yorumlar yapmalarına neden olabilir. Hatta öğrenciler bu bilimsellikten uzak yorumları destekleyici nedenler de öne sürebilirler. Bunlara ek olarak eğitimde ölçme ve değerlendirme ders içeriğinde öğrencilerin karıştırabilecekleri kavram ikilileri de yer alabilmektedir. Buna örnek olarak geçerlik ve güvenirlik kavramları verilebilir. Bu kavramların tam olarak ne olduğu, birbirleri ile ilişkileri vb. genel olarak öğrencilerin zihinlerinde soru işaretleri bırakabilen konular arasındadır. Bahsedilen örnekler ve daha niceleri ölçme ve değerlendirme uzmanları tarafından incelenebilir. Bu incelemeler sonucunda öğrencilerde yaygın bir şekilde var olan kavram yanılgıları tespit edilebilir. Ulusal düzeydeki çalışmalar incelendiğinde ölçme ve değerlendirme dersindeki kavram yanılgıları ile doğrudan ilişkili çalışmaların oldukça az olduğu görülmektedir. Doğrudan ilişki çalışmalardan ilki Arık'ın (2006)'da yürüttüğü çalışmasıdır. Arık yaptığı çalışmada öğretmenlerin ölçme ve değerlendirme kavramlarını nasıl algıladıklarını belirlemek üzere kendi geliştirdiği 13 soruluk iki aşamalı kavram testini kullanmıştır. Üztemur (2013)'da ölçme ve değerlendirme dersindeki kavram yanılgılarını belirlediği çalışmasında Arık'ın 2006'da geliştirdiği kavram testini kullanmıştır. Bu çalışma kapsamında geliştirilen testin ölçme ve değerlendirme dersindeki kavram yanılgılarını belirlemede kullanılabilecek ikinci test olduğu söylenebilir.

Araştırmanın Amacı

Araştırmanın amacı eğitimde ölçme ve değerlendirme dersini almış öğretmen adaylarının ders kapsamında belirlenen bazı kavramlardaki yaygın yanılgılarının neler olduğunu iki aşamalı bir test ile belirlemektir.

YÖNTEM

Bu araştırma öğretmen adaylarının kavram yanılgılarını ortaya koymaya yönelik betimsel bir araştırmadır.

Çalışma Grubu

Çalışma grubu 2014-2015 öğretim yılında Hacettepe, Kırıkkale ve Akdeniz Üniversitelerinin Eğitim Fakültelerinde öğrenim gören öğretmen adayları arasından seçilen 328 öğrenciden oluşmaktadır. Araştırmaya katılan bütün öğrenciler eğitimde ölçme ve değerlendirme dersini çalışmaya katılmadan önce almış ve geçerli not alarak derslerinde başarılı olmuşlardır. Uygulamaya katılan öğrencilerin tümü çalışmaya gönüllü olarak dâhil olmuştur.

Veri Toplama Araçları

Bu araştırmada veri toplama aracı olarak araştırmacılar tarafından geliştirilen iki aşamalı kavram testi kullanılmıştır. Tanı testinde araştırmacılar tarafından geliştirilen soruların yanında alan yazında daha önce kullanılmış dört soruya da yer verilmiştir. Alan yazındaki bu sorular ilgili araştırmacıdan gerekli izinler alınarak teste dâhil edilmiştir. Bu sorulara, teste dâhil edilme aşamasında birtakım değişiklikler yapılmıştır. Sevimli (2010)'nin uyarlama çalışması yaparak tezinde kullandığı test

istatistik dersindeki kavram yanılgılarını ölçmek amaçlı geliştirilmiş çoktan seçmeli bir testtir. Bu testten alınan sorular iki aşamalı test formuna dönüştürülmüştür.

Tanı testinin geliştirilme aşamasında ilk olarak, test kapsamında yer alacak kavramlar uzman görüşleri ve literatür taraması ile belirlenmiş, ve iki aşamalı testin ilk aşaması oluşturulmuştur. İlk aşaması çoktan seçmeli ikinci aşaması açık uçlu test 78 öğrenciye uygulanarak ikinci aşama için gerekli veriler toplanmıştır. Son aşamada ise tanı testi geliştirilmiştir. Testin puanlarının geçerliliği uzman görüşü alınarak incelenmiştir. Uzmanlar geliştirilen testin ölçme ve değerlendirme dersindeki kavram yanılgılarını ortaya koymak amacıyla kullanılabileceği yönünde görüş bildirmişlerdir. Testin kararlılık anlamında güvenilir sonuçlar üretip üretmediğine ise test-tekrar test yöntemiyle bakılmıştır. Testin hesaplanan güvenirlik katsayısı 0.81'dir.

Her iki aşaması dört seçenekli kavram testindeki her bir soruda 16 farklı seçenek ikilisi oluşabilmektedir. Bu seçenek ikililerinden biri doğru 15'i yanlıştır. Seçilen her yanlış kombinasyonu kavram yanılgısı olarak değerlendirmek doğru değildir. Kavram yanılgılarının belirlenmesinde Tablo 1'deki sınıflandırılmadan faydalanılmıştır.

Tablo 1. İki Aşamalı Kavram Testlerinde Seçilen Kombinasyonların Sınıflandırılması

		İkinci Aşamanın Sonuçları	
		Doğru	Yanlış
Birinci Aşamanın Sonuçları	Doğru	Doğru kavram bilgisi	Çoğunlukla yanlış gerekçeli doğru *Nadiren kavram yanılgısı
	Yanlış	Doğru gerekçeli yanlış	*Çoğunlukla kavram yanılgısı Nadiren hata

Kaynak: Eryılmaz, A. (2010). Development and application of three-tier heat and temperature test: Sample of bachelor and graduate students. *Egitim Arastırmaları – Eurasian Journal of Educational Research*, 40, 17-31.

Tablo 1 incelendiğinde kavram yanılgıların çoğunlukla iki aşamanın da yanlış olduğu seçenek ikileri arasında olabileceği belirtilmiştir. Bununla birlikte ilk aşamanın doğru ikinci aşamanın yanlış olduğu seçenek ikilileri de nadiren kavram yanılgısı barındırmaktadır. Çalışma kapsamında öğrenciler tarafından işaretlenen seçenek ikilileri Tablo 1 doğrultusunda incelenmiş ve kendi içinde bir tutarlılık barındırıran yanlış ikililer kavram yanılgısı olarak değerlendirmeye alınmıştır. Daha sonra bu ikililerden %10 ve üzerinde olanlar yaygın kavram yanılgısı olarak kabul edilmiş ve bulgu olarak sunulmuştur.

İşlem

Testin uygulanması için gerekli sınıf ortamı araştırmaya katılan grupların öğretim elemanları tarafından sağlanmıştır. Testin hem iki aşamalı olması hem de 17 sorudan oluşması öğrencilerin cevaplama isteklerini oldukça düşürmüştür. Gerekli motivasyonun sağlanması için araştırmaya katılan öğrencilere testte en yüksek puanı alan kişiye hediye verileceği çalışmanın başında belirtilmiştir. Testin cevaplanması için öğrencilere zaman kısıtlaması verilmemiştir. Bununla birlikte öğrenciler yaklaşık olarak 25-30 dakika arasında testteki soruları tamamlamıştır.

Verilerin Analizi

On altı seçenek ikilisi oluşan testin verileri 1'den 16'ya numaralandırılarak kayıt edilmiştir. Verilerin analizinde frekans ve yüzde kullanılmıştır.

BULGULAR

Bu bölümde çalışmada tespit edilen yaygın yanılgılar (grubun %10 ve daha fazlası tarafından seçilen kombinasyonlar) seçilme yüzdeleri ve soru için geçerli doğru cevap ikilisi ile birlikte verilecektir. Makale içeriğinde testte yer alan bütün sorulara yer verilmeyecektir, bütün sorular ekler bölümünde yer almaktadır. Soruların genel şeklini anlamak amaçlı testin birinci sorusu ve bu sorunun öğrenciler tarafından işaretlenme yüzdeleri örnek olarak sunulmuştur.

Soru 1: Bir çalışmada yemek yeme miktarı ve kilo arasındaki korelasyon +0.90, kandaki tiroit hormonu ve kilo arasındaki korelasyon -0.90 olarak hesaplanıyor.

Bu bulgular doğrultusunda aşağıdaki yorumlardan hangisi yapılabilir?

A)Tiroit hormonu ve kilo arasında bir ilişki bulunamamıştır.

B)Yeme miktarı ile kilo arasındaki ilişki daha güçlüdür.

C)Kilo alma ile yemek yeme arasında %90 ilişki vardır.

D)Kandaki tiroit hormonu ile kilo arasında yüksek düzeyde bir ilişki bulunmuştur.

Çünkü

1)İki değişken arasında, istatistiksel bir ilişki varsa, pozitif bir korelasyon vardır.

2)Korelasyon yüzde olarak ifade edilebilir.

3) Korelasyon en yüksek ilişki değerine -1.00 ve +1.00 de ulaşır.

4)Korelasyon katsayısı -1.00'den 1.00'e doğru gittikçe güçlenir.

Tablo 2'de öğrencilerin birinci soruda seçtikleri kombinasyonların seçilme yüzdeleri yer almaktadır.

Tablo 2. Öğrencilerin Birinci Soruya Verdikleri Cevapların Dağılımı

			Birinci Aşama			
			A	B	C	D*
İkinci Aşama	1	N	5	20	6	5
		%	1.5 %	6.1 %	1.8 %	1.5 %
	2	N	1	0	24	2
		%	0.3 %	0 %	7.3 %	0.6 %
	3*	N	3	24	5	108*
		%	0.9 %	7.3 %	1.5 %	32.9 %
	4	N	3	86	4	18
		%	0.9 %	26.2 %	1.2 %	5.5 %

Tablo 2 incelendiğinde doğru cevap ikilisi olan D3 dışında öğrencilerin %10'dan daha fazla işaretledikleri tek kombinasyon B4'tür. Bu çalışmada öğrencilerin %10'dan fazla işaretledikleri seçenek ikilileri yaygın kavram yanılgısı olabilir yönünde değerlendirmeye alınmış ve toplam 30 seçenek ikilisi yaygın yanılgı olarak sunulmuştur. Tablo 3'de yaygın yanılgılardan önce ilgili sorudaki doğru kavram ikilisi verilmiştir. Daha sonra 17 soruda tespit edilen 30 yaygın yanılgı seçilme yüzdesine göre büyükten küçüğe sıralanarak sunulmuştur.

Tablo 3. Doğru Kavram İkilisi, Yaygın Yanılgılar ve Seçilme Yüzdeleri

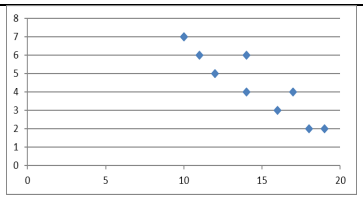
Soru Numarası	Seçilme yüzdesi	Doğru kavram ikilisi	Seçilme yüzdesi	Yaygın yanılgılar
17	%14,6	<table><tr><td>Soru</td><td>Güçlük</td><td>Ayırtedicilik</td></tr><tr><td>1</td><td>0,75</td><td>0,80</td></tr><tr><td>2</td><td>0,25</td><td>0,20</td></tr><tr><td>3</td><td>0,75</td><td>0,20</td></tr><tr><td>4</td><td>0,25</td><td>0,80</td></tr></table> "Tablodaki istatistiklere göre en ayırt edici ve zor olan soru 4'üncü sorudur. Çünkü sorunun hem ayırt edici hem de zor olması için ayırt edicilik gücü 1'e, güçlük indeksi 0'a yaklaşmalıdır."	%50,9	"Tablodaki istatistiklere göre en ayırt edici ve zor olan soru 1'inci sorudur. Çünkü sorunun hem ayırt edici hem de zor olması için ayırt edicilik gücü 1'e, güçlük indeksi 1'e yaklaşmalıdır."
6	%14,3	"İlköğretim 3. sınıf öğrencilerinin zekâ düzeylerinin normal dağılım göstermesini beklerim; çünkü bir veri grubunun normal dağılım göstermesi frekans eğrisinin çan şeklinde ve simetrik olmasına bağlıdır."	%50,6 %9,8	"İlköğretim beşinci sınıf öğrencilerinin yaşlarının normal dağılım göstermesini beklerim; çünkü bir veri grubunun normal dağılım göstermesi frekans değerlerinin bir değer etrafında toplanmasına bağlıdır." "İlköğretim beşinci sınıf öğrencilerinin yaşlarının normal dağılım göstermesini beklerim; çünkü bir veri grubunun normal dağılım göstermesi her bir alt grubun frekans değerlerinin eşit olmasına bağlıdır"
15	%42,7	"-0,85 düzeyindeki korelasyon yüksek bir ilişkiyi gösterir. Çünkü ilişkinin gücü yorumlanırken korelasyon katsayısının işareti göz önünde bulundurulmaz."	%37,2	"-0,85 olarak hesaplanan korelasyon katsayısı düşük bir ilişkiyi gösterir. Çünkü korelasyon eksi çıktığı için aralarında zayıf bir ilişki olduğu sonucuna ulaşılabilir."
10	%30,8	"Ayşe öğretmenin her öğrenciye almış olduğu puanın %10 fazlasını vermesi ölçme sonuçlarına sistematik hata karıştırmaktır. Çünkü ölçme sonuçları sabit oranlarda değişmiştir."	%36,1	"Her öğrenciye almış olduğu puanın %10 fazlasını vermek sabit hatadır çünkü bütün ölçme sonuçları aynı miktarda etkilenmiştir."
1	%32,9	"Kandaki tiroit hormonu ile kilo arasında bulunan -0.90 korelasyon yüksek düzeyde bir ilişkiyi gösterir. Çünkü korelasyon en yüksek ilişki değerine -1.00 ve +1.00'de ulaşır."	%26,2	"+0.90 korelasyon, -0.90'a göre daha güçlü bir ilişkiyi gösterir. Çünkü korelasyon katsayısı -1.00'den +1.00'e doğru gittikçe güçlenir."
14	%4,6	"Maddenin güçlüğü 1'e çok çok yakın ise madde ayırt ediciliği 0'a çok yakın olur. Çünkü maddeyi doğru cevaplayanların çok olması, iyi ve iyi olmayan öğrencileri ayırt etmeyi zorlaştırır."	%24,4	"Madenin güçlüğü 1'e çok çok yakın ise ayırt ediciliği 1'e çok yakın olur. Çünkü maddeyi doğru cevaplayanların çok az olması nedeni ile zayıf öğrenciler daha iyi ayırt edilir."
12	%29,9	"Tutarlık, duyarlık ve kararlık kavramları ölçme aracının güvenirlik özelliği ile ilişkilidir. Çünkü bu kavramlar teste daha az hata karıştığının göstergesidir."	%23,5 %11,6	"Tutarlılık, duyarlık ve kararlık kavramları ölçme aracının geçerlik özelliği ile ilişkilidir. Çünkü bu kavramlar testin ölçmek istediği davranışları ölçtüğünün göstergesidir." "Tutarlılık, duyarlık ve kararlık kavramları ölçme aracının güvenirlik özelliği ile ilişkilidir. Çünkü bu kavramlar testin ölçmek istediği davranışları ölçtüğünün göstergesidir."

Tablo 3. Doğru Kavram İkilisi, Yaygın Yanılgılar ve Seçilme Yüzdeleri (Devam)

<table>
<tr><th>Soru Numarası</th><th>Seçilme yüzdesi</th><th>Doğru kavram ikilisi</th><th>Seçilme yüzdesi</th><th>Yaygın yanılgılar</th></tr>
<tr><td>13</td><td>%6,4</td><td>A ve B grubuna uygulanmış iki ayrı testin istatistikleri<table><tr><th></th><th>Puan Aralığı</th><th>Standart Sapması</th><th>Ortalaması</th></tr><tr><td>A</td><td>0-100</td><td>10</td><td>50</td></tr><tr><td>B</td><td>0-10</td><td>2</td><td>5</td></tr></table>“Tabloya göre B testini alan grup A testini alan gruba göre ölçülen özellik bakımından daha çok farklılık göstermiştir. Çünkü iki grubun farklılaşması karşılaştırılırken standart sapma ve ortalama birlikte dikkate alınır.”</td><td>%22,6

%13,1</td><td>“A testini alan grup B testini alanlara göre daha çok fark göstermiştir. Çünkü iki grubun farklılaşması karşılaştırılırken standart sapma dikkate alınır.”
“Verilen istatistikler yeterli değildir. Çünkü iki grubun farklılaşması karşılaştırılırken basıklık katsayısı verilmeden yorum yapılamaz.”</td></tr>
<tr><td>4</td><td>%17,4</td><td>“3, 4, 5, 6, 6, 8, 10, 12, 19, 26, 83 verilerinden oluşan kümenin ortanca ile temsil edilmesi en uygunu olur. Çünkü veri kümesinin yanlış değerlendirilmesine neden olacak uç değerlerin etkisini azaltmalıyız.”</td><td>%20,1</td><td>“3, 4, 5, 6, 6, 8, 10, 12, 19, 26, 83 verilerinden oluşan kümenin aritmetik ortalama ile temsil edilmesi en uygunu olur. Çünkü bütün değerlerin kullanılarak hesaplandığı bir istatistik seçmeliyiz.”</td></tr>
<tr><td>16</td><td>%12,8</td><td><table><tr><th></th><th>A</th><th>B</th><th>C*</th><th>D</th><th>N</th></tr><tr><td>ÜST GRUP</td><td>11</td><td>30</td><td>15</td><td>4</td><td>50</td></tr><tr><td>ALT GRUP</td><td>15</td><td>11</td><td>4</td><td>20</td><td>50</td></tr></table>“Bu soruda en iyi çalışan çeldirici D seçeneğidir. Çünkü alt grup-üst grup farkının fazla olduğu seçenek en iyi çeldiricidir.”</td><td>%19,8

%17,1</td><td>“Bu soru için en iyi çalışan çeldirici B seçeneğidir. Çünkü üst gruptan en çok kişiyi çeken şık en iyi çeldiricidir.”
“Bu soru için en iyi çalışan çeldirici B seçeneğidir. Çünkü toplamda en çok işaretlenen şık en iyi çeldiricidir.”</td></tr>
<tr><td>8</td><td>%22,6</td><td>“Metrenin birimleri yapaydır. Çünkü ölçme aracı birimleri ortak kararla belirlenmiştir.”</td><td>%14,3

%17,4</td><td>“Metre aracının birimi doğaldır. Çünkü aracın sıfır noktası mutlak sıfırdır.”
“Metre aracının birimi doğaldır. Çünkü ölçme aracı ölçmek istediği özelliği direk olarak ölçer.”</td></tr>
<tr><td>2</td><td>%25,6</td><td>“Düşük puandan yüksek puana doğru sıralamada grubun 95’inci yüzdeliğinde bulunan öğrenci grubun %95’ini geride bırakmıştır. Çünkü yüzdelik küçükten büyüğe sıralanmış verilerin belli bir yüzdesini altında bırakan noktadaki değerdir.”</td><td>%17,1</td><td>“Düşük puandan yüksek puana doğru sıralamada grubun 95’inci yüzdeliğinde bulunan öğrenci grubun %5’ini geride bırakmıştır. Çünkü düşük puandan yüksek puana doğru sıralanmış veri grubunda 95’inci yüzdelik ifadesi ölçümlerin %95’ini üstünde %5’ini altında bulunduran noktadaki değerdir.”</td></tr>
<tr><td>5</td><td>%17,4</td><td> 

%17,4 “Dağılım grafiğinden yararlanarak X ve Y arasındaki korelasyon katsayısını -0,9 olarak tahmin edebiliriz. Çünkü veriler arasında güçlü bir ilişki vardır.”</td><td>%16,2

%13,7

%11,3

%10,7</td><td>“X ve Y arasındaki korelasyon katsayısını 0.0 olarak tahmin edebiliriz. Çünkü veriler arasında ilişki yoktur.”
“X ve Y arasındaki korelasyon katsayısını 0.8 olarak tahmin edebiliriz. Çünkü veriler arasında güçlü bir ilişki vardır.”
“X ve Y arasındaki korelasyon katsayısını -0.9 olarak tahmin edebiliriz. Çünkü veriler arasında zayıf bir ilişki vardır.”
“X ve Y arasındaki korelasyon katsayısını -0.1 olarak tahmin edebiliriz. Çünkü veriler arasında zayıf bir ilişki vardır.”</td></tr>
</table>

Tablo 3. Doğru Kavram İkilisi, Yaygın Yanılgılar ve Seçilme Yüzdeleri (Devam)

Soru Numarası	Seçilme yüzdesi	Doğru kavram ikilisi	Seçilme yüzdesi	Yaygın yanılgılar
9	%11,9	%11,9 "Uygulama sonunda sınav kâğıtlarına 0 ile 100 arasında puanlar verme işlemi ölçmedir. Çünkü gözlem sonuçları sayılarla ifade edilmiştir."	%14,2	"Uygulama sonunda sınav kâğıtlarına 0 ile 100 arasında puanlar verme işlemi değerlendirmedir. Çünkü bu işlemde ölçümler ölçütlerle karşılaştırılarak karara varılmıştır."
			%14,2	"Uygulama sonunda sınav kâğıtlarına 0 ile 100 arasında puanlar verme işlemi ölçüt belirlemedir. Çünkü ölçme kuralı belirlenmiştir."
3	%51,2	"Serideki en yüksek puanın değişmesi ile medyan değişmez, standart sapma değişir. Çünkü bir serideki en yüksek puanın değişimi medyanı etkilemez."	%13,1	"Serideki en yüksek puanın değişmesi ile bir tek ranj değişmez. Çünkü bir serideki en yüksek puanın değişiminden bir tek ranj etkilenmez."
			%11,6	"Serideki en yüksek puanın değişmesi ile medyan değişir, standart sapma değişmez. Çünkü bir serideki en yüksek puanın değişimi standart sapmayı etkilemez."
11	%21,6	"Geçerli her test güvenilirdir. Çünkü testin ölçmek istediği davranışı ölçebilmesi için önce hatalardan arınması gerekir."	%13,1	"Geçerlik ve güvenilirlik birbirini etkileyen kavramlar değildir. Çünkü bunlar bağımsız kavramlardır."
			%11,9	"Güvenilir her test geçerlidir. Çünkü hatalardan arınık bir test ölçmek istediği özelliği ölçebilir."
7	%31,4	"50 ölçüm ile hesaplanan standart sapma değeri -0,8 çıkmış ise standart sapma yanlış hesaplanmıştır; çünkü standart sapma negatif değer alamaz."	%12,5	"Verilerin standart sapması -0,8 ise ölçümlerin çoğu negatiftir. Çünkü standart sapma ölçümler negatif iken negatif değer alır."
			%12,5	"Verilerin standart sapması -0,8 ise ölçümlerin tümü ortalamadan daha küçüktür. Çünkü standart sapma ortalamadan farkların toplamıdır."
			%10,7	"Verilerin standart sapması -0,8 ise ölçümlerin tümü ortalamadan daha küçüktür. Çünkü standart sapma -1 ile +1 arasında değerler alabilir."

Tablo 3 incelendiğinde 17 soruda 30 yaygın yanılgı tespit edildiği görülebilir. Test kapsamında ölçülen kavramlardan en fazla yanılgı yaşananlar maddenin güçlüğü, normal dağılım, korelasyon ve hata türleri olmuştur.

SONUÇLAR ve TARTIŞMA

Bu çalışmada öğretmen adaylarının ölçme değerlendirme dersi ile ilgili kavram yanılgılarının belirlenmesi amaçlanmıştır. Bu bağlamda toplam 17 sorudan oluşan iki aşamalı kavram yanılgıları testi geliştirilmiştir. Araştırma kapsamında geliştirilen bu testin güvenirlik ve geçerlik çalışmaları yapılmış ve testin öğretmen adaylarının ölçme değerlendirme ile ilgili kavram yanılgılarını ortaya çıkaran güvenilir ve geçerli bir ölçme aracı olduğu ortaya konmuştur.

Arık (2006) tarafından geliştirilen 13 soruluk kavram testi ilk aşaması iki, ikinci aşaması 4 seçenekli bir formda hazırlanmıştır. Üztemur (2013) da çalışmasında Arık'ın geliştirdiği testi kullanmıştır. Bu çalışmalarda ilk aşamada yanlış, ikinci aşama doğru seçim yapan bireylerin kavram yanılgısına sahip

olduğu belirtilmiştir. Bu doğrultuda iki çalışmada da en çok "düzeltme puanı" kavramında kavram yanılgısı olduğu belirtilmiştir. Bu çalışmalarda örneğin; "4 yanlışın 1 doğruyu götürmesi işlemi sonucunda elde edilen puana" birinci aşamada "düzeltme puanı" yerine "ham puanı", ikinci aşamada ise bu işlemin yapılma nedeninin "test puanlarının şans başarısından arındırılması" şeklinde işaretleme yapanlar kavram yanılgısına sahip olarak şeklinde kategorileştirilmiştir. Alanyazında ölçme ve değerlendirme alanında yer alan çalışmalardan farklı olarak bu çalışmada, bu tür yanlışlar daha çok kavramların karıştırılması olarak ele alınmış, her iki aşamada da kendi içinde tutarlılık barındıran yanlış ikililer kavram yanılgısı olarak değerlendirilmiştir. Kavram yanılgılarının bu şekilde değerlendirilmesinde Tablo 1'deki sınıflandırmadan faydalanılmıştır.

Hacettepe, Kırıkkale ve Akdeniz Üniversitelerinde öğrenim görmekte olan Eğitim Fakültesi öğrencileri ile yürütülen çalışmanın sonuçları, öğretmen adaylarının bu alandaki birçok kavramda zorluk yaşadıklarını ortaya çıkarmıştır. Testte kavram yanılgısı içeren ve katılımcıların en az %10'u tarafından işaretlenen toplam 30 seçenek ikilisi tespit edilmiştir.

Test genel olarak incelendiğinde öğrencilerin en çok güçlük indeksi kavramında ve normal dağılım gösteren veri kümesini belirlemede yanılgı yaşadıkları söylenebilir. Öğrencilerin güçlük indeksi kavramında yanılgı yaşamalarının nedeni, kavramın Türkçe karşılığının yanıltıcı olması olabilir. 'Güçlük indeksi 0'dan 1'e doğru gittikçe sorunun zorlaşacağı' kulağa birçok kişi tarafından mantıklı gelmektedir. Fakat bu yanlış bir bilgidir. Bu kavramın ismi sorunun 'kolaylık indeksi' olsa belki soru birçok öğrenci tarafından doğru cevaplanmış olacaktı. Bu kavramın Türkçe ifadesinin uzmanlar tarafından tekrar gözden geçirilmesi yararlı olabilir. Çalışmaya katılan öğrencilerin %50,6'sı ilköğretim beşinci sınıf öğrencilerinin yaşlarının, bir değer etrafında toplanması gerekçesiyle normal dağılım göstermesini beklemektedir. Bu noktada normal dağılımla ilgili somut örnekler işe koşulabilir. Bu sayede öğrencilerin kavramı daha net anlaması sağlanabilir. Korelasyon kavramını ölçen sorularda ortak yanlışlar genel olarak eksi korelasyonun ilişki belirtmediği veya artı korelasyonun eksiye göre daha güçlü bir ilişkiyi gösterdiği yönündedir. Korelasyon grafiğini okuma becerisinin ölçüldüğü 5'inci soruda ise öğrencilerin grafiği okumada yetersiz kaldıkları gözlenmiştir. Eksi 0,92 düzeyinde korelasyonu temsil edecek şekilde çizilen grafikte öğrencilerin %33,4'ü grafiği zayıf veya ilişki yok şeklinde yorumlamıştır. Grubun %13,7'si ise veriler arasındaki güçlü ilişkiyi fark etmiş, fakat ilişkinin yönünü pozitif olarak değerlendirerek yine grafiği okumada yetersiz kalmışlardır. Bu yanılgıların önüne bilgisayar desteğini sınıf ortamında kullanarak ve öğrencilere farklı şekillerde verilen veri gruplarının grafikleri çizdirilerek geçilebilir. Öğrencilerin sabit ve sistematik hata kavramlarını karıştırdıkları da testten elde edilen bir sonuçtur. Öğrenciler merkezi eğilim ve dağılım ölçülerinde de yaygın yanılgılar yaşamışlardır. Öğrencilerde yaygın olarak yer alan bir yanılgı da bir maddeyi doğru cevaplayanların çok az olması sonucunda o maddenin çok iyi ayırt edici olacağı yönündedir. Ayrıca testte yer alan ve kapsamı geçerlik ve güvenirlik kavramların özelliklerini ölçen sorularda öğrencilerin bu kavramlara ait özellikleri karıştırdıkları söylenebilir. Katılımcılar puanlama sistemleri farklı iki sınavı, bağıl değişkenlik katsayısı ile yorumlama konusunda da yeterli başarıyı gösterememişlerdir. Öğrencilerin 5'te 1'i veri kümesinde uç değerler olsa dahi aritmetik ortalamanın veri kümesini en iyi temsil eden merkezi eğilim ölçüsü olduğunu düşünmektedir. On altıncı soruda en iyi çalışan çeldirici özelliklerini yansıtmada yaygın yanılgılar yaşayan öğrenciler, üst grup- alt grup yöntemi ile madde ayırt ediciliğini belirlemede de uygun seçenekleri işaretlemede yetersiz kalmışlardır. Testten elde edilen bir diğer sonuçta öğrencilerin ölçme ve değerlendirme kavramlarını karıştırabilecekleri yönündedir. Ayrıca katılımcılar merkezi dağılım ölçülerinden standart sapma kavramında örneğin standart sapmanın eksi çıkabileceği şeklinde yanılgılar göstermişlerdir.

Çalışmanın bulgularında en az doğru cevaplanan sorular ile en yaygın kavram yanılgılarını barındıran soruların eşleşmediği görülmektedir. Bu noktadan hareketle öğrencilerin en az doğru yaptıkları sorularda en yaygın kavram yanılgısı yaşadıklarını savunmanın yanlış olduğu çalışma kapsamında ulaşılan bir sonuçtur. Bu tür bir bulguya çoktan seçmeli testlerle ulaşılamaz. Bir öğrencinin tesadüfi hata ya da dikkat dağınıklığından dolayı çoktan seçmeli bir soruyu yanlış cevaplaması olağandır. Fakat iki aşamalı testlerdeki gibi birbiri ile ilişkili iki yanlış cevap, öğrencinin kavram yanılgısına sahip olduğunu savunmak için daha güçlü deliller sunar. Bunun

yanında iki aşamalı testlerde doğru cevap verilmesi kavramın büyük olasılıkla doğru yapılandırıldığının da kanıtı sayılabilir. Bu noktalar dikkate alındığında iki aşamalı testlerin çoktan seçmeli testlere göre öğrencilerin kavramları nasıl algıladıklarını anlamada daha elverişli olduğu söylenebilir.

Çalışmada kullanılan iki aşamalı kavram testinin her aşaması 4 seçenekli çoktan seçmeli bir formda hazırlanmıştır. Bu şekilde her bir soru için yapısı standart olan bir testin bulgularını karşılaştırmanın anlamlı olacağı söylenebilir. Ayrıca testin her aşamasının 4 seçenekli olması öğrencilere iki aşamada 16 farklı seçenek kombinasyonu sunmuştur. Bu kadar seçenek arasından en az %10 tarafından seçilen yanlış kombinasyonların neredeyse hepsinin kavram yanılgısı özelliği taşıdığı görülmüştür. Chen, Lin ve Lin (2002)'nin lise öğrencilerinin düzlem aynalarla ilgili kavram yanılgılarını ölçmeyi amaçladıkları çalışmalarında da %10 dan fazla işaretlenen seçenek ikilileri olası kavram yanılgısı olarak değerlendirilmeye alınmış ve bu seçenek ikililerinin çoğunlukla kavram yanılgısı olduğu görülmüştür.

Kavram yanılgıların tespit edilmesi ve giderilmesi noktasında öğreticilere büyük görev düşmektedir. Her konunun sonunda kavram yanılgılarını ortaya koyacak kısa testler uygulanabilir ve bu sayede kavram yanılgıları saptanabilir. Bu saptamalar doğrultusunda düzenlenecek eğitim ortamları ile yanılgıların giderilmesi kolaylaşabilir.

İki aşamalı testlerin hazırlanması ve geliştirilmesi zorlu bir süreçtir. Fakat test geliştirildikten sonra uygulanması ve bulgularının analiz edilmesi oldukça hızlı olabilir. Alan uzmanları öğrencilerin kavramları nasıl algıladıklarını ortaya koyacak geçerlik ve güvenirliği yüksek iki veya üç aşamalı testler geliştirebilirler.

İki aşamalı testlerin kavram yanılgılarını belirlemenin yanında eğitimin birçok alanında kullanımının da yararlı olacağını düşünülebilir. Çünkü bu testler yüzeysel öğrenmelerden ziyade anlamlı öğrenmelere odaklanmaktadır. Yani bu testlerin doğruları da, yanlışları da araştırmacıya değerli sonuçlar sunmaktadır denebilir.

Araştırmada örneklem uygun örnekleme yöntemi ile oluşturulduğu için bölümler arası homojen bir dağılım sağlanamamıştır, bu konudaki başka araştırmalarda örneklem evreni daha iyi temsil edilecek şekilde seçilebilir. Araştırmaya dâhil edilemeyen eğitim fakültesi bölümleri mevcuttur, o bölümler başka çalışmalarda incelenebilir.

Araştırmada kullanılan test eğitimde ölçme ve değerlendirme dersi kapsamındaki kavramların birçoğunu kapsayacak şekilde geliştirilmiştir. Bu konuyla ilgilenecek araştırmacılar konuları bölerek daha kısa ve katılımcıları daha kolay motive edebilecekleri testler geliştirebilirler. Konuları bölmek soruları birbiri ile daha rahat ilişkilendirip daha kapsamlı sonuçlar elde edilmesine yardımcı olabilir.

Testten özellikle yaygın kavram yanılgılarını ortaya koyan sorular seçilip üzerine yeni sorular eklenebilir. Bu işlem öğretmen adaylarının ölçme ve değerlendirme dersinde yaşadıkları zorlukları daha geniş bir perspektiften görmeye imkân sağlayabilir.

Çalışmada bazı kavram yanılgıları %35'in üzerinde tespit edilmiştir. Bu yanılgıların derinlemesine incelenmesi başka bir araştırmanın konusu olabilir. Yanılgıların sebepleri ve çözüm önerileri incelenebilir.

KAYNAKÇA

Arık, S. R. (2006). *İlköğretim öğretmenlerinin ölçme ve değerlendirme alanındaki kavram yanılgılarının belirlenmesi* (Yayımlanmamış Yüksek Lisans Tezi, Ankara). http://tez2.yok.gov.tr/ adresinden edinilmiştir.

Eryılmaz, A. (2010). Development and application of three-tier heat and temperature test: Sample of bachelor and graduate students. *Eğitim Arastirmalari – Eurasian Journal of Educational Research*, 40, 17-31.

Eryılmaz, A. ve Sürmeli. E. (2002, Eylül). *Üç aşamalı sorularla öğrencilerin ısı ve sıcaklık konularındaki kavram yanılgılarının ölçülmesi*. V. Ulusal Fen Bilimleri ve Matematik Kongresi, ODTÜ, Ankara.

Chen, C. C., Lin, H. S., & Lin, M. L. (2002). Developing a two-tier diagnostic instrument to assess high school students' understanding-the formation of images by a plane mirror. *Proceedings of National Science Council, 12*(3), 106-121.

Karataş, F. Ö., Köse, S. ve Coştu, B. (2003). Öğrenci yanılgılarını ve anlama düzeylerini belirlemede kullanılan iki aşamalı testler. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 13*(1), 54-69.

Özalp, D. ve Kahveci, A. (2011). Maddenin tanecikli yapısı ile ilgili iki aşamalı tanılayıcı soruların ontoloji temelinde geliştirilmesi. *Milli Eğitim Dergisi, 40*(191), 135-155.

Senemoğlu, N. (2003). *Gelişim öğrenme ve öğretim.* Ankara: Gazi Kitabevi.

Sevimli, N. E. (2010). *Matematik öğretmen adaylarının istatistik dersi konularındaki kavram yanılgıları; istatistik dersine yönelik öz yeterlilik inançları ve tutumlarının incelenmesi* (Yüksek Lisans Tezi, Marmara Üniversitesi Eğitim Bilimleri Enstitüsü, İstanbul). http://tez2.yok.gov.tr/ adresinden edinilmiştir.

Şahin, Ç. ve Ersoy, E. (2009). Sınıf öğretmeni adaylarının yeni ilköğretim programındaki ölçme-değerlendirme konusundaki yeterlilik düzeylerine ilişkin algıları. *Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 18*(2), 363-386.

Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' mis-conception in science. *International Journal of Science Education, 10*(2), 159-169.

Ülgen, G. (2001). *Kavram geliştirme: Kurallar ve uygulamalar.* Ankara: Pegem.

Ülgen, G. (2004). *Kavram geliştirme kuramlar ve uygulamalar* (4. Baskı) Ankara: Nobel.

Üztemur, S. S. (2013). *Sosyal bilgiler öğretmenlerinin ölçme ve değerlendirme alanındaki kavram yanılgıları ve öz-yeterlik inançlarının incelenmesi* (Yüksek Lisans Tezi). Celal Bayar Üniversitesi, Manisa. http://tez2.yok.gov.tr/ adresinden edinilmiştir.

White, R. T., & Gunstone, R. F. (1992). *Probing understanding.* London: The Falmer Press.

EXTENDED ABSTRACT

Introduction

The ability to implement activities properly is an important part of both the assessment as well as evaluation processes within all areas of education. The first condition in order for these processes to function properly is (thus) the proper structuring of concepts within a given field. Newly acquired concepts may however be improperly structured instead of either being associated with the knowledge present in students' minds, or instead of the application of proper configurations. In this study, these improperly structured concepts have been labelled as "misconception(s)"

The situation of having misconceptions entails the structuring of a concept as being from that concept's academic/scientific definition, together with and with the implanting of facts in a manner that is resistant to chance into the minds of individuals. In essence, all misconceptions are mistakes, however not all mistakes are misconceptions (Eryılmaz and Sürmeli, 2002). Consequently, the process of distinguishing misconceptions from other mistakes becomes more challenging. This challenging process has led to the development of many tools and methods for accurately identifying misconceptions. Methods that are developed with the aim of accurate identification should be of a sufficient level of quality in order to distinguish misconceptions from other forms of mistakes.

Two-tier concept tests, as developed in 1988 by Treagust, are now used by many researchers in order to assess misconceptions. The second tier of this test asks participants to specify their reasons for in the choices they made in the first tier. These tests may be more successful than commonly known multiple-choice tests with respect to revealing how students understand concepts, as they reveal the thinking patterns of students in a more explicit manner. Moreover, two-tier tests carry the advantage of offering researchers results that are easy to interpret.

Misconceptions either hinder or delay students' acquisition of relevant concept as well as other concepts that correlate with these concepts (Özalp and Kahveci, 2011). Therefore, studies regarding the determining and eliminating of misconceptions of utmost importance. An assessment and evaluation course within education is a comprehensive course that also encompasses concepts that are part of statistics and mathematics. There may be concepts that should be properly structured correctly within the course content that, if not, could otherwise lead to misconceptions. For example, a standard deviation formula memorized without reason may lead to students interpret concepts in an

unscientific manner. Students may offer reasons supporting the logic behind such unscientific interpretations. In addition, the course in question may contain concept pairs that students can easily confuse. The concepts of validity and reliability are good examples of this. What these concepts are exactly and how they are related to each other etc., are among any number of things that leave students with more questions than answers. These examples alongside many others can be investigated by assessment and evaluation experts. Moreover, misconceptions that are commonly found in and among students may be identified in the results of analysis.

Method

This is a descriptive study aimed at revealing misconceptions held by student teachers. The study group was formed by selecting 328 prospective student teachers studying in the respectful education faculties of Hacettepe, Kırıkkale and Akdeniz University during the 2014-2015 academic year.

We used the two-tier diagnostic test as a data collection tool developed by researchers in this study. 16 different combinations could be formed for each question in the concept test, which consists of four choices in the first tier as well as four choices at the second tier. One of these choice combinations is correct, while the remaining 15 are wrong. It is not right to consider every wrong choice combination as being a misconception since a choice combination needs to have consistency within itself in order to be identified as a misconception. Therefore, we determined the combinations chosen by the participants as matching misconceptions. We then considered those that were 10% or above as being common misconceptions, and then presented these as part of the findings.

Results and Discussion

We determined that a total 30 choice combinations contained misconceptions (10% or above) from the 17 questions contained within the test.

Here, one can note that in the overall examination of the test that the students most commonly have misconceptions when it comes to the concepts of difficulty index and determining data set with normal distribution. Common wrong choices in the questions assessing the concept of correlation generally states that negative correlation does not indicate any relationship whatsoever, or positive correlation shows stronger relation when compared to negative correlation. Students were observed as not being fully capable of reading the graphic in the fifth question assessing correlation graphic reading skills. This misconception may be avoided by using computer assistance in classroom, as well as by making students plot graphics of data sets provided in various ways. Another misconception commonly found among students was that an item is more distinctive when answered correctly by few. Moreover, it can be noted that students confuse attributes of the concepts of validity and reliability in the questions assessing knowledge of these concepts. Students furthermore demonstrated misconceptions when it came to standard deviation, which is one of the central distribution measurements—this can be seen as being negative.

It is very normal for a student to choose a wrong answer due to a lack or distraction of attention. However, two wrong answers that correlate with one other provide us with stronger evidence that the student has a misconception. Similarly, answering two-tier tests correctly may be evidence for the proper structuring of a/the concept in question. Given these facts, it can be noted that two-tier tests are better in understanding how students comprehend concepts as compared to multiple choice tests.

Four choices at each tier of the test provided 16 different choice combinations at two tiers to the students. We observed that a minimum of 10% of wrong combinations made by students bears the attributes of misconception.

Instructors have an important role in determining and eliminating misconceptions. Short tests to reveal misconceptions can be applied at the end of each subject. Misconceptions can thus be

identified/determined. The elimination of misconceptions can be facilitated by means of the education environments improved in line with such determinations.

EKLER

Ek 1

EĞİTİMDE ÖLÇME VE DEĞERLENDİRME KAVRAM TESTİ

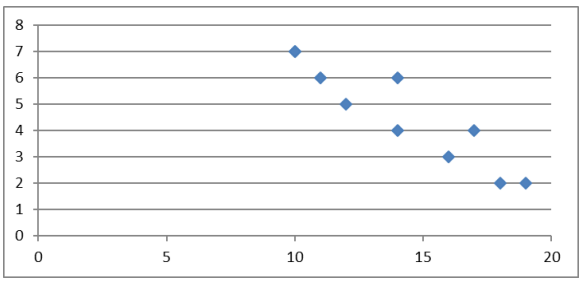
1. Bir çalışmada yemek yeme miktarı ve kilo arasındaki korelasyon +0.90, kandaki tiroit hormonu ve kilo arasındaki korelasyon -0.90 olarak hesaplanıyor.
Bu bulgular doğrultusunda aşağıdaki yorumlardan hangisi yapılabilir?
A) Tiroit hormonu ve kilo arasında bir ilişki bulunamamıştır.
B) Yeme miktarı ile kilo arasındaki ilişki daha güçlüdür.
C) Kilo alma ile yemek yeme arasında %90 ilişki vardır.
D) Kandaki tiroit hormonu ile kilo arasında yüksek düzeyde bir ilişki bulunmuştur.
Çünkü
1) İki değişken arasında, istatistiksel bir ilişki varsa, pozitif bir korelasyon vardır.
2) Korelasyon yüzde olarak ifade edilebilir.
3) Korelasyon en yüksek ilişki değerine -1.00 ve +1.00 de ulaşır.
4) Korelasyon katsayısı -1.00'den 1.00'e doğru gittikçe güçlenir.

2. Bir öğrenci matematik sınavı için hesaplanan puanı ile düşük puandan yüksek puana doğru sıralamada grubun 95'inci yüzdeliğinde ise aşağıdakilerden hangisi daima doğrudur?
A) Öğrencinin notu A olacaktır.
B) Öğrenci en azından alınabilecek bütün puanların %95'ini kazanmıştır.
C) Öğrenci grubun %95 ini geride bırakmıştır.
D) Öğrenci grubun %5'ini geride bırakmıştır.
Çünkü
1) Yüzdelik küçükten büyüğe sıralanmış verilerin belli bir yüzdesini altında bırakan noktadaki değerdir.
2) 95'inci yüzdelik ifadesi ölçümlerin %95 ini üstünde %5 ini altında bulunduran noktadaki değerdir.
3) 95'inci yüzdelik en fazla 100 alınabilecek bir sınavdan 95 alan bireylerin yerinin gösterildiği noktadır.
4) 95'inci yüzdelik grubun baştan 5. sırasındaki bireydir.

3.Bir sınıftaki öğrencilerin notları 83, 88, 90, 88, 86, 84 tür. Bu sınıfta 90 alan öğrencinin puanı arttırılırsa,
A) Medyan değişir, standart sapma değişmez.
B) Medyan değişmez, standart sapma değişir.
C) Medyan da standart sapma da değişir.
D) Bir tek ranj değişmez.
Çünkü
1) Bir seride medyanı ve standart sapmayı değiştirmeden serideki sayılar arttırılamaz.
2) Bir seride ki en yüksek puanın değişiminden bir tek ranj etkilenmez.
3) Bir seride ki en yüksek puanın değişimi medyanı etkilemez.
4) Bir seride ki en yüksek puanın değişimi standart sapmayı etkilemez.

4.Aşağıda sayıların bir kümesi verilmiştir. Bu dağılımın hangi merkezi eğilim ölçüsü ile temsil edilmesi en uygundur?
3, 4, 5, 6, 6, 8, 10, 12, 19, 26, 83
A) Aritmetik Ortalama
B) Ortanca
C) Tepe değeri
D) Geometrik Ortalama
Çünkü
1) 83 uç değerini de dâhil edebileceğimiz bir istatistik seçmeliyiz.
2) Bütün değerlerin kullanılarak hesaplandığı bir istatistik seçmeliyiz.
3) En fazla frekansın gözlendiği değerler bize grubu daha iyi temsil etme şansı verir.
4) Veri kümesinin yanlış değerlendirilmesine neden olacak uç değerlerin etkisini azaltmalıyız.

5.



Yukarıda X ve Y değişkenleri için verilen dağılım grafiğinden yararlanarak X ve Y arasındaki korelasyon katsayısını ……. olarak tahmin edebiliriz.
A) -0.1 B) 0.0 C) 0.8 D) -0.9
Çünkü
1) Veriler arasında ilişki yoktur.
2) Veriler arasında güçlü bir ilişki vardır.
3) Verilerden biri artarken diğeri de artmaktadır.
4) Veriler arasında zayıf bir ilişki vardır.

6. Aşağıdaki istatistiklerden hangisinin normal dağılım göstermesini beklersiniz?
A) İlköğretim 5. Sınıf öğrencilerinin yaşları
B) İlköğretim 3. Sınıf öğrencilerin zekâ düzeyleri
C) A şehrindeki insanların cinsiyetleri
D) 30 kişiyi aşan grupların özellikleri
Çünkü bir veri grubunun normal dağılım göstermesi;
1) Frekans değerlerinin bir değer etrafında toplanmasına bağlıdır.
2) Her bir alt grubun frekans değerlerinin eşit olmasına bağlıdır.
3) Frekans eğrisinin çan şeklinde ve simetrik olmasına bağlıdır.
4) Gruptaki kişilerin sayıca 30'u aşmasına bağlıdır.

7. Bir bilim insanı yaptığı deney sonunda 50 ayrı ölçme sonucu elde ediyor. Bu verilerin standart sapması -0,8 olduğuna göre aşağıdakilerden hangisi kesinlikle doğrudur?
A) Ölçümlerin çoğu negatiftir
B) Ölçümlerin tümü ortalamadan daha küçüktür.
C) Standart sapma yanlış hesaplanmıştır.

D) Ölçümlerin tümü negatiftir.
Çünkü
1) Standart sapma ortalamadan farkların toplamıdır.
2) Standart sapma -1 ile +1 arasında değerler alabilir.
3) Standart sapma ölçümler negatif iken negatif değer alır.
4) Standart sapma negatif değer alamaz.

8.Masanın uzunluğunu ölçmek için kullanılan ölçme aracı metre ise bu ölçme aracının birimleri hakkında ne söylenebilir?
A) Birimi yapaydır.
B) Birimi kullanışlı değildir.
C) Birimi doğaldır.
D) Birimleri eşit değildir.
Çünkü
1) Ölçme aracının sıfır noktası mutlak sıfırdır.
2) Ölçme aracı birimleri eşit olarak ayrılır.
3) Ölçme aracı birimleri ortak kararla belirlenir.
4) Ölçme aracı ölçmek istediği özelliği direk olarak ölçer.

9. Ali öğretmen, her sorunun 10 puan olmasına karar verdiği sınavı için bir uygulama yapmış. Uygulama sonunda sınav kâğıtlarına 0 ile 100 arasında puanlar vermiştir. Ali öğretmenin yaptığı işlem nedir?
A) Ölçmedir.
B) Değerlendirmedir.
C) Ölçüttür.
D) Ölçümdür.
Çünkü
1) Ölçme sonuçlarını bir ölçütle karşılaştırmıştır.
2) Ölçme kuralını belirlemiştir.
3) Gözlem sonucunu sayılarla ifade etmiştir.
4) Ölçümleri ölçütlerle karşılaştırarak karara varmıştır.

10. Ayşe öğretmenin yaptığı sınavda öğrencileri düşük puanlar almıştır. Bu nedenle Ayşe öğretmen her öğrenciye almış olduğu puanın %10 fazlasını vermiştir. Ayşe öğretmenin yaptığı işlem nedir?
A) Ölçme sonucu ham puan elde etmektir.
B) Ölçme sonuçlarına sabit hata karıştırmaktır.
C) Ölçme sonuçlarına sistematik hata karıştırmaktır.
D) Ölçme sonuçlarına tesadüfî hata karıştırmaktır.
Çünkü
1) Hatanın tespit edilmesi zorlaşmıştır.
2) Bütün ölçme sonuçları aynı miktarda etkilenmiştir.
3) Öğrencilerin başarısızlıklarından kaynaklanmıştır.
4) Ölçme sonuçları belli oranlarda değişmiştir.

11. Aşağıda güvenirlik ve geçerlik ile ilgili olarak verilen karşılaştırmalardan hangisi doğrudur?
A) Geçerli her test güvenilirdir.
B) Geçerlik ve güvenirlik birbirini etkileyen kavramlar değildir.

C) Güvenilir her test geçerlidir.
D) Geçerli her test güvenilir, güvenilir her test de geçerlidir.
Çünkü
1) Bu özelliklerden biri sağlanıyor ise diğeri de sağlanmış olur.
2) Hatalardan arınık bir test ölçmek istediği özelliği ölçebilir.
3) Testin ölçmek istediği davranışı ölçebilmesi için önce hatalardan arınması gerekir.
4) Bunlar bağımsız kavramlardır.

12. Tutarlılık, duyarlılık, kararlılık gibi kavramlar ölçme aracının hangi özelliği ile ilgilidir?
A) Kullanışlılık
B) Geçerlilik
C) Güvenilirlik
D) Objektiflik
Çünkü bu özellikler
1) Teste daha az hata karıştığının göstergesidir.
2) Testin ölçme istediği davranışları ölçtüğünün göstergesidir.
3) Testin çok yorulmadan kullanılabileceğinin göstergesidir.
4) Puanlamaya puanlayıcı hatasının karışmamasıdır.

13.Aşağıda A ve B guruplarına uygulanmış iki ayrı testin istatistikleri yer almaktadır.

A ve B grubuna uygulanmış iki ayrı testin istatistikleri			
	Puan Aralığı	Standart Sapması	Ortalaması
A Testi	0-100	10	50
B Testi	0-10	2	5

Gruplardan hangisi ölçülen özellik bakımından daha çok farklılık göstermiştir?
A) A testini alan grup
B) B testini alan grup
C) İkisi de eşittir.
D) Verilen istatistikler yeterli değildir.
Çünkü iki grubun farklılaşması karşılaştırılırken;
1) Standart sapma dikkate alınır.
2) Puan aralıklarının genişliği dikkate alınır.
3) Basıklık katsayısı verilmeden yorum yapılamaz.
4) Standart sapma ve ortalama birlikte dikkate alınır.

14. Bir testin 8. maddesinin güçlüğü 1'e çok yakınsa, bu maddenin ayırt ediciliği kaç olabilir?
A) 0'a çok yakın
B) -1'e çok yakın
C) +1'e çok yakın
D) Tahmin için uygun veri yoktur.
Çünkü
1) Maddeyi doğru cevaplayanların çok az olması nedeniyle zayıf olan öğrenciler daha iyi ayırt edilebilir.

2) Maddeyi doğru cevaplayanların çok olması, iyi ve iyi olmayan öğrencileri ayırt etmeyi zorlaştırır.
3) Maddeye neredeyse herkesin yanlış cevap vermiş olması, bilenle bilmeyeni ayırt etmeyi zorlaştırır.
4) Madde güçlüğünden elde edilen veri ile madde ayırt ediciliğini tahmin edilemez.

15. Yapılan bir çalışmada lise 1 öğrencilerinin eleştirel düşünme ve problem çözme becerileri arasındaki korelasyon - 0,85 çıkmıştır. Bu sonuçlara göre eleştirel düşünme ve problem çözme arasında nasıl bir ilişki vardır?
A) Düşük bir ilişki vardır.
B) Bir ilişki bulunamamıştır.
C) Yüksek bir ilişki vardır.
D) Korelasyon yanlış hesaplanmıştır.
Çünkü
1) Korelasyon eksi çıktığı için aralarında hiç ilişki yoktur sonucuna varabiliriz.
2) Korelasyon eksi çıktığı için aralarında zayıf bir ilişki olduğu sonucuna varabiliriz.
3) İlişkinin gücü yorumlanırken korelasyon katsayısının işareti göz önünde bulundurulmaz.
4) Korelasyon eksi çıkamayacağı için yanlış hesaplanmıştır diyebiliriz.

16. Bu soru için doğru cevap C ise en iyi çalışan çeldirici hangisidir?
A) A
B) B
C) C
D) D
Çünkü
1) Üst gruptan en çok kişiyi çeken şık en iyi çeldiricidir.
2) Toplamda en çok işaretlenen şık en iyi çeldiricidir.
3) Alt grup- üst grup farkının fazla olduğu seçenek en iyi çeldiricidir.
4) Üst grup- alt grup farkının fazla olduğu seçenek en iyi çeldiricidir.

17.

Soru	Güçlük indeksi	Ayırt edicilik gücü
1	0,75	0,80
2	0,25	0,20
3	0,75	0,20
4	0,25	0,80

Yukarıdaki istatistiklere göre en ayırt edici ve zor olan soru hangisidir?
A) 1.soru
B) 2.soru
C) 3.soru
D) 4.soru
Çünkü sorunun hem ayırt edici hem de zor olması için;
1) Ayırt edicilik gücü 1'e, güçlük indeksi 0'a yaklaşmalıdır.
2) Ayırt edicilik gücü 0'a, güçlük indeksi 1'e yaklaşmalıdır.
3) Ayırt edicilik gücü 1'e, güçlük indeksi 1'e yaklaşmalıdır.
4) Ayırt edicilik gücü 0'a, güçlük indeksi 0'a yaklaşmalıdır.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(4);421-434

Investigation of the Performance of Multidimensional Equating Procedures for Common-Item Nonequivalent Groups Design*

Çok Boyutlu Eşitleme Yöntemlerinin Eşdeğer Olmayan Gruplarda Ortak Madde Deseni için Performanslarının İncelenmesi

Burcu ATAR ** Gonca YEŞİLTAŞ ***

Abstract
In this study, the performance of the multidimensional extentions of Stocking-Lord, mean/mean, and mean/sigma equating procedures under common-item nonequivalent groups design was investigated. The performance of those three equating procedures was examined under the combination of various conditions including sample size, ability distribution, correlation between two dimensions, and percentage of anchor items in the test. Item parameter recovery was evaluated calculating RMSE (root man squared error) and BIAS values. It was found that Stocking-Lord procedure provided the smaller RMSE and BIAS values for both item discrimination and item difficulty parameter estimates across most conditions.

Keywords: Multidimensional equating, mean/mean, mean sigma, Stocking-Lord

Öz
Bu çalışmada çok boyutlu veri için adapte edilen Stocking-Lord, ortlama/ortlama ve ortalama/sigma eşitleme yöntemlerinin performansları eşdeğer olmayan gruplarda ortak madde deseni göz önüne alınarak incelenmiştir. Bu üç eşitleme yöntemin performansları orneklem büyüklüğünün, yetenek dağılımının, boyutlar arasındaki korelasyon değerlerinin ve testteki ortak madde yüzdelerinin kombinasyonlari altinda araştırılmıştır. Madde parametre kestirimlerinin değerlendirilmesinde RMSE ve yanlılık değerleri kullanılmıştır. Bu çalışmada çoğu koşul icin hem madde ayırt edicilik parametre kestirimlerinde hem de madde güçlük parametre kestirimlerinde Stocking-Lord yönteminin diger iki yönteme gore daha kucuk RMSE ve yanlılık değerleri verdiği bulunmuştur.

Anahtar Kelimeler: Çok boyutlu eşitleme, ortalama/ortalama, ortalama/sigma, Stocking-Lord

INTRODUCTION

The main reason to administer tests under standardized conditions is to assess the abilities of examinees fairly and objectively. Scores obtained from the large-scale standardized achievement tests administered in the field of education are sometimes used in important decisions such as selecting students to place into educational programs or institutes based on their abilities. These tests are administered once or more than once within a year. A new form of the test is generally used in each administration for the security purposes in those high-stakes tests. Even though the forms are developed to measure the same construct, they may exhibit differences in their statistical characteristics such as item difficulties and reliabilities. Those differences may give advantage to examinees who take the easier and more reliable form of the test. When the different forms of a test are used, the scores obtained from one administration of the test need to be converted into the scores

* Bu araştırma 1059B191300504 kodlu TUBITAK projesi tarafından desteklenmiştir.
** Doç. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, burcua@hacettepe.edu.tr, ORCID ID: orcid.org/0000-0003-3527-686X
*** Dr., Cumhuriyet Üniversitesi, Sivas-Türkiye, goncayesiltas@cumhuriyet.edu.tr, ORCID ID: orcid.org/0000-0002-7291-7083

DOI: 10.21031/epod.335284
Geliş Tarihi: 18.08.2017
Kabul Tarihi: 21.11.2017

obtained from the previous administration of the test in order to prevent the reflection of the statistical differences among the forms of the test into examinees' scores. Scores obtained from different forms of a test can be compared after a statistical process called equating (Kolen & Brennan, 2004). Otherwise, it is not appropriate to compare the scores obtained from different forms of the test.

There are different test equating procedures for different data collection designs. Common-item nonequivalent groups design is one of the most widely used data collection designs in equating of the different forms of the large-scale standardized achievement tests by testing companies. Among procedures based on the classical test theory, Tucker linear equating procedure (Gulliksen, 1950), Levine linear equating procedure (Levine, 1955), frequency estimation and chained equipercentile equating procedures (Angoff, 1971) can be used with common-item nonequivalent groups design. Procedures based on item response theory (IRT) can also be used with common-item nonequivalent groups design. Item response theory has prominent properties in test equating. One of the advantages of item response theory is the invariance of item and ability parameters when the model fits the data (Lord,1980). Inavariance property of item response theory has an important role in test equating especially under common-item nonequivalent groups design (Skaggs & Lissitz, 1986). The invariance property of item response theory depends on the tenability of the assumptions. One of the assumptions that should be considered in test equating studies based on item response theory is the unidimensionality assumption. However, tests in actual administrations exhibit a multidimensional structure. As a result, unidimensionality assumption is mostly violated in many testing situations (Li & Lissitz, 2000). Depending on the degree of the violation of the assumption and depending on the conditions as sample size, ability distribution, number of anchor item in the test, and so on, the performace of the equating procedure will be effected. Investigating the robustness of those procedures to the violation of unidimensionality assumption is essential. (Camilli, Wang & Fesq, 1995; De Champlain, 1996; Li & Lissitz, 2000; Oshima, Davey & Lee, 2000). When the unidimensionality assumption is not met, procedures based on multidimensional item response theory (MIRT) might be conducted. It is possible with MIRT to model the interaction of items that can discriminate between different ability levels and examinees that have different proficiencies on those levels (Ackerman, 1994). When the forms of a test exhibit multidimensional structure, MIRT equating can be considered. In MIRT equating, the accuracy of parameter estimates after equating process is critical to address (Li & Lissitz, 2000). The performance of procedures based on MIRT should be investigated under different conditions that are similar to actual testing conditions (Yao & Boughton, 2009). By this way, for different test conditions, a practical equating procedure that provides accurate estimates and minimun equating error can be determined.

There are several publications on multidimensional equating/linking (Hirsch, 1989; De Champlain, 1996; Bolt 1999; Li & Lissitz, 2000; Yao & Boughton, 2009; Yao, 2011; Eser & Gelbal, 2015). In some of those studies real test data were used and in the others simulated data was considered. De Champlain (1996) examined the equating results of unidimensional IRT true-score equating procedure on different subgroups of a two-dimensional real test data. Data used in De Champlain's analyses is the one obtained from the administration of two forms of the Law School Admission Test (LSAT). When the equating functions obtained from subgroups are compared with the ones obtained from the whole group, it was seen that the differences along the scale are small although those differences increases toward the lower end of the scale. Bolt (1999) simulated two-dimensional data using the parameter estimates obtained from two forms of the LSAT data. He compared the performance of unidimensional IRT true-score equating procedure with the performance of unidimensional linear and equipercentile equating procedures under different levels of correlation between dimensions. As a result of that study, it was found that IRT true-score equating procedure performs as well as other 2 conventional procedures when the correlation between dimensions is higher. It was also found that IRT true-score equating procedure performs better than the others when the correlation between dimensions is lower. Yao and Boughton (2009) examined the linking accuracy of test response function procedure under multidimensional perspective for tests including both dichotomously and polytomously scored items. In their study, they used simulated two-dimensional data under different conditons of population distribution, anchor set length, and item

structure. They found that parameter recovery was good across all conditons with a well chosen anchor set. Yao (2011) investigated the linking accuracy of multidimensional test response function linking procedure for dichotomously scored items using a simulated data with five-dimension. Sample size, population distribution, and anchor set length were varying conditions in that study. It was found that overall score and domain score recovery was good even for condition with the smallest anchor set. In their study, Eser and Gelbal (2015) examined the performance of multidimesional item response theory model parameter estimates of two and three dimensional tests under the combination of different levels of sample size and test length conditions. They found that the parameters were estimated more accurately as the sample size and test length increased.

Considering the multidimensional structure of many testing data, it is critical to investigate the performance of different multidimensional equating/linking procedures in depth and observe which procedure performs better under different conditions. Since it is possible to encounter with various types of data under different settings in real testing, it is important to conduct studies with simulated data under conditions that are close to real testing situations. In this study, sample size, ability distibution, correlation between two dimesions, and percentage of anchor items in a test with 40 dichotomously scored items are considered as varying conditions that are close to real testing situations to compare the performance of multidimensional extensions of Stocking-Lord, mean/mean, and mean/sigma equating procedures under common-item nonequivalent groups design.

Purpose of the Study

The purpose of this study was to investigate the performance of three multidimensional equating procedures in the recovery of item and ability parameter estimates under various simulation conditions. Different simulation conditions were formed with the combination of sample size, ability distribution, correlation between dimensions, and the percentage of anchor items in the test. The equating procedures compared in this study under common-item nonequivalent groups design were the extentions of unidimenisonal equating procedures – Stocking-Lord, mean/mean, and mean/sigma.

METHOD

Simulated data under common-item nonequivalent groups design was used in this study. For the purpose of the study, examinee responses to a 40 dichotomously scored items in a two-dimensional test were generated based on the item parameter estimates from the study of Yao & Boughton (2009). As the structure of the data, some of the items loaded on a single dimension (simple structure), some of the items loaded on both dimensions (complex structure).

SimuMIRT program (Yao, 2003) was used to generate response data under various conditions. In their study, Li & Lissitz (2000) found that a sample size of 2000 with 20 anchor items from a 40-item test was adequate for the multidimensional test response function equating procedure. Factors manipulated and factors held constant were adapted considering real testing situations in this study. Sample size and number of anchor items in the test were two of the factors that were manipulated. Two levels of sample size (1000 and 2000) and four levels of percentage of anchor items in the test (15%, 30%, 60%, and 100%) were considered. In addition to sample size and percentage of anchor items, ability distribution and correlation between dimensions were manipulated. Two levels of ability distribution (for a base of comparison, multivariate normal distribution with mean of 0 and standard deviation of 1 in both dimensions was generated; for other level, mean of -0.5 and 0.5 with standard deviation of 1 on both dimensions were generated) and three levels of correlation between two dimensions (0, 0.5 and 0.8) were taken into account. As a result of the combination of those factors, 48 simulation conditions were generated. Simulation conditions for each level of percentage of anchor items in the test were shown in Table 1. Each simulation condition was replicated 20 times. The computer softwares used here for data simulation, parameter estimation, and linking purposes are softwares developed for data simulation, parameter estimation, and linking. Linking

software has to be run seperately for each analysis as opposed to other softwares. In addition, each output must be examined individually to obtain the essential values. Because of these reasons, if the number of replications increases, it requires longer time to complete the analysis. The number of replications has been limited by Yao and Boughton (2009) as well as by Yue and Hongyun (2013). Test length and the number of dimensions were the factors that were held constant.

BMIRT program (Yao, 2003) was used to estimate the item and ability parameters of the response data under compensatory multidimensional three-parameter (M-3PL) IRT model. BMIRT was found to produce accurate parameter estimates under the M-3PL model (Yao & Boughton, 2007; Yao & Schwarz, 2006).

Equating was conducted using multidimensional Stocking-Lord, mean/mean, and mean/sigma equating procedures. LinkMIRT program (Yao, 2004) was used for equating in all simulation conditions.

To evaluate the item parameter recovery, root mean square error (RMSE), and bias (BIAS) values were calculated using the following formulas:

$$RMSE(\hat{\tau}) = \sqrt{\frac{\sum_{r}^{R}(\hat{\tau}_r - \tau)^2}{R}} \quad \text{and}$$

$$Bias(\hat{\tau}) = \frac{\sum_{r}^{R}\hat{\tau}_r}{R} - \tau$$

where τ is the true value of the parameter, $\hat{\tau}_r$ is the estimated value of the parameter at the r^{th} replication, and R is the number of replications.

Table 1. Simulation Conditions

Condition	Sample Size	Mean1	Mean2	Var-Cov Matrix
C1	1000	0	0	1,0,0,1
C2	2000	0	0	1,0,0,2
C3	1000	0	0	1,0.5,0.5,1
C4	2000	0	0	1,0.5,0.5,2
C5	1000	0	0	1,0.8,0.8,1
C6	2000	0	0	1,0.8,0.8,2
C7	1000	0.5	-0.5	1,0,0,1
C8	2000	0.5	-0.5	1,0,0,2
C9	1000	0.5	-0.5	1,0.5,0.5,1
C10	2000	0.5	-0.5	1,0.5,0.5,2
C11	1000	0.5	-0.5	1,0.8,0.8,1
C12	2000	0.5	-0.5	1,0.8,0.8,2

Note. These twelve conditions were repeated for each level of percentage of anchor item

RESULTS

Item Parameter Recovery

RMSE and BIAS values of the first item discrimination parameter related to the first dimension estimates calculated for the three equating procedures across all conditions are given in Table 2 and Table 3, respectively. RMSE and BIAS values of the first item discrimination parameter estimates

for the levels of the percentage of anchor items are shown in Figure 1 and Figure 2, respectively for each equating procedure

Table 2. RMSE for the First Item Discrimination Parameter

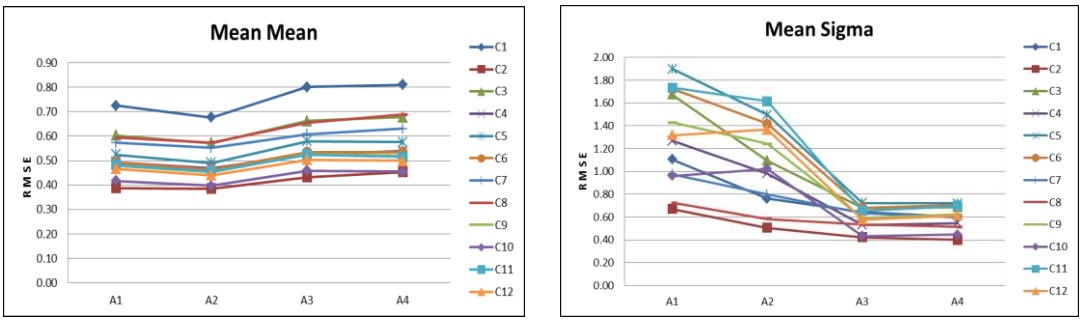
	15%			30%			60%			100%		
	MM	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**
C1	0.73	1.11	0.62	0.68	0.76	0.60	0.80	0.65	0.63	0.81	0.60	0.65
C2	0.39	0.67	0.35	0.39	0.50	0.35	0.43	0.42	0.39	0.45	0.40	0.37
C3	0.60	1.67	0.56	0.57	1.10	0.57	0.66	0.68	0.55	0.68	0.69	0.62
C4	0.49	1.27	0.43	0.47	0.98	0.45	0.52	0.53	0.43	0.54	0.55	0.49
C5	0.52	1.90	0.51	0.49	1.49	0.50	0.58	0.72	0.55	0.58	0.72	0.57
C6	0.49	1.72	0.44	0.46	1.42	0.44	0.54	0.68	0.52	0.54	0.70	0.46
C7	0.57	0.98	0.53	0.55	0.80	0.50	0.61	0.63	0.47	0.63	0.59	0.51
C8	0.59	0.72	0.51	0.57	0.58	0.51	0.66	0.53	0.48	0.69	0.51	0.52
C9	0.48	1.43	0.43	0.45	1.24	0.44	0.53	0.59	0.44	0.52	0.62	0.45
C10	0.42	0.96	0.40	0.40	1.02	0.39	0.46	0.43	0.37	0.46	0.45	0.36
C11	0.48	1.73	0.48	0.50	1.61	0.50	0.52	0.65	0.52	0.52	0.70	0.52
C12	0.47	1.32	0.44	0.44	1.37	0.44	0.50	0.58	0.47	0.50	0.61	0.45

Table 3. BIAS for the First Item Discrimination Parameter

	15%			30%			60%			100%		
	MM	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**
C1	0.53	0.76	0.42	0.50	0.44	0.40	0.59	0.44	0.39	0.59	0.43	0.40
C2	0.25	0.48	0.20	0.25	0.31	0.16	0.26	0.27	0.19	0.27	0.25	0.16
C3	0.50	1.38	0.41	0.47	0.88	0.40	0.55	0.48	0.37	0.56	0.50	0.40
C4	0.40	1.10	0.29	0.38	0.81	0.28	0.43	0.37	0.27	0.44	0.38	0.27
C5	0.45	1.72	0.40	0.42	1.35	0.38	0.51	0.47	0.39	0.51	0.50	0.38
C6	0.44	1.61	0.37	0.42	1.33	0.37	0.49	0.45	0.36	0.49	0.49	0.34
C7	0.39	0.66	0.30	0.38	0.47	0.26	0.42	0.36	0.26	0.43	0.34	0.26
C8	0.41	0.46	0.31	0.40	0.35	0.29	0.45	0.37	0.31	0.47	0.36	0.32
C9	0.40	1.21	0.33	0.38	1.06	0.31	0.44	0.38	0.32	0.44	0.41	0.30
C10	0.34	0.83	0.27	0.33	0.89	0.23	0.37	0.31	0.24	0.37	0.32	0.22
C11	0.41	1.56	0.38	0.39	1.42	0.35	0.46	0.48	0.37	0.45	0.52	0.36
C12	0.42	1.19	0.37	0.39	1.25	0.35	0.46	0.43	0.35	0.46	0.47	0.34

Based on the RMSE values in Table 2 and the BIAS values in Table 3 for the first discrimination parameter, it can be said that Stocking-Lord procedure produced smaller RMSE and BIAS values than mean/mean and mean/sigma procedures under all conditions. Percentage of anchor items in the test factor affected the results of mean/sigma procedure the most. The RMSE and BIAS values calculated for mean/sigma procedure were much higher than the ones calculated for mean/mean and Stocking-Lord procedures when the percentage of anchor items were 15% and 30%. The RMSE values were decreased for mean/sigma procedure as the percentage of anchor items were incerased. Under the 15% and 30% anchor items conditions, the RMSE values for mean/sigma procedure were the smallest when the sample size was 2000 and the correlation between dimensions was 0.

(a) (b)



(c)

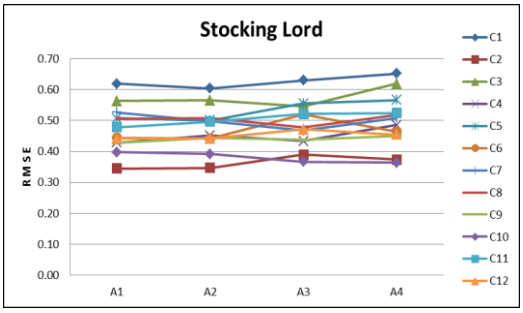
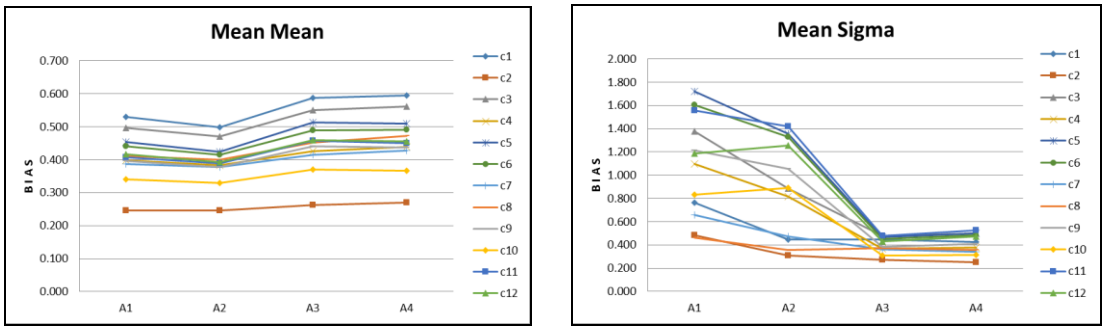


Figure 1. RMSE Values of the First Item Discrimination Parameter for Three Equating Procedures by the Percentage of Anchor Items Under All Conditions

(a) (b)



(c)

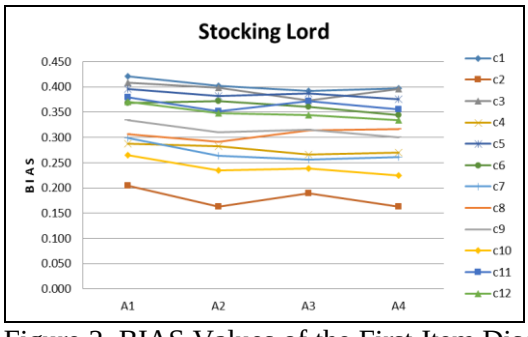


Figure 2. BIAS Values of the First Item Discrimination Parameter for Three Equating Procedures by the Percentage of Anchor Items Under All Conditions

It can be seen in Figure 1 that the percentage of anchor items factor did not effect the results of mean/mean and Stocking-Lord procedures for the first item discrimination parameter much. The most effective factor for those two procedures was the sample size. Mean/mean and Stocking-Lord procedures produced smaller RMSE values when the sample size was larger. Sample size was also an effective factor for mean/sigma method.

RMSE and BIAS values of the second item discrimination parameter related to the second dimension estimates calculated for the three equating procedures across all conditions are given in Table 4 and Table 5, respectively. RMSE and BIAS values of the second item discrimination parameter estimates for the levels of the percentage of anchor items are shown in Figure 3 and Figure 4, respectively for each equating procedure.

Table 4. RMSE for the Second Item Discrimination Parameter

	15%			**30%**			**60%**			**100%**		
	MM	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**
C1	0.74	0.94	0.70	0.87	0.76	0.78	0.82	0.80	0.85	0.88	0.77	0.93
C2	0.45	0.40	0.37	0.57	0.54	0.51	0.55	0.68	0.55	0.59	0.64	0.55
C3	0.66	1.27	0.73	0.78	0.85	0.81	0.74	0.74	0.79	0.81	0.72	0.91
C4	0.53	0.88	0.55	0.60	0.68	0.61	0.58	0.70	0.63	0.63	0.66	0.72
C5	0.61	1.35	0.65	0.67	1.01	0.65	0.63	0.82	0.69	0.67	0.77	0.74
C6	0.54	1.03	0.57	0.60	0.83	0.56	0.57	0.75	0.67	0.60	0.72	0.64
C7	0.69	0.91	0.73	0.83	0.76	0.75	0.80	0.81	0.61	0.88	0.77	0.83
C8	0.73	0.77	0.74	0.89	0.68	0.77	0.87	0.80	0.68	0.95	0.77	0.88
C9	0.62	1.10	0.64	0.71	0.82	0.70	0.65	0.73	0.59	0.72	0.71	0.66
C10	0.52	0.73	0.54	0.61	0.69	0.57	0.57	0.70	0.47	0.62	0.67	0.51
C11	0.62	1.20	0.66	0.71	0.98	0.74	0.66	0.77	0.72	0.72	0.74	0.79
C12	0.60	0.91	0.59	0.66	0.84	0.67	0.63	0.73	0.62	0.68	0.70	0.63

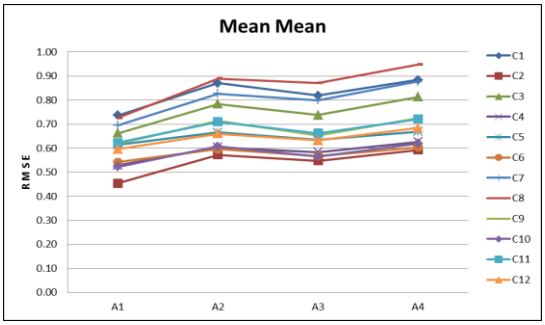
Table 5. BIAS for the Second Item Discrimination Parameter

	15%			**30%**			**60%**			**100%**		
	MM	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**
C1	0.48	0.52	0.50	0.57	0.36	0.51	0.53	0.67	0.58	0.58	0.65	0.55
C2	0.31	0.36	0.32	0.34	0.27	0.31	0.32	0.58	0.34	0.34	0.53	0.98
C3	0.51	0.90	0.54	0.57	0.58	0.54	0.54	0.55	0.54	0.60	0.53	0.54
C4	0.39	0.63	0.38	0.43	0.44	0.38	0.41	0.61	0.39	0.44	0.55	0.36
C5	0.49	1.10	0.50	0.53	0.81	0.47	0.50	0.57	0.45	0.54	0.54	0.41
C6	0.45	0.85	0.43	0.49	0.68	0.43	0.47	0.62	0.40	0.51	0.57	0.37
C7	0.47	0.53	0.47	0.54	0.40	0.47	0.51	0.65	0.46	0.57	0.64	0.49
C8	0.51	0.46	0.54	0.60	0.43	0.51	0.58	0.71	0.51	0.64	0.69	0.56
C9	0.48	0.80	0.46	0.54	0.58	0.47	0.50	0.60	0.44	0.56	0.60	0.42
C10	0.40	0.56	0.37	0.46	0.52	0.38	0.43	0.65	0.37	0.47	0.61	0.33
C11	0.52	1.01	0.52	0.58	0.78	0.52	0.54	0.57	0.47	0.60	0.59	0.48
C12	0.52	0.73	0.48	0.55	0.69	0.49	0.53	0.65	0.45	0.58	0.62	0.41

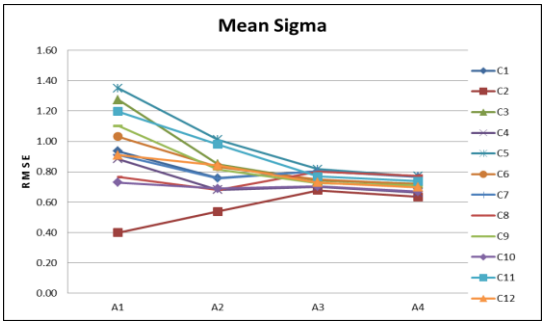
Based on the RMSE values in Table 4 and the BIAS values in Table 5 for the second discrimination parameter,it can be said that mean/mean, mean/sigma, and Stocking-Lord procedures generally produced close RMSE and BIAS values under manipulated simulation conditions. When the sample size is smaller and the correlation between two dimension is higher, mean/sigma procedure had higher RMSE and BIAS values than other two procedures.

As seen in Figure 3 and Figure 4, the RMSE and the BIAS values tend to decrease as the percentage of anchor items increases. However, the larger sample size and no correlation between dimensions condition produced the smallest RMSE and BIAS values for mean/sigma procedure. The RMSE values for mean/mean and Stocking-Lord procedures tend to increase as the percentage of anchor items increases. The BIAS values for mean/mean and Stocking-Lord procedures are more stable across the levels of percentage of anchor items. As was the case for the first discrimination parameter, the larger sample size conditions produced smaller RMSE and BIAS values for the second discrimination parameter for three methods.

(a)



(b)



(c)

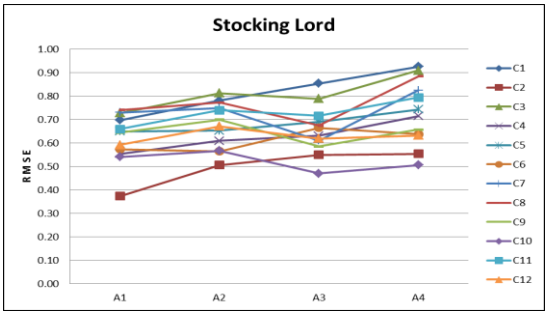
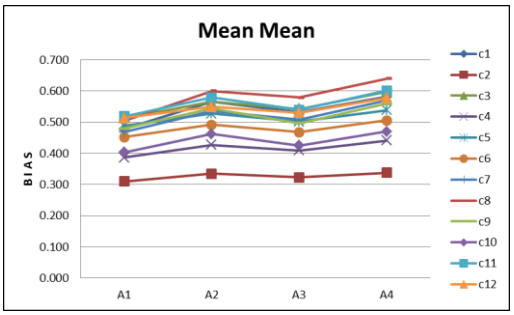
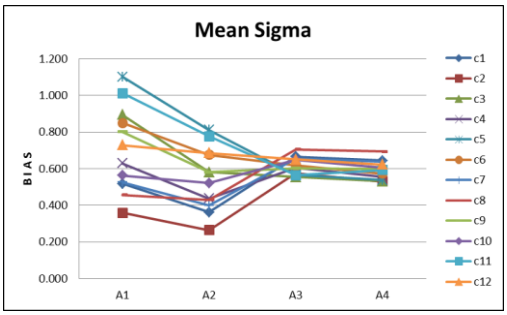


Figure 3. RMSE Values of the Second Item Discrimination Parameter for Three Equating Procedures by the Percentage of Anchor Items Under All Conditions

(a)



(b)



(c)

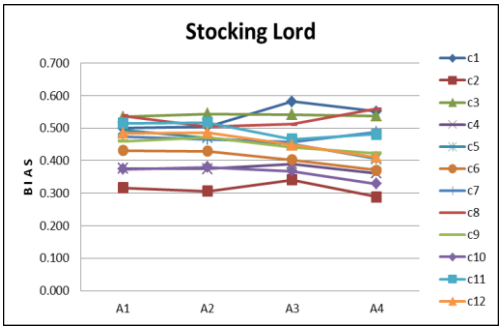


Figure 4. BIAS Values of the Second Item Discrimination Parameter for Three Equating Procedures by the Percentage of Anchor Items Under All Conditions

RMSE and BIAS values of the item difficulty parameter estimates calculated for the three equating procedures across all conditions are given in Table 6 and Table 7, respectively. RMSE and BIAS values of the item difficulty parameter estimates for the levels of the percentage of anchor items are shown in Figure 5 and Figure 6, respectively for each equating procedure.

Table 6. RMSE for the Item Difficulty Parameter

	15%			**30%**			**60%**			**100%**		
	MM	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**
C1	0.71	0.61	0.63	1.02	0.78	0.63	0.92	0.68	0.81	0.87	0.68	0.79
C2	0.45	0.40	0.37	1.47	0.70	0.53	0.96	0.61	0.63	0.89	0.63	0.59
C3	0.77	0.58	0.62	0.89	0.83	0.61	0.89	0.70	0.70	0.71	0.71	0.73
C4	0.69	0.51	0.56	0.79	0.71	0.55	0.81	0.64	0.62	0.64	0.65	0.64
C5	0.87	0.58	0.56	1.03	0.88	0.55	1.04	0.74	0.57	0.75	0.76	0.60
C6	0.73	0.51	0.49	0.89	0.74	0.49	1.45	0.64	0.52	0.85	0.66	0.50
C7	0.72	0.64	0.58	1.06	0.77	0.58	1.15	0.68	0.61	0.92	0.69	0.66
C8	0.68	0.61	0.58	0.86	0.66	0.57	1.23	0.61	0.62	0.99	0.62	0.69
C9	0.82	0.65	0.62	0.87	0.83	0.61	2.07	0.79	0.63	1.34	0.77	0.63
C10	0.67	0.59	0.55	0.72	0.66	0.55	1.04	0.66	0.57	0.80	0.66	0.57
C11	0.86	0.65	0.62	0.91	0.82	0.63	1.01	0.79	0.64	0.87	0.78	0.66
C12	0.73	0.58	0.56	0.78	0.68	0.56	0.76	0.70	0.57	0.65	0.69	0.57

Table 7. BIAS for the Item Difficulty Parameter

	15%			30%			60%			100%		
	MM	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**	**MM**	**MS**	**SL**
C1	0.57	0.50	0.54	0.53	0.65	0.53	0.55	0.60	0.59	0.53	0.61	0.58
C2	0.54	0.48	0.50	0.50	0.62	0.45	0.59	0.56	0.47	0.49	0.58	0.46
C3	0.71	0.49	0.52	0.79	0.71	0.50	0.49	0.63	0.52	0.51	0.64	0.54
C4	0.63	0.46	0.49	0.73	0.63	0.46	0.46	0.6	0.46	0.45	0.61	0.47
C5	0.81	0.49	0.47	0.98	0.79	0.46	0.47	0.68	0.46	0.52	0.71	0.46
C6	0.69	0.44	0.44	0.86	0.67	0.43	0.58	0.59	0.42	0.43	0.63	0.42
C7	0.58	0.55	0.47	0.52	0.63	0.48	0.67	0.61	0.50	0.59	0.62	0.49
C8	0.54	0.54	0.46	0.55	0.55	0.47	0.85	0.56	0.47	0.72	0.56	0.47
C9	0.73	0.57	0.52	0.77	0.71	0.51	0.63	0.72	0.54	0.58	0.71	0.54
C10	0.61	0.54	0.46	0.65	0.57	0.47	0.66	0.60	0.51	0.58	0.61	0.51
C11	0.80	0.56	0.53	0.85	0.72	0.52	0.57	0.73	0.55	0.57	0.73	0.55
C12	0.67	0.51	0.49	0.73	0.60	0.48	0.52	0.66	0.50	0.52	0.65	0.50

Based on Table 6 and Table 7, RMSE values for mean/mean procedures are higher than the other two equating procedures across all conditions. Moreover, in terms of BIAS values mean/sigma procedure has higher values than the other two equating procedures under the simulation conditions. RMSE and BIAS values of item difficulty parameter estimates incease as sample size decreases. On the other hand, mean difference and correlation between two dimensions does not have clear effect on the RMSE and BIAS values of item difficulty parameter estimates under the manipulated conditions.

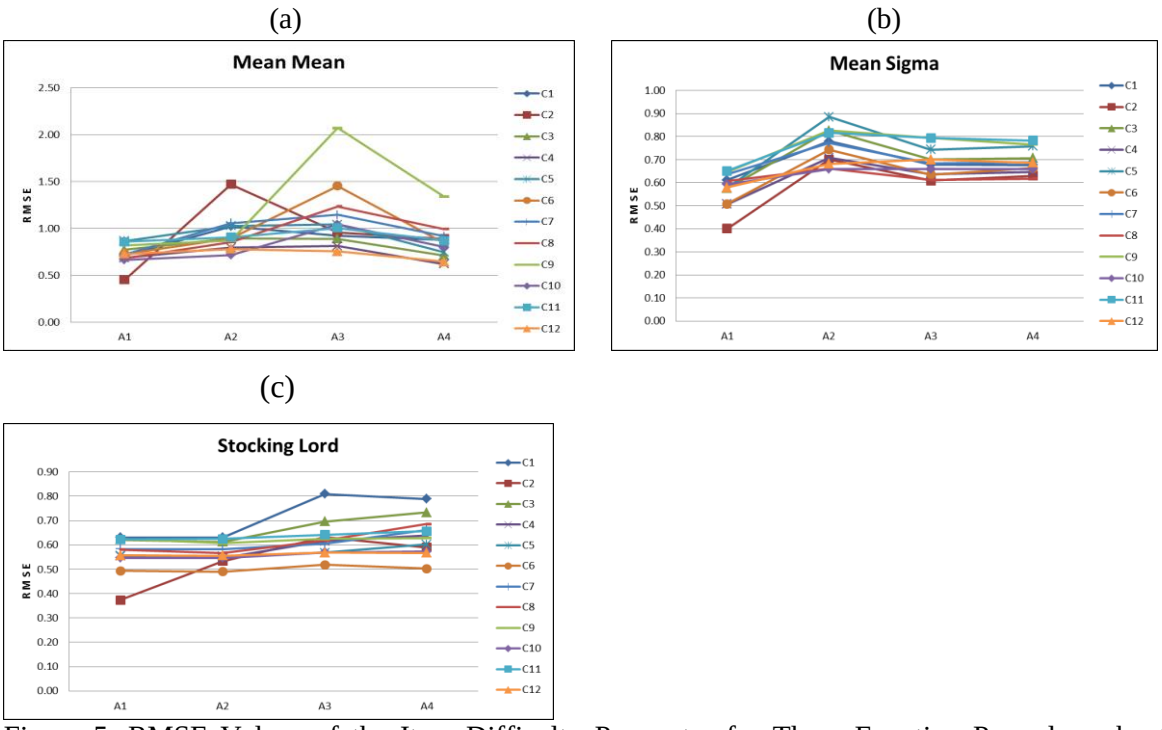
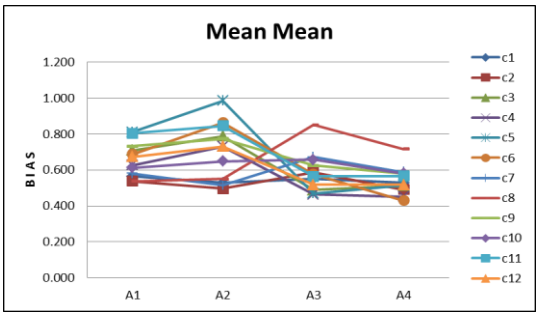
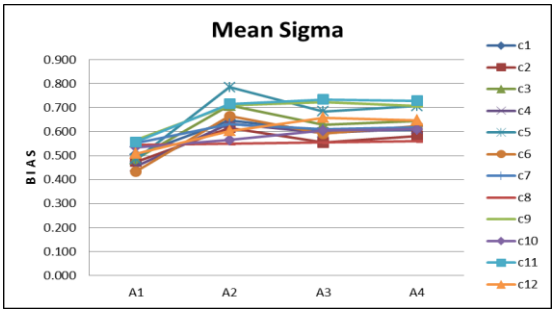


Figure 5. RMSE Values of the Item Difficulty Parameter for Three Equating Procedures by the Percentage of Anchor Items Under All Conditions

(a)



(b)



(c)

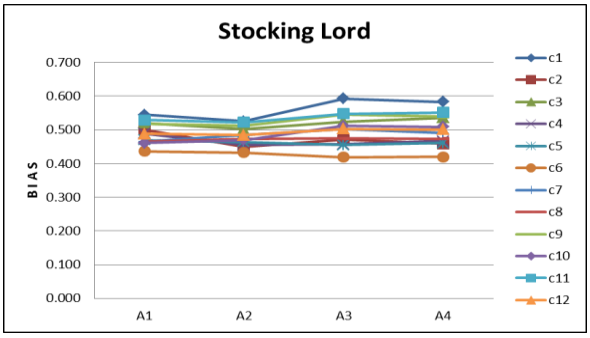


Figure 6. BIAS Values of the Item Difficulty Parameter for Three Equating Procedures by the Percentage of Anchor Items Under All Conditions

CONCLUSION and DISCUSSION

When RMSE and BIAS values of item parameter estimates were examined, it was found that the ability distribution factor did not have considerable effect on the accuracy of item discrimination and item difficulty parameter estimates for the three equating procedures. Those findings are similar to the findings of Yao & Boughton (2009). They found in their study that with a well-chosen anchor set including at least one simple structured item per dimension, parameter estimates are more accurate. On the other hand, sample size factor affected RMSE and BIAS values of item parameter estimates for the three procedures with larger sample size conditions produced smaller RMSE and BIAS values. Those findings are consistent with the findings of Eser and Gelbal (2015). They suggested in their study to run the analyses with a sample size of 2000 examinees for a two-dimensional test. Percentage of anchor items factor affected the results of mean/sigma procedure the most. RMSE and BIAS values of item discrimination parameter estimates tend to decrease as the percentage of anchor items increases for that procedure. RMSE and BIAS values of item discrimination parameter are similar for the levels of the percentage of anchor items factor for mean/mean and Stocking-Lord procedures.

It can be concluded that Stocking-Lord and mean/mean procedures provided better estimates for the item discrimination parameters while Stocking-Lord and mean/sigma procedures provided better estimates for item difficulty parameter.

When the test forms are in multidimensional structure, it is critical to use appropriate methods in the estimation of item and person parameters and equating/linlking test forms. In that sense, the performance of different methods should be investigated in depth under different conditions that are similar to real testing situations.

In this study, only three of the multidimensional equating procedures were compared for item parameter recovery. In future researches, accuracy of parameter estimates can be compared among

unidimensional and multidimensional equating procedures. Conditions may vary in the test length and the number of dimensions considering the real test settings.

REFERENCES

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67-91.

Angoff, W. H. (1971). Scales, norms, and equaivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp.508-600). Washington, DC: American Council on Education. (Reprinted as W. H. Angoff, *Scales, norm, and equivalent scores*. Princeton, NJ: Educational Testing Service, 1984).

Bolt, D. M. (1999). Evaluating the effects of multidimensionality on IRT true score equating. *Applied Measurement in Education, 12*(4), 383-407.

Camilli, G., Wang, M., & Fesq, J. (1995). The effects of dimensionality on equating the law school admission test. *Journal of Educational Measurement*, *32*(1), 79-96.

Davey, T., Oshima, T. C., & Lee, K. (1996). Linking multidimensional item calibrations. *Applied Psychological Measurement*, *11*(3), 221-224.

De Champlain, A. F. (1996). The effect of multidimensionality on IRT true-score equating for subgroups of examinees. *Journal of Educational Measurement*, *33*(2), 181-201.

Eser, D. Ç. & Gelbal, S. (2015). Examining parameter estimations of simple and complex structured tests with various dimensionality properties based on multidimensional item response theory. *Journal of Measurement and Evaluation in Education and Psychology*, *6*(2), 331-350.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hirsch, T. M. (1989). Multidimensional equating. *Journal of Educational Measurement*, *26*(4), 337-349.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Levine, R. (1955). Equating the score scales of alternate forms administered to samples of different ability (Research Bulletin, 55-23). Princeton, NJ: Educational Testing Service.

Li, Y. H., & Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement*, *24*(2), 115-138.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education*, *3*(1), 73-95.

Lord, F. M. (1980). Applications of Item Response Theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56*(4), 495-529.

Yao, L. (2003). *BMIRT: Bayesian multivariate item response theory* [Computer software and manual]. Monteray, CA: CTB/McGraw Hill.

Yao, L. (2003). *SimuMIRT* [Computer software]. Monteray, CA: DMDC DoD Center.

Yao, L. (2004). *LinkMIRT: Linking of multivariate item response model* [Computer software]. Monteray, CA: DMDC DoD Center.

Yao, L., & Boughton, K. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psycholoical Measurement*, *31*, 1-23.

Yao, L., & Boughton, K. (2009). Multidimensional linking for tests with mixed item types. *Journal of Educational Measurement*, *46*(2), 177-197.

Yao, L., & Schwarz, R. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed format tests. *Applied Psychological Measurement*, *30*, 469-492.

Yao, L. (2011). Multidimensional linking for domain scores and overall scores for nonequivalent groups. *Applied Psychological Measurement, 35*(1), 48-66.

Yue, L., & Hongyun, L. (2013). Comparison of MIRT linking methods for different common item designs. *Acta Psychologica Sinica*, *45*(4), 466-480.

GENİŞ ÖZET

Giriş

Testlerin standart koşullar altında uygulanmasının başlıca nedeni, adayların yeteneklerini adil ve objektif olarak değerlendirmektir. Adaylar ile ilgili değerlendirmeler önemli kararlarda kullanılır. Güvenlik açısından her bir uygulamada yeni bir test formu kullanılır. Formlar aynı yapıyı ölçmek

için geliştirilmiş olsa da, madde güçlükleri ve güvenilirlikleri gibi istatistiksel özelliklerinde farklılıklar gösterebilir. Bunu önlemek için yeni testin puanları önceki uygulamada elde edilen puanlara dönüştürülmelidir. Farklı formlarından elde edilen puanlar eşitleme denilen istatistiksel bir işlemin ardından karşılaştırılabilir (Kolen & Brennan, 2004).

Farklı veri toplama desenleri için farklı test eşitleme yöntemleri vardır. Eşdeğer olmayan gruplarda ortak madde deseni, geniş ölçekli standartlaştırıolmış testlerin farklı formlarının eşitlenmesinde/bağlanmasında en yaygın kullanılan veri toplama desenlerinden birisidir.

Maddeler tepki kuramına MTKdayanan yöntemler, eşdeğer olmayan gruplarda ortak madde deseni kullanılabilir. Madde tepki kuramı eşitlemede belli özelliklere sahiptir. Madde tepki kuramını avantajlarından birisi, model-veri uyumu sağlandığında madde ve yetenek parametrelerinin değişmezliliğidir (Lord, 1980). Madde tepki kuramının değişmezlik özelliği, özellikle eşdeğer olmayan gruplarda ortak madde deseninde test eşitlemede önemli bir role sahiptir (Skaggs & Lissitz, 1986). Madde tepki kuraminin değişmezlik özelliği, varsayımların karşılanmasına bağlıdır. Madde tepki kuramına dayanan eşitleme çalışmalarında dikkate alınması gereken varsayımlardan biri, tek boyutluluk varsayımıdır. Halbuki gerçek uygulamalardaki testler çok boyutlu bir yapı sergilemektedir ve tek boyutluluk varsayımı birçok test durumunda ihlal edilmektedir (Li & Lissitz, 2000). Tek boyutluluk varsayımı karşılanmadığında, çok boyutlu madde tepki kuramına (MIRT) dayanan yöntemler uygulanabilir. Bir testin formları çok boyutlu bir yapı sergilediğinde, MIRT eşitleme düşünülebilir. Eşitleme işlemi sonrasında parametre kestirimlerinin doğruluğu, MIRT eşitleme önemlidir (Li & Lissitz, 2000). MIRT temelli işlemlerin performansı, gerçek test koşullarına benzer farklı koşullar altında araştırılmalıdır (Yao ve Boughton, 2009).

Çok boyutlu eşitleme/bağlama ile ilgili birçok yayın bulunmaktadır (Hirsch, 1989; De Champlain, 1996; Bolt 1999; Li & Lissitz, 2000; Yao ve Boughton, 2009; Yao, 2011). Bu araştırmalardan bazılarında gerçek test verileri ve bazilarinda simüle edilen veriler kullanılmış. Bolt (1999) LSAT verilerinin iki formundan elde ettiği parametre kestirimlerini kullanılarak iki boyutlu verileri simüle ettmiştir. Tek boyutlu madde tepki kuramı gerçek puan eşitlemenin performansını, boyutlar arasındaki farklı korelasyon seviyeleri altında tek boyutlu doğrusal ve eşit yüzdelikli eşitleme yöntemlerinin performansıyla karşılaştırmıştır. Bu çalışma sonucunda, boyutlar arasındaki korelasyonun daha yüksek olduğu durumlarda, MTK gerçek puan eşitleme yönteminin performansinin diğer 2 geleneksel yöntemin performansi kadar iyi olduğunu saptamıştır. Ayrıca, MTK gerçek puan eşitleme yönteminin boyutlar arasındaki korelasyonun daha düşük olduğu durumlarda diğerlerinden daha iyi performans gösterdiğini bulmustur. Yao ve Boughton (2009), test yanıt fonksiyonu yönteminin bağlantılı doğruluğunu, hem iki kategorili puanlanan hem de çok kategorilii puanlanan maddeleri içeren testler için çok boyutlu perspektifle incelemistir. Çalışmalarında, popülasyon dağılımı, ortak madde set uzunluğu ve madde yapısı farklı koşullar altında simüle edilmiş iki boyutlu verileri kullanilmisitr. Parametre iyileştirmesinin, iyi seçilmiş bir ortak madde kümesiyle tüm koşullarda iyi olduğunu bulmuşlardır. Yao (2011), beş boyutlu simüle edilmiş verileri kullanarak iki ayrı puanlamalı maddeler için bağlama yöntemini birbirine bağlayan çok boyutlu test yanıt fonksiyonunun bağlantı doğruluğu üzerinde araştırmalar yapmıştır. Bu çalışmada örneklem büyüklüğü, populasyon dağılımı ve ortak madde seti uzunluğu değişik koşullarda simule edilmistir. En küçük ortak madde seti olan koşullar için bile genel puanın ve alan puanı iyileşmesinin iyi olduğu bulunmustur. Pek çok test verisinin çok boyutlu yapısını göz önünde bulundurarak, farklı çok boyutlu eşitleme/bağlama yöntemlerinin performansını derinlemesine araştırmak ve hangi yöntemin farklı koşullar altında daha iyi performans gösterdiğini gözlemlemek çok önemlidir.

Bu çalışmanın amacı çeşitli simülasyon koşulları altında madde parametre kestirimlerinin kararlılığında üç çok boyutlu eşitleme yönteminin performansını araştırmaktır. Farkli orneklem büyüklüğü, yetenek dağılımı, boyutlar arasındaki korelasyon ve testteki ortak maddelerin yüzdesi bir araya getirilerek simülasyon koşulları oluşturulmustur. Bu çalışmada, eşdeğer olmayan gruplarda ortak madde deseni altında çok boyutlu Stocking-Lord, ortalama/ortalama ve ortalama/sigma karşılaştırılmıştır.

Yöntem

Bu çalışmada, eşdeğer olmayan gruplarda ortak madde deseni altında üretilen veriler kullanılmıştır. Çalışmanın amacı doğrultusunda, Yao ve Boughton'un (2009) çalışmasında yer alan madde parametre kestirimlerine dayanarak, iki kategorili puanlanan 40 maddeli iki boyutlu bir teste verilen yanıtlar oluşturulmuştur.

SimuMIRT programı (Yao, 2003) çeşitli koşullar altında yanıt verisi üretmek için kullanılmıştır. Li & Lissitz (2000) çalışmalarında, 20 ortak madde içeren 40 maddelik bir testten için 2000 örneklem büyüklüğünün, çok boyutlu test yanıt fonksiyonu eşitleme yöntemi için yeterli olduğunu bulmuştur. Bu çalışmada, örneklem büyüklüğü (1000 ve 2000), yetenek dağılımı ((0,0), (-0.5,0.5)), ortak madde yuzdesi (% 15,% 30,% 60 ve% 100) ve boyutlar arasindaki korelasyon (0,0.5, 0.8) manipüle edilen faktörlerdir. Her simülasyon koşulu 20 kere tekrarlanmıştır. Test uzunluğu ve boyut sayısı, sabit tutulan faktörlerdir. Parametreleri kestirmek etmek için BMIRT programı (Yao, 2003) kullanılmıştır. Eşitleme, çok boyutlu Stocking-Lord, ortalama/ortalama ve ortalama/sigma eşitleme yöntemleri kullanılarak gerçekleştirilmistir. LinkMIRT programı (Yao, 2004) tüm simülasyon koşullarında eşitleme için kullanılmıştır.

Maddelerin parametre kestirimlerimi değerlendirmek için, RMSE ve yanlılık (BIAS) değerleri hesaplanmıştır.

Sonuç ve Tartışma

Madde parametre kestirimlerinin RMSE ve BIAS değerleri incelendiğinde, yetenek dağılımı faktörünün, üç eşitleme yöntemi için madde ayırt edicilik ve madde güçlük parametreleri kestirimlerine önemli bir etkisi olmadığı bulunmuştur. Diğer yandan, örneklem büyüklüğü faktörü daha büyük örneklem büyüklüğü koşulları altında üç yöntem için de madde parametre kestirimlerinin RMSE ve BIAS değerlerini etkilerken, daha küçük RMSE ve BIAS değerleri üretmiştir. Ortak madde faktörünün yüzdesi en fazla ortalama/sigma yönteminin sonuçlarını etkilemiştir. Bu yöntem için madde ayırt edicilik parametre kestirimlerinin RMSE ve BIAS değerleri, ortak maddelerinin yüzdesi arttıkça azalma eğilimi göstermiştir. Madde ayırt edicilik parametre kestirimlerinin RMSE ve BIAS değerleri, ortalama/ortalama ve Stocking-Lord yöntemleri için ortak madde faktörünün seviyeleri bakımından benzerdir.

Stocking-Lord ve ortalama/ortalama yöntemlerinin madde ayırt edicilik parametresi için daha iyi kestirimler sağladığı, Stocking-Lord ve ortalama/sigma yöntemlerinin madde güçlüğü parametresi için daha iyi kestirimler sağladığı sonucuna varılabilir.

Bu çalışmada çok boyutlu eşitleme yöntemlerinden sadece üçü madde parametrekestirimlerinin kararlılığı bakımından karşılaştırılmıştır. Gelecekteki araştırmalarda parametre kestirimlerinin kararlılığı tek boyutlu ve çok boyutlu eşitleme yöntemleri arasında karşılaştırılabilir. Koşullar gerçek test değerleri göz önünde bulundurularak test uzunluğu ve boyut sayısı çeşitlendirilerek oluşturulabilir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(4);435-452

PISA 2012 Problem Çözme Yeterliğine Etki Eden Okul Değişkenlerinin İncelenmesi: Türkiye-Sırbistan Karşılaştırması

PISA 2012 Analysis of School Variables Affecting Problem-Solving Competency: Turkey-Serbia Comparison

Emine YAVUZ * Bayram ÇETİN **

Öz
OECD'nin PISA 2012 Türkiye problem çözme raporuna göre Türkiye ve Sırbistan aynı matematik okuryazarlık düzeyindedir. Fakat Sırbistan'ın ortalama problem çözme yeterliğinin, Türkiye'den daha yüksek olduğu ifade edilmiştir. Bu doğrultuda bu çalışmada iki ülkenin problem çözme okuryazarlığına etki eden okul değişkenleri belirlenip karşılaştırılmıştır. Nedensel karşılaştırma yöntemi ile yürütülen bu çalışmada Türkiye örnekleminde 147 okuldan 4494 öğrenciye, Sırbistan örnekleminde ise 132 okuldan 4059 öğrenciye ait veri üzerinde ayrı ayrı HLM analizi yapılmıştır. HLM analizi sonucunda, Sırbistan için "engel ve aile bağışı" değişken etkileri, Türkiye için ise "terk, öğretmen morali ve matematik yarışı" değişken etkileri istatistiksel olarak anlamlı bulunmuştur. İki ülkede farklı değişkenlerin problem çözme okuryazarlığı üzerinde manidar etkileri olduğu görülse de bu değişkenlerin okul iklimi kavramının birer bileşeni olması oldukça dikkate değerdir.

Anahtar Kelimeler: PISA 2012, problem çözme yeterliği, Türkiye, Sırbistan.

Abstract
According to the OECD's PISA 2012 Turkey problem-solving report, Turkey and Serbia are at the same mathematical literacy level. However, Serbia's average of problem-solving competency is said to be higher than Turkey's. In this study, school variables that affect problem-solving competency of the two countries were examined and compared. The method of the study was causal comparison method, and HLM analysis was performed on data of 4494 students from 147 schools in Turkey sample and 4059 students from 132 schools in Serbia sample separately. As a result of HLM analysis, "obstacle and family donation" variable for Serbia and "abandon, teacher morale and mathematics competition" variable for Turkey were statistically significant. Although it was found that for each countries different variables influence the problem-solving competency, it was quite remarkable that these variables are in common in that they are components of the school climate concept.

Keywords: PISA 2012, problem solving competency, Turkey, Serbia.

GİRİŞ

Çok fazla değişkenin etkisinde sürekli değişen ve gelişen dünyada, toplumlar arası etkileşim kaçınılmaz hale gelmiştir. Bu durum, sınırların ortadan kalkmasına ve hiçbir toplumun ve ulusun, dış dünyadan bağımsız olarak kendi içerisinde kapalı kalamamasına neden olmuştur. Bu karmaşık ve rekabetin en üst düzeye çıktığı ortamlarda sağlanacak başarı, öncelikle nitelikli insan gücüne bağlıdır. Bu bağlamda eğitim, küresel dönemin dinamikleri doğrultusunda çok boyutlu ve çok yönlü nitelikli insan modelinin yetiştirilmesinde, en etkin ve önemli araçlardan biridir (Demir, 2010:1-5).

* Araştırma Görevlisi, Erciyes Üniversitesi, Eğitim Fakültesi, Kayseri-Türkiye, yavuzemine0@gmail.com, ORCID ID: https://orcid.org/0000-0002-1991-1416
** Doçent Doktor, Milli Eğitim Bakanlığı, Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü, Ankara-Türkiye, bcetin27@gmail.com, ORCID ID: https://orcid.org/0000-0001-5321-8028

DOI: 10.21031/epod.334610
Geliş Tarihi: 14.08.2017
Kabul Tarihi: 17.11.2017

Ekonomik İşbirliği ve Kalkınma Örgütü (Organisation for Economic Cooperation and Development - OECD) ülkelerin ihtiyaç duyduğu nitelikli insan gücü yetiştirme kapasitesini ortaya koymak ve bu kapasiteyi etkileyen değişkenleri belirlemek amacıyla, yetişmekte olan öğrencilerin başarılarının dikkate alındığı çalışmalar ve sınavlar yapmaktadır. Yapılan çalışmalar, projeye katılan ülkelerin genç nüfuslarını yetiştirmede karşılaştıkları sorunlara ışık tutmakta ve küresel bazda problemleri olup olmadığına ilişkin bilgi vermektedir. Problemlerin varsa çözümünü araştırmak için başka araştırmalara ihtiyaç olduğunu ortaya koymaktadır. Var olan eğitim sistemlerinin mevcut durumunu tespit etmek, öğrencilerin bilgi toplumunun ihtiyaçlarına uygun yetişip yetişmediklerini anlamak, var olan eğitim sistemini geliştirmek ve diğer ülkelerin eğitim sistemleriyle karşılaştırma yapmak isteyen ülkeler de bu çalışmalara ve sınavlara katılmaktadırlar (Eğitimde Araştırma ve Geliştirme Daire Başkanlığı: EARGED, 2005). Eğitim sistemlerinin kalitesinin değerlendirilmesinde ve karşılaştırılabilirliklerini sağlayan bu uluslararası sınav, Uluslararası Öğrenci Başarısını Belirleme Programı'dır (The Program for International Student Assessment-PISA). PISA ilk defa 2000 yılında uygulanan, ülkelerdeki 15 yaşındaki öğrencilerin eğitimleri süresince, günlük hayatta karşılaşabilecekleri durumlar konusunda ne derece hazırlıklı yetiştirildiklerini belirlemek amacıyla geliştirilmiş bir programdır.

PISA ve Uluslararası Matematik ve Fen Eğilimi Çalışması (Trends in International Mathematics and Science Study - TIMSS) gibi karşılaştırmalı çalışmalar, özellikle vizyon geliştirme ve eğitimin planlanması süreçlerinde eğitimcilere ve politika üretenlere büyük katkı sağlamaktadır (Aydın, Sarıer ve Uysal, 2012: 22). Ayrıca öğrenci, okul ve eğitim sistemlerinin ortak bazı özelliklerini açıklamasıyla PISA; eğitimde kaliteyi, eşitliği ve verimliliği artırmak için kullanılabilecek yararlı bir araç olup öğrenci başarısında iyi bir kestiricidir (Schleicher, 2007). Bu tür çalışmalardan elde edilen verilerin analiz edilmesiyle ülkeler, mevcut eğitim sistemlerinin güçlü ve zayıf yönlerini, eğitim politikalarını, öğretim programlarını ve öğretmenlerin yeterliklerini gözden geçirilebilmektedir (Çelen, Çelik ve Seferoğlu, 2011:764-770).

PISA ile öğrencilerin sahip oldukları bilgi ve becerileri kullanabilme yeteneği, analiz edebilme, akıl yürütme ve okulda öğrenilen fen ve matematik kavramlarını kullanarak etkin bir iletişim kurma becerisine sahip olup olmamaları ölçülmeye çalışılmaktadır (OECD, 2009). Bu doğrultuda PISA'da öğrenciler, üç ana alan (okuma becerileri, fen ve matematik okuryazarlığı) ve yan alanlarda (problem çözme, vb.) değerlendirilmektedirler. Üç yılda bir yapılan bu araştırmaların her uygulamasında her konu alanıyla ilgili alt test bulunmakla birlikte, ağırlık verilen, daha detaylı incelenen alan değişmektedir. 2000 yılında başlayan bu uygulamalarda matematik ve problem çözme alanları ilk olarak 2003 yılında ikinci olarak 2012 yılında detaylı bir şekilde incelenmiştir (OECD, 2004b; 2013a).

PISA 2012'de problem çözme yeterliği, çözümü belli olmayan problem durumlarını öğrencinin önceden kazandığı bilgi, beceri, yetenek ve psikolojik kaynaklar yardımıyla çözmesi olarak ifade edilmiştir (OECD, 2013a). Tanımdan anlaşılacağı üzere bu yeterlik sadece özel bir alana ait olan bilgileri öğrenmeyi değil, öğrencinin sahip olduğu ilgili bilgi ve becerilerini kullanabilmesini içermektedir. Nitekim sürekli değişen ve gelişen dünyada bireylerin başarılı olabilmeleri için değişime ayak uydurmaları, bu esnada önceden edinilmiş bilgi ve becerilerini kullanabilmeleri gerekmektedir. Bu bağlamda, bireylerin sürekli değişim içerisinde olan hayata aktif katılımlarında problem çözme yeterliği oldukça önemlidir (Yavuz ve Atar, 2016). Bu gibi nedenlerle problem çözme yeterliliğinin öğrencilere kazandırılması, birçok ülkenin eğitim programlarının temel amaçlarından biridir (Lesh ve Zawojewski, 2007).

OECD'nin PISA 2012 Türkiye problem çözme alanına ait rapora (OECD, 2013b) göre Türkiye ve Sırbistan aynı matematik okuryazarlık düzeyindedir. Fakat raporda Türkiye ve Sırbistan'ın aynı matematik okuryazarlık düzeyinde olmalarına rağmen Sırbistan'ın ortalama problem çözme yeterliğinin, Türkiye'den daha yüksek olduğu ifade edilmiştir. Türkiye ve Sırbistan PISA'ya ilk defa 2003 yılında katılmışlardır. Sırbistan eğitim sistemi incelendiğinde Türkiye ile pek çok benzerlik gösterdiği söylenebilir. Bu benzerlikleri açıklamak için Sırbistan eğitim sisteminden kısaca bahsetmek gerekir. Sırbistan'da eğitim 4 düzeyden meydana gelir: Okul öncesi, ilkokul, ortaöğretim ve yükseköğretim. Okul öncesi eğitime 7 yaşına kadar her öğrenci gidebilmektedir. İlkokul 7-14 yaş aralığını ve 8 sınıf düzeyini kapsar. Ayrıca her öğrenci için zorunludur. İlkokul iki kademeden oluşmaktadır. Birinci kademe 1-4. sınıfları, ikinci kademe ise 5-8. sınıfları içermektedir. Birinci

kademeyi bitiren öğrenciler, ikinci kademeyi genellikle aynı okulda devam etmektedirler. Beşinci sınıfa başlayan öğrenciler tarih, geometri, biyoloji, fizik ve kimya gibi özel alan dersleri almaya başlarlar. Öğrenciler dersleri bir önceki sınıf düzeyinde olduğu gibi grup şeklinde alırlar fakat bazı derslere alan öğretmenleri girmektedir. Sekizinci sınıf sonunda öğrenciler okul ortalamaları ile birlikte ilkokul diploması alarak mezun olurlar (Baucal, Pavlovic-Babic ve Willms, 2007: 539). Sırbistan'da ortaöğretim 15-18 yaş aralığını kapsamaktadır. Ortaöğretimde üç tür lise bulunmaktadır: Genel lise, mesleki lise ve sanat okulları. Genel liseler sosyal ve fen bilimlerin okutulduğu dört yıllık liselerdir. Bu liselere giden öğrenciler üniversite için hazırlanırlar. Meslek liselerine giden öğrenciler üniversiteye hazırlanabilecekleri gibi genellikle 15 farklı alanda iş hayatına atılmak için hazırlanırlar. Bu liseler ilgili alana göre 1, 2, 3 veya 4 yıl sürebilir. Dört yıllık meslek lisesinden ve 2 yıl çalışmadan sonra bu öğrenciler bir yıl daha uzmanlık eğitimi alabilmektedirler. Sanat eğitimi okulları müzik, bale ve güzel sanatlar olarak dallara ayrılmıştır (Baucal ve Pavlovic-Babic, 2009; Teodorovic, 2005: 12-16). Sırbistan'da öğrenciler meslek liselerine ilköğretim ikinci kademe ortalaması ile girebilirken genel liselere veya sanat okullarına yerleşebilmek için bir "giriş sınavına" girmeleri gerekmektedir. Bu sınav matematik ve dil testinden oluşmaktadır. Bu sınavın %40'ı ve öğrencilerin ilkokul ikinci kademe ortalamalarının %60'ı ile elde edilen ağırlıklı ortalama ile öğrenciler genel liselere veya sanat okullarına yerleşmektedir (Baucal ve diğerleri 2007: 540). Öğrencilerin ilgili sınıf düzeylerine devam etme yaşları dışında Türkiye ve Sırbistan eğitim sisteminin oldukça benzer olduğu görülmektedir. Sırbistan'da 15 yaş grubu öğrencileri lise düzeyinde öğrenimlerine devam ederken Türkiye'de 15 yaş grubu öğrencilerin bir kısmı ilköğretim son sınıf düzeyindedir (MEB, 2015).

PISA'da toplanan veri yapısı ve veri türüne bağlı olarak yapılan araştırmalar ile ülkeler belirlenen alanlarda birbirleriyle karşılaştırılabilmektedirler. İlgili literatür tarandığında Türkiye ve Sırbistan'ın herhangi bir yıla ait PISA problem çözme yeterliğini karşılaştıran bir çalışma bulunamazken, iki ülkenin PISA 2009 matematik başarısını karşılaştıran bir çalışma bulunmuştur. Kılıç, Çene ve Demir'in (2012) yaptığı bu çalışmada sekiz ülkedeki (Türkiye, Bulgaristan, Yunanistan, Azerbaycan, Rusya, İsrail, Sırbistan, Romanya ve Ürdün) öğrencilerin matematik başarısına etki eden ülke, okul ve öğrenci düzeyi değişkenleri ile öğrencilerin kullandıkları öğrenme stratejileri karşılaştırılmıştır. Yapılan üç düzeyli hiyerarşik lineer model analizi sonucunda cinsiyet, sosyo-ekonomik durum, ev eğitim kaynağı, kültürel aitlik, ezberleme, eleme ve kontrol stratejileri Sırbistan ile Türkiye'de istatistiksel olarak anlamlı etkiye sahip olduğu görülmüştür. Ayrıca öğrenci/öğretmen oranı değişkeni Türkiye'de istatistiksel olarak anlamlı etkiye sahiptir. Türkiye ve Sırbistan'ın eğitim sistemlerinin ve PISA matematik başarılarının paralellik göstermesi, öğrencilerin matematik başarılarına etki eden değişkenlerin benzerlik göstermesi beklentisini oluşturmaktadır. Nitekim Kılıç ve diğerlerinin (2012) yaptığı bu çalışma bu beklentiyi destekler niteliktedir.

Problem çözme yeterliliğinin öğrencilere kazındırılması konusu eğitim programlarına dahil edilmesi ile bu yeterliliğin kazandırılmasında okulların önemli işlevlerinin olması beklenir. Literatürde ülkeler arasında öğrenci başarısına etki eden okul değişkenlerinin karşılaştırıldığı çalışmalar mevcuttur. Örneğin İş (2003), farklı kültürlerde 15 yaşındaki öğrencilerin PISA 2000 matematik okuryazarlıklarına etki eden öğrenci, aile ve okul faktörlerini yapısal eşitlik modeli ile incelemiştir. Çalışması sonucunda okul kaynaklarından teknoloji kullanımı ve okul iklimi etkileri, Norveç'te anlamlı değilken, Japonya'da negatif yönde Brezilya'da ise pozitif yönde bir etkiye sahiptir. Ailenin okul ile ilişkisi ise öğrenci başarısı üzerinde her üç ülkede anlamlı etkiye sahiptir. Bu çalışmanın bulguları dikkate alındığında bir okul değişkeninin öğrenci başarısı üzerindeki etkisi farklı ülkelerde farklılaştığı görülmektedir. Satıcı da (2008), İş'in (2003) çalışmasına benzer şekilde yapısal eşitlik modeli ile ülkeleri karşılaştırmıştır. Satıcı (2008) çalışmasında Türkiye ve Hong Kong-Çin'de öğrencilerin PISA 2003 matematik başarılarını etkileyen öğrenci, öğretmen ve okul ile ilgili değişkenleri incelemiştir. Çalışması sonucunda Hong Kong-Çin'de öğrencilerin dersteki başarısı ile ilgili rekabetçi düşünceleri; Türkiye de ise okula ait olma örtük değişkeni öğrencilerin başarısına etki eden en önemli yordayıcılar olarak bulunmuştur. En önemli yordayıcının Hong Kong-Çin'de öğrenci düzeyi değişkeni, Türkiye'de ise okul düzeyi değişkenin bulunması oldukça manidardır.

Fuller ve Clarke (1994), gelişmiş ülkelerdeki okul etkisi ile ilgili 100 çalışmayı incelemişlerdir. İncelemeleri sonucunda yönetici eğitim düzeyi veya çalışan değerlendirmesi 100 çalışmanın içinde

etkili okul değişkenleri olarak üç ya da dört kez karşılaşılmıştır. Farrell ve Oliveira (1993) ise incelediği 50 çalışmanın içerisinde 7 analizden 4'ünde okul düzeyi değişkeni olan yönetici niteliğinin öğrenci başarısıyla anlamlı ilişkisi olduğunu gözlemlemiştir. Endüstrileşmiş ülkelerde öğrenci başarısını etkileyen okul değişkenlerinin belirlenmesi için yapılmış 100'den fazla çalışma üzerinde Bosker ve Witziers (aktaran Scheerens ve Bosker, 1997) bir meta-analiz yapmışlardır. Çalışmalarının sonucunda başarıdaki varyansın %19'unun okul düzeyi tarafından açıklanabileceğini fakat bu oranın çalışmadan çalışmaya değişkenlik gösterdiğini belirlemişlerdir. Ryoo (2001), 35 ülkenin TIMSS 1995 verisini kullanarak öğrenci başarısına etki eden öğrenci, okul ve ülke düzeyi değişkenlerini incelemiştir. Araştırmasında HLM analizi kullanan Ryoo (2001), diğer çalışmalardan farklı olarak öğrenci başarısındaki varyansı üç parçaya ayırmıştır ve okul düzeyinin öğrenci başarısındaki varyansın % 11.51'ini açıkladığını tespit etmiştir. Benzer şekilde uygulamaya katılan tüm ülkelerin TIMMS 2007 verisini kullandıkları çalışmalarında Mohammadpour ve Abdul Ghafar (2014), öğrenci başarısındaki varyansı üçe ayırmış ve bu oranı %20.61 olarak tespit etmişlerdir. Bu oran OECD'nin (2004a) PISA 2003 verileri ile çoklu regresyon analizi kullanarak yaptığı çalışmasında yaklaşık %28 olarak hesaplanmıştır. Ülkeler kendi verileri ile yaptıkları çalışmalarda ise bu oran değişmektedir. Ayrıca OECD'nin (2001) PISA 2000 çalışmasına katılan OECD ülkelerinin verileri ile yaptığı çalışmasında okul düzeyi değişkenlerinin (öğretmen morali, katılımı, okul anatomisi, seçiciliği ve öğretmen anatomisi) başarıda açıkladığı varyans yaklaşık %0.5-2 olarak hesaplanmıştır.

Yukarıda verilen farklı analizlerin kullanıldığı, başarıya etki eden okul düzeyi değişkenlerin incelendiği çalışmalar ışığında okul düzeyinin başarıdaki varyansı açıklama oranlarının ülkeden ülkeye değiştiği söylenebilir. Bu bağlamda başarıyı etkileyen okul düzeyi değişkenlerin incelenmesi ülke başarıları arasındaki farklılıkları açıklamaya yardımcı olacaktır. Ayrıca OECD'nin PISA 2012 Türkiye problem çözme alanına ait rapora göre Türkiye ve Sırbistan aynı matematik okuryazarlık düzeyindedir (sırasıyla 408, 420). Fakat raporda Türkiye ve Sırbistan'ın aynı matematik okuryazarlık düzeyinde olmalarına rağmen Sırbistan'ın ortalama problem çözme yeterliği (473), Türkiye'den (454) daha yüksek olduğu ifade edilmiştir. Coğrafi konum olarak Türkiye'ye yakın, eğitim sistemi olarak Türkiye eğitim sistemine benzer bir sistemi olan bu ülkenin ortalama problem çözme yeterliğinin Türkiye'den yüksek olması düşündürücüdür. Bu doğrultuda Türkiye ve Sırbistan'ın ortalama problem çözme yeterliğine etki eden okul özelliklerinin belirlenmesi ve karşılaştırılması amaçlanmıştır.

Araştırmanın Amacı

Bu çalışmanın amacı Türkiye ve Sırbistan'ın ortalama problem çözme yeterliklerine etki eden okul değişkenlerini belirlemek ve karşılaştırmaktır. Bu amaçla aşağıdaki araştırma sorularına cevap aranmıştır.

Araştırma soruları:

1. Türkiye ve Sırbistan'da PISA 2012 çalışmasına katılan okulların problem çözme yeterliği puanları arasında fark var mıdır?

2. Türkiye ve Sırbistan'da PISA 2012 çalışmasına katılan okulların, eğer varsa, problem çözme yeterliği puanları arasındaki farkı açıklayan okul düzeyi değişkenleri nelerdir?

3. Türkiye ve Sırbistan'da etkisi manidar bulunan okul değişkenleri, PISA 2012 problem çözme yeterliği puanlarındaki varyansın ne kadarını açıklamaktadır? Etkisi manidar bulunan okul değişkenleri her iki ülkede benzer midir?

Bu bölümde araştırmanın kapsamı ve yapılma gerekçesi kısaca açıklanmıştır. Çalışmanın amacıyla benzer amaçlar güden diğer çalışmalar literatür başlığı altında detaylı olarak incelenmiştir. Çalışmanın amacı ve literatür incelemesi doğrultusunda yöntem kısmına geçilmiş, çalışmanın örneklemi, ölçme araçları ve veri analizi yöntemi hakkında bilgiler verilmiştir. Veri analizi sonucunda elde edilen bulgular raporlandıktan sonra, bulgular tartışılmış ve sonuçlar özetlenmiştir. Çalışmanın son kısmında ise araştırma sonuçlarına dayalı önerilerde bulunulmuştur.

YÖNTEM

Bu araştırmada Türkiye ve Sırbistan'ın 2012 yılı PISA sınavındaki problem çözme yeterliğine etki eden değişkenlerin belirlenmesi ve karşılaştırılması amaçlandığından nicel araştırma yöntemlerinden nedensel karşılaştırma deseni kullanılmıştır. Nedensel karşılaştırma araştırmaları, ortaya çıkmış/var olan durumun veya gruplar arasındaki farklılıkların nedenlerini, bu nedenleri etkileyen değişkenleri ya da etkinin sonuçlarını koşullar ve katılımcılar üzerinde herhangi bir müdahale olmaksızın belirlemeyi amaçlayan çalışmalardır (Büyüköztürk, Çakmak, Akgün, Karadeniz, ve Demirel, 2008).

Örneklem

Ülkeler PISA uygulamalarını kendi olanakları ile PISA Yönetim Kurulu tarafından belirlenen PISA Teknik Standartlarına uygun bir şekilde yaparlar. Uygulama sonucunda elde edilen verilerin güvenirlikleri ise ilgili yılın PISA teknik raporunda raporlanmaktadır. Çalışmaya katılan ülkelerde iki basamaklı tabakalamalı örnekleme yöntemi kullanılır. İlk olarak her ülke için 15 yaş öğrencilerin bulunduğu okullar oluşturulan listelerden sistematik seçilerek, seçilen her okuldan eşit olasılıkla rastgele olarak 35 öğrenci belirlenir. Türkiye'de PISA 2012'ye, 170 okuldan 4848 öğrenci katılırken; Sırbistan'da 142 okuldan 4353 öğrenci katılmıştır.

Veri Toplama Araçları

Problem çözme alanına ait alt test

Problem çözme yeterliğini ölçmek için geliştirilen ölçme aracı maddelerin hazırlanmasından puanlanmasına kadar geçen süreçte bir dizi uygulamadan geçmiştir. Bu bölümde bu aşamalar kısaca özetlenmiştir. Daha önceki dört PISA değerlendirmesinde kazanılan tecrübeler, mümkün olan en yüksek seviyedeki kültürlerarası ve uluslar arası çeşitliliğe sahip kavramsal olarak titiz materyallere ulaşılmasına yardımcı olmak için çeşitli test merkezlerinin geliştirme uzmanlığını kullanmanın önemini göstermiştir. Bu nedenle PISA 2012 yeni problem çözme maddelerinin hazırlanması için ACER, kültürel olarak farklı ve tanınmış kurumlardan dokuz test geliştirme merkezi ile anlaşmıştır. Süreçte ACER bu merkezler arasında materyallerin dağıtımını koordine etmiş ve her bir merkezdeki madde yazarlarının katıldığı kooperatif geliştirme süreçlerini yönetmiştir. Yeni geliştirilen problem çözme maddelerinin değerlendirmesi için PISA Uluslararası Konsorsiyumu ve ulusal başvurular sonucu kabul edilen uzmanlardan bir uzman grubu oluşturulmuştur. Uzmanı Grubu, test geliştirme merkezlerinin ilk geliştirme çalışmalarından sonra, problemleri çözme yetkinlikleri ile ilgili tüm materyalleri gözden geçirmiştir. Bunun için küçük ölçekli bilişsel laboratuvar etkinlikleri yapılmış ve bunlar Ulusal Merkezler tarafından incelenmiş, saha testi yapılmıştır. Daha sonra test geliştirmenin ikinci aşamasında her bir uzman grubun, ilk geliştirme aşamasından sorumlu olmadığı en az bir testi gözden geçirmiştir. Madde hazırlama sürecine yönelik bir rehber hazırlanmış ve bu konu üzerine çalıştay düzenlenmiştir. Uluslararası ve ulusal gözden geçirmeler sonucunda maddeler, alan uygulamasına tabi tutulmuşlardır. Problem çözme yeterliğini ölçmek için geliştirilen araçta toplamda 42 madde bulunmaktadır. Bu maddeler, ana uygulamadan bir yıl önce, 2011 yılında tüm katılımcı ülkelerde yapılan alan uygulamasında test edilen, yeni geliştirilen 79 adet problem çözme maddesinin oluşturduğu bir havuzundan seçilmiştir. Tüm soruların yanıtlanması için geçen süre 80dakikadır. Ana uygulamada problem çözme soruları, her birinin cevaplanma süresi yaklaşık 20 dakika olacak şekilde 4 gruba ayrılmıştır. Bu dört gruptan farklı ikili kombinasyonlar ile formlar oluşturulmuştur. Uygulamada her bir öğrencinin bir formu yaklaşık 40dakikada cevaplaması gerekmiştir. Uygulamaya katılan ülkeler veri kümesinin güvenilirliği ve ölçme hatası ağırlıklandırılmış olabilirlik kestirimi yöntemi ile sırasıyla 0.79 ve 1.27 olarak kestirilmiştir. Ağırlıklandırılmış olabilirlik kestirimi yöntemi, iç tutarlılığın madde tepki kuramına (MTK) dayalı olarak kestirilmesi şeklinde ifade edilebilir. Ayrıca problem çözme alanının güvenirliği Türkiye için 0.86, Sırbistan için 0.85 olarak PISA 2012 teknik raporunda yayınlanmıştır (OECD, 2014: 32-230).

PISA 2012'de öğrencilerin problem çözme yeterliklerinin değerlendirilmesinde üç alan belirlenmiştir: Problem içeriği/bağlamı, problem durumunun doğası ve problemin çözümündeki bilişsel süreç (OECD, 2013a). Bu çalışmada problem çözme yeterliği puanlarına etki eden değişkenler incelendiği için öğrencilere ait toplam puanlar üzerinden işlem yapılmıştır. Bu nedenle öğrencilerin PISA 2012'de problem çözme yeterliklerinin değerlendirildiği üç alan hakkında detaylı bilgi verilmemiştir.

PISA'da 2003'ten 2012'ye kadar geçen süreçte problem çözmenin değerlendirme çerçevesi çeşitli konularda (kompleks problem çözümü, bilgi aktarımı, bilgisayar destekli problem çözümü gibi) genişletilmiştir. PISA 2012'de bireysel problem çözme kapasitelerin belirlenmesi üzerinde yoğunlaşılmıştır. 2003 yılında kağıt-kalem sınavı olarak uygulanan problem çözme alanına ait alt testi, 2012 yılında bilgisayar destekli olarak uygulanmaya başlanmıştır (OECD, 2013a).

Okul anketi

Bu çalışmada yer alan tüm değişkenler, PISA uygulamalarına katılan öğrencilerin okullarıyla ilişkili verilerin elde edilmesi için OECD tarafından oluşturulan okul anketinden elde edilmiştir. Cevaplaması yaklaşık 30 dakika alan bu anket okul yöneticileri tarafından doldurulmaktadır. Anket 8 bölümden oluşmaktadır. Bu bölümler (OECD, 2013a): 1) Okulun yapısı ve organizasyonu, 2) Öğrenci ve öğretmen yapısı, 3) Okul kaynakları, 4) Okulun eğitimi, müfredatı ve değerlendirmesi, 5) Okul iklimi, 6) Okul politikaları ve uygulamaları, 7) Okuldaki finansal eğitim ve 8) Online okul anketi için ek sorudur.

Verilerin Analizi

Çalışmanın başında PISA 2012 okul anketinde bulunan tüm değişkenlerin problem çözme yeterliği puanına etkilerinin incelenmesi düşünülmüştür. İki ülkeye ait okul değişkenleri aşağıdaki işlemler uygulanarak çalışma amacına uygun şekilde düzenlendikten sonra HLM analizine dahil edilmiştir.

1. Finansal okuryazarlık uygulamasına Türkiye ve Sırbistan katılmadığı için okuldaki finansal eğitim ile ilgili maddeler veriden silinmiştir.
2. Endeks puanı olan değişkenlerin (matematik sınıfı için yetenek grupları, eğitimsel liderlik, öğretmen morali, sorumluluk, vb.) maddeleri (gözlenen değişkenleri) veriden silinmiştir.
3. Çok kategorili değişkenlerin kategori sayısı ikiye indirilmiştir.
4. Bağımlı değişken ile bağımsız değişkenler arasındaki korelasyon incelenmiş ve bağımlı değişken ile ilişkisi olmayan okul değişkenleri veri setinden silinmiştir. Son durumda veri setinde problem çözme yeterliği puanı ile ilişkili Sırbistan verisinde 21 değişken; Türkiye veri setinde ise 47 değişken kalmıştır.
5. Yapılan inceleme sonucunda bu değişkenlerden 12'sinin her iki ülke için ortak olduğu görülmüştür.
6. Kayıp veriler incelenmiştir. Öncelikli olarak kayıp verilerin sistematik olarak dağılıp dağılmadığı kontrol edilmiştir. SPSS programında yapılan kayıp veri analizi ile Little'ın MCAR testi manidar bulunmamıştır. Bu değerin manidar bulunmaması verilerin sistematik olarak dağılmadığını göstermektedir. Kayıp verilerde herhangi bir sistematik dağılım olmadığı görüldüğü ve kategorik veriler için de kayıp veri ataması yapılması gerektiği için SPSS'te çoklu atama (multiple imputation-MI) yapılmıştır.
7. Veride bazı okullarla ilgili çok az bilgi olduğu başka bir ifade ile bu okul müdürlerinin anketteki maddelerin birçoğunu cevaplamadıkları görülmüştür. Bu okullar ve 5'den az öğrenciye sahip okullar analiz dışı bırakılmıştır. Buna paralel olarak ilgili okullardaki öğrencilere ait bilgiler de veriden temizlenmiştir.
8. Uç değer ve çoklu bağlantılılık kontrol edilmiştir ve çoklu bağlantı olmadığı görülmüştür. Mahanalobis uzaklığına göre uç değerler serbestlik derecesi değişken sayısına (12) eşit olmak

üzere, söz konusu serbestlik derecesi için elde edilen ki-kare, üst sınır olarak kabul edilir. Veri setinde bu sınırın üzerindeki değerler uç değer olarak kabul edilmiştir. Türkiye'de 10 okul; Sırbistan'da 9 okul uç değer olarak belirlenmiştir.

Tablo1. HLM Analizine Dahil Edilen Değişkenlere Ait Betimsel İstatistikler

Analiz Düzeyi	Ülkeler	Değişkenler	N	Ortalama	SD	Minimum	Maksimum
Düzey-1	Türkiye	PV1	4494	454.07	78.63	220.55	712.40
		PV2	4494	454.22	79.17	171.85	732.69
		PV3	4494	453.84	78.27	167.79	718.89
		PV4	4494	453.84	78.70	196.20	713.21
		PV5	4494	454.35	79.20	235.16	721.33
	Sırbistan	PV1	4059	470.67	87.79	141.01	770.83
		PV2	4059	471.74	89.47	145.06	742.43
		PV3	4059	471.11	87.47	127.21	757.85
		PV4	4059	471.27	87.58	156.43	739.99
		PV5	4059	470.32	89.27	116.66	743.24
Düzey-2	Türkiye	Engel	147	0.05	1.03	-1.67	3.57
		Nmo	147	4.98	3.43	0.00	18.00
		Terk	147	3.96	5.93	-9.52	30.00
		Bagıs	147	8.48	13.04	0.00	60.00
		Uzmanlık	147	18.24	35.04	0.00	100.00
		OranMo	147	0.12	0.05	0.00	0.23
		OgrtMorali	147	-0.24	1.04	-2.47	1.45
		Eksiklik	147	0.89	1.00	-1.09	3.60
		MatematikYarisi	147	1.82	0.39	1.00	2.00
		SanatKulubu	147	1.51	0.50	1.00	2.00
		Aktivite	147	1.55	0.50	1.00	2.00
		EkstraAktivite	147	1.70	1.07	0.00	3.00
	Sırbistan	Engel	132	-0.21	0.92	-2.37	2.60
		Nmo	132	4.54	2.49	0.00	15.00
		Terk	132	1.70	2.27	-2.46	10.00
		Bagıs	132	18.30	28.71	-12.98	100.00
		Uzmanlık	132	47.63	42.33	-37.41	110.51
		OranMo	132	0.08	0.03	0.00	0.17
		OgrtMorali	132	-0.35	0.85	-2.15	1.45
		Eksiklik	132	-0.77	0.60	-1.77	1.19
		MatematikYarisi	132	1.28	0.45	1.00	2.00
		SanatKulubu	132	1.51	0.50	1.00	2.00
		Aktivite	132	1.13	0.34	1.00	2.00
		EkstraAktivite	132	2.04	0.92	0.00	3.00

9. Son durumda HLM analizinde öncelikle varsayımlar kontrol edilmiştir. Bunun için HLM programının yardımı ile düzey 1 ve düzey 2 için artık dosyaları oluşturulmuştur. Bu dosyalardaki verilerin kullanılmasıyla artıkların normal dağılımı, homojenlikleri ve birbirleriyle olan ilişkileri incelenmiştir. İncelemeler sonucunda varsayımların sağlandığı yani, artıkların normale yakın, homojen bir şekilde dağıldıkları ve birbirleriyle ilişkisiz oldukları görülmüştür. Varsayımların sağlandığı görüldükten sonra araştırma problemlerinin cevaplanması ile ilgili modeller analiz edilmiştir.
10. Son olarak 12 değişkenden hangilerinin analize dahil edilmesi gerektiğini belirlemek için açımlayıcı analiz yapılmıştır. Açımlayıcı analiz HLM programının seçeneklerinden biri olup, modele hangi değişkenlerin dahil edilmesinin uygun olduğuna karar vermek için kullanılan

bir özelliktir. Bu analizde değişkenlerin t değerleri hesaplanmış ve 12 değişkenin de t değeri manidar bulunmuş. T değeri manidar bulunan tüm değişkenler modele dahil edilmiştir.

İncelemeler sonucunda Türkiye örnekleminde 147 okuldan 4494 öğrenci, Sırbistan örnekleminde ise 132 okuldan 4059 öğrenci ait veri kalmıştır. PISA'da her bir öğrencinin problem çözme yeterliği puanı için beş olası değer (plausible value, PV1-5) raporlanmaktadır. Bu değerler OECD tarafından öğrenci başarılarının farklı ağırlıklandırma yöntemleriyle hesaplanmış halidir. HLM7.01 programı her bir öğrenciye ait beş puanı aynı anda analize dahil ederek tek bir bağımlı değişken gibi model testini gerçekleştirmektedir. Bu nedenle öğrencilerin problem çözme yeterliği puanlarına ait beş olası değere herhangi bir işlem yapılmadan, doğrudan HLM analizine dahil edilmiştir.

HLM analizinde 1. düzeyi öğrenci değişkenleri, 2. düzeyi ise okul değişkenleri oluşturmaktadır. "Matematik yarışmaları, sanat kulübü ve matematikle ilgili ülkeye özgü aktiviteler" değişkenleri kategorik olarak (evet=1, hayır=2), diğer değişkenler ise sürekli değişken olarak analize dahil edilmiştir. HLM analizine dahil edilen değişkenlere ait betimsel istatistiklere Tablo 1'de yer verilmiştir. Tablo 1'de "Engel" değişkeni, öğrencilerin öğrenmelerini engelleyici özelliklerini; "NMO" değişkeni örneklemdeki okullarda görev yapmakta olan toplam matematik öğretmeni sayısını; "Terk" değişkeni öğrencilerin gittikleri okuldaki yıllık ortalama okulu bırakan öğrenci sayısını, "Bagıs" değişkeni velilerin okulu destekleme miktarını, "Uzmanlık" değişkeni öğretmenlerin uzmanlıklarını, tecrübelerini; "OranMo" değişkeni bir okuldaki matematik öğretmeni sayısının toplam öğretmen sayısına oranını; "OgrtMorali" değişkeni öğretmenlerin moralini; "Eksiklik" değişkeni okuldaki öğretmen eksikliğini; "MatematikYarisi" değişkeni okullarda düzenlenen matematik yarışmalarını; "Aktivite" değişkeni ülkeye özgü matematiksel etkinlikleri; "EkstraAktivite" değişkeni ise her ülkenin matematik ile ilgili yaptığı ekstra aktiviteleri ifade etmektedir. Ayrıca veri analizinde bazı değişkenlerin indeks puanı kullanıldığı için betimsel istatistiklerinde farklı minimum ve maksimum değerlere sahip oldukları görülmektedir.

HLM analizi

Sosyal bilimlerde veri analizi için, kullanılan analiz yöntemine bağlı olarak veriler bir üst düzeyde toplanabilmekte (aggregation) veya üst düzeye ait veriler alt düzeye yayılabilmektedir (disaggregation). Verilerin tek bir düzeye toplanmasının bir amacı, tek düzeyli analiz yöntemlerinin verilerin analizinde kullanabilmelerini sağlamaktır. Tek düzeyli analiz yöntemleri gözlemlerin bağımsızlığı ve varyansların homojenliği sayıltılarının sağlanmasını gerektirmektedirler. Geniş örneklemlerden elde edilen veriler bu sayıltıların sağlanmasını yapıları gereği zorlaştırmaktadırlar. Aynı okula giden öğrencilerin birbirlerine ilgili değişken konusunda farklı okula giden öğrencilerden daha çok benzemeleri gözlemlerin bağımsızlığı varsayımının ihlaline işaret eder ve bu duruma örnek verilebilir (Raudenbush ve Bryk, 2002). Ayrıca geniş örneklemlerde belirlenen özellik bakımından bir okul homojenken diğer okul heterojen olabilmektedir. Böyle bir durumda ise varyansların homojenliği sayıltısı ihlal edilmiş olur (Hox, 2010, s.4-7). Bu varsayımların ihlali sonucunda tek düzeyli analiz yöntemleri yanlı kestirimler yapabilmektedir (Bryk ve Raudenbush, 1988). Bu nedenle geniş örneklemli çalışmalarda tek düzeyli analiz yöntemlerinin olası yanlı sonuçlarından kaçınmak için çok düzeyli analiz yöntemlerinin kullanılması önerilmektedir (Raudenbush ve Bryk, 2002, s. 3-6). Bu çalışmada verilerin analizi için çok düzeyli analiz yöntemlerinden hiyerarşik lineer modeller (HLM) kullanılmıştır. HLM analizinde kullanılan algoritma standart hataları düzelterek (adjust) sonuçların yanlılığını ortadan kaldırır. Bu nedenle HLM analizinde tek düzeyli analizler (OLS, doğrusal regresyon, vb.) gibi katsayı tahminlerine ait standart hata değerleri olduğundan daha düşük kestirilmez. Ayrıca HLM'de bireysel etkiler incelenebildiği gibi grup etkileri de incelenebilmektedir ve farklı düzeyler arasındaki hipotezler eş zamanlı test edilebilmektedir (Cadiz, 2001; Raudenbush ve Bryk, 2002, s. 38).

Bu araştırmada, araştırma problemlerinin cevaplanması için HLM7.01 programı kullanılmıştır. Aşağıda bu araştırma kapsamında test edilen HLM modellerine yer verilmiştir.

1) Tek-yönlü varyans analizi rastgele etkiler modeli: Bu model "PISA 2012 çalışmasına katılan okulların problem çözme yeterliği puanları arasında fark var mıdır?" sorusunu cevaplamak için

kurulmuştur. Literatürde "boş model" olarak da bilinen bu model ile ilk araştırma problemi cevaplanırken eş zamanlı olarak veri analizinde HLM kullanımının uygun olup olmadığı incelenir. Model eşitlikleri aşağıdaki gibidir:

Düzey 1: $Y_{ij}=\beta_{0j}+r_{ij}$ Düzey 2: $\beta_{0j}=\gamma_{00}+u_{0j}$

Burada,

Y_{ij} : Her bir öğrencinin problem çözme yeterliği puanını,

β_{0j} : j. okulun problem çözme yeterliği puan ortalamasını,

r_{ij} : j. okuldaki i. öğrencinin hata puanını, yani i. öğrencinin j okulunun ortalama problem çözme yeterliği puanından farkıdır. "0" ortalama ve σ^2 varyansı ile normal dağılım gösterdiği varsayılır.

γ_{00}: Tüm öğrencilerin problem çözme yeterliği puan ortalamalarını,

u_{0j} : j. okuldaki hata puanını göstermektedir, yani j okulunun genel başarı ortalamasından farkıdır. "0" ortalama ve τ_{00} varyansı ile normal dağılım gösterdiği varsayılır.

2) Ortalamaların bağımlı olduğu model: İkinci araştırma probleminin cevaplanması için kurulan bu modelde ikinci düzeye problem çözme yeterliği ile ilişkili okul değişkenleri atanarak HLM analizi yapılır. Analiz sonucuna bağlı olarak hangi okul değişkenlerinin problem çözme yeterliğine etki ettiği belirlenir.

Düzey 1: $Y_{ij}=\beta_{0j}+r_{ij}$

Düzey 2:

$\beta_{0j}=\gamma_{00}+\gamma_{01}(Engel)+\gamma_{02}(Nmo)+\gamma_{03}(Terk)+\gamma_{04}(Bagıs)+\gamma_{05}(Uzmanlık)+\gamma_{06}(OranMo)+\gamma_{07}(OgrtMrrali)+\gamma_{08}(Eksiklik)+\gamma_{09}(MatematikYarisi)+\gamma_{10}(SanatKulubu)+\gamma_{11}(Aktivite)+\gamma_{12}(EkstraAktivite)+u_{0j}$

Bu model için birinci modelden farklı olarak γ_{0k} sembolü tanımlanabilir.

γ_{0k}: Okul düzeyi bağımsız değişkenlerindeki (k=1,2,3,4) bir birim değişikliğin okul ortalama problem çözme yeterliği puanlarında meydana getirdiği değişiklik olarak yorumlanır.

BULGULAR

Bu bölümle araştırma problemleri ile ilgili bulgulara yer verilmiştir.

1. Araştırma Problemi ile İlgili Bulgular

PISA 2012 çalışmasına katılan okulların problem çözme yeterliği puanları arasında bir farkın olup olmadığının belirlenmesi için oluşturulan tek-yönlü varyans analizi rastgele etkiler modeli analiz sonuçları Tablo 2a ve Tablo 2b'de verilmiştir.

Tablo 2a. Tek-Yönlü Varyans Analizi Sabit Etkiler Modeli Analiz Sonuçları

Ülkeler	Sabit etkiler	Katsayılar	Standart hata(SH)	*T*	Yaklaşık *s.d.*
Türkiye	Kesim noktası, γ_{00}	449.43*	4.85	92.73	146
Sırbistan	Kesim noktası, γ_{00}	468.71*	5.05	92.86	131

*$p<0.05$

Tablo 2a incelendiğinde PISA 2012 problem çözme yeterliği için genel ortalama Türkiye için 4.85 standart hata ile 449.43 olarak; Sırbistan için 5.05 standart hata ile 468.71 olarak kestirilmiştir. Kestirilen genel ortalamalar için güven aralığı hesaplandığında (%95CI(γ_{00}) = $\gamma_{00}\pm$ (1.96)(SH))

Türkiye'nin genel ortalamasının gerçek değerinin 439.92-458.94 puanları arasında; Sırbistan'ın genel ortalamasının gerçek değerinin ise 458.81-478.61 puanları arasında olması beklenir. Ayrıca analiz sonucunda Türkiye genel ortalamasının güvenirlik katsayısı 0.97 olarak, Sırbistan'ın ise 0.94 olarak bulunmuştur.

Tablo 2b. Tek-Yönlü Varyans Analizi Rastgele Etkiler Modeli Analiz Sonuçları

	Rastgele Etkiler	Standart Sapma	Varyans bileşenleri	*s.d.*	χ^2
Türkiye	Düzey-2, (u_0)	57.76*	3335.68	146	5047.44
	Düzey-1, (r_{ij})	54.59	2980.00		
Sırbistan	Düzey-2, (u_0)	55.74*	3107.07	131	2533.34
	Düzey-1, (r_{ij})	70.04	4904.92		

*p<0.05

Tablo 2b'de Türkiye için ortalama problem çözme yeterliği puanının okullar içi değişkenliğin (σ^2) 2980.00 olarak ve okullar arası değişkenliğin (τ_{00}) 3335.68 olarak kestirildiği görülmektedir (χ^2=5047.44, sd=146). Sırbistan için ortalama problem çözme yeterliği puanının okullar içi değişkenliğin (σ^2) 4904.92 olarak ve okullar arası değişkenliğin (τ_{00}) 3107.07olarak kestirildiği görülmektedir (χ^2=2533.34, sd=131). Her iki ülke için (u_{00}) katsayısının istatistiksel olarak anlamlılığı incelendiğinde bu katsayıya ait p değerinin anlamlı olduğu (p<0.05) görülmüştür. Bu katsayının anlamlı olması her iki ülkede problem çözme yeterliğinin okuldan okula değiştiğini göstermektedir. Başka bir ifade ile PISA 2012 çalışmasına katılan okulların ortalama problem çözme yeterliği puanları arasında fark vardır. Bu doğrultuda verinin yuvalanmış bir yapı gösterdiği, bunun için verinin çok düzeyli modeller ile analiz edilmesi gerektiği belirlenmiştir.

Problem çözme yeterliği puanlarındaki değişkenliğin düzeyler tarafından ne kadar açıklandığı sınıflar arası korelasyon (ICC) yardımıyla hesaplanır (Raudenbush ve Bryk, 2002).

Türkiye için:

Okullar içi açıklanan varyans oranı: $\hat{p} = 2980.00/(3335.68 + 2980.00)=0.47$

Okullar arası açıklanan varyans oranı: $\hat{p} = 3335.68/(3335.68 + 2980.00)=0.53$

Sırbistan için:

Okullar içi açıklanan varyans oranı: $\hat{p} = 4904.92/(3107.07 + 4904.92)=0.61$

Okullar arası açıklanan varyans oranı: $\hat{p} = 3107.07/(3107.07 + 4904.92)=0.39$

Sınıflar arası korelasyonun hesaplanması sonucunda problem çözme yeterliği puanlarındaki değişkenliğin okul düzeyi tarafından açıklanma oranı Türkiye için yaklaşık 0.53 iken Sırbistan için yaklaşık 0.39'dur.

2. Araştırma Problemi ile İlgili Bulgular

PISA 2012 çalışmasına katılan okulların, eğer varsa, problem çözme yeterliği puanları arasındaki farkı açıklayan okul düzeyi değişkenlerinin belirlenmesi için ortalamaların bağımlı olduğu model kurulmuş, modelin analiz sonuçları Türkiye için Tablo 3'te, Sırbistan için Tablo 4'te verilmiştir. Tablo 3 incelendiğinde öğrencilerin problem çözme yeterliği puanlarına 3 değişken etkisinin (terk, öğretmen morali ve matematik yarışı) manidar (p<0.05); 9 (engel, toplam matematik öğretmeni sayısı, bağış, uzmanlık, matematik öğretmeni oranı, eksiklik, sanat kulübü, aktiviteler ve ekstra aktiviteler) değişken etkisinin ise manidar olmadığı (p>0.05) görülmektedir.

Tablo 3. Türkiye için Ortalamaların Bağımlı Olduğu Modeli Sabit Etkiler Analizi Sonucu

Sabit etkiler	Katsayılar	Standart hata	*t*	Yaklaşık s.d	Etki Büyük.
Kesim Noktası, β_{00}	501.14	36.97	13.56	134	
Engel, β_{01}	1.17	4.07	0.29	134	---
Nmo, β_{02}	2.82	1.54	1.835	134	---
Terk, β_{03}	-1.32*	0.64	-2.06	134	-0.02
Bagıs, β_{04}	0.18	0.25	0.70	134	---
Uzmanlık, β_{05}	0.16	0.13	1.23	134	---
OranMo, β_{06}	72.95	82.24	0.89	134	---
OgrtMorali, β_{07}	10.56*	4.65	2.27	134	0.18
Eksiklik, β_{08}	1.44	3.93	0.37	134	---
MatematikYarisi, β_{09}	-48.65*	14.21	-3.42	134	-0.84
SanatKulubu, β_{10}	6.39	12.74	0.50	134	---
Aktiviteler,β_{11}	12.67	8.16	1.55	134	---
EkstraAktiviteler, β_{12}	4.30	5.59	0.77	134	---
Rastgele Etkiler	Standart sapma	Varyans bileşenleri	s.d	χ^2	
Düzey2 hata, u_0	47.95	2299.61	134	3170.36	
Düzey1 hata, *r*	54.59	2980.25			

*$p<0.05$

Tablo 3 incelendiğinde Terk değişkeni katsayısı (β_{03}) yaklaşık 0.64 standart hata ile yaklaşık -1.32 olarak kestirildiği görülmüştür. Bu katsayının *p* değeri istatistiksel olarak manidar bulunduğu ($p<0.05$, $sd=134$) ve negatif bir değer aldığı için okullardaki öğrenci terklerinin artması ile öğrencilerin problem çözme yeterliği puanlarının düştüğü söylenebilir. Başka bir ifade ile okul terkinin fazla olduğu bir okula giden öğrencinin problem çözme yeterliği puanı, okul terkinin az olduğu bir okula giden öğrencinin problem çözme yeterliği puanından 1.32 birim daha düşüktür. Terk değişkeni katsayısının (β_{03}) %95 güven aralığı oluşturulduğunda, gerçek değerinin -0.07 ile -2.57 aralığında olması beklenir. Bu değişkenin gerçek hayatta ne kadar etkiye sahip olduğunu incelemek için etki büyüklüğü hesaplanmıştır. Terk değişkenin etki büyüklüğü göz önüne alındığında (-0.02) bu değişkenin etkisinin günlük hayatta hissedilmeyecek kadar az olduğu söylenebilir.

Ortalamaların bağımlı olduğu model analizi sonucunda istatistiksel olarak manidar etkiye sahip olan bir diğer değişken Ogrtmorali'dir ($p<0.05$, $sd=134$). Ogrtmorali değişkeninin katsayısı (β_{07}) yaklaşık 4.65 standart hata ile yaklaşık 10.56 olarak kestirilmiştir. Bu doğrultuda öğretmen moralinin yüksek olduğu okula giden bir öğrencinin problem çözme yeterliği puanı, öğretmen moralinin düşük olduğu okula giden bir öğrenciden 10.56 birim daha fazladır yorumu yapılır. Ogrtmorali değişkeni katsayısının (β_{03}) %95 güven aralığı oluşturulduğunda, bu katsayının gerçek değerinin 1.45 ile 19.67 aralığında olması beklenir. Bu değişkenin gerçek hayatta etkisinin hissedilip hissedilmediğini incelemek için etki büyüklüğü hesaplanmıştır. Bu değer 0.18 olarak bulunmuştur. Bu değer öğretmen morallerinin yüksek olduğu okullarda öğrenim gören öğrencilerin problem çözme yeterliği puanları, öğretmen moralleri düşük olan okullarda öğrenim gören öğrencilere göre 0.18 standart sapma daha fazla olduğunu ifade etmektedir. Ogrtmorali değişkenin etki büyüklüğü göz önüne alındığında (0.18) bu değişkenin küçük bir etkiye sahip olduğu, başka bir ifade ile günlük hayatta etkisinin çok az hissedildiği söylenebilir.

Son olarak Matyarisi değişkeni katsayısının (β_{09}) istatistiksel olarak manidar olduğu görülmektedir ($p<0.05$, sd=134). Matyarisi değişkeni katsayısı yaklaşık 14.21 standart hata ile yaklaşık -48.65 olarak kestirilmiştir ve bu katsayının gerçek değerinin -76.5 ile -20.8 aralığında olması beklenir. Bu durum matematik dersi için sık sık yarışma düzenlenen bir okula giden öğrencinin, matematik dersi için yarışma düzenlenmeyen bir okula giden öğrenciden 48.65 birim daha düşük problem çözme yeterliği puanına sahip olduğu şeklinde yorumlanır. Bu değişkenin etki büyüklüğü incelendiğinde, değişken etkisinin günlük hayatta hissedilebildiği söylenir. Aynı zamanda matematik dersi için yarışma

düzenlenmeyen okullarda öğrenim gören öğrencilerin problem çözme yeterliği puanları, matematik dersi için yarışma düzenlenen okullarda öğrenim gören öğrencilere göre 0.84 standart sapma daha yüksektir yorumu yapılabilir.

Sırbistan'da PISA 2012 çalışmasına katılan okulların, problem çözme yeterliği puanları arasındaki farkı açıklayan okul düzeyi değişkenlerinin belirlenmesi için kurulmuş olan modelin analiz sonuçları Tablo 4'te verilmiştir. Tablo 4 incelendiğinde öğrencilerin problem çözme yeterliği puanlarına 2 değişken etkisinin (engel ve bağış) manidar (p<0.05); 9 (toplam matematik öğretmeni sayısı, terk, uzmanlık, matematik öğretmeni oranı, öğretmen morali, eksiklik, matematik yarışı, sanat kulübü, aktiviteler ve ekstra aktiviteler) değişken etkisinin ise manidar olmadığı (p>0.05) görülmektedir.

Tablo 4. Sırbistan için Ortalamaların Bağımlı Olduğu Model Analizi Sonucu

Sabit etkiler	Katsayılar	Standart hata	t	Yaklaşık s.d	Etki Büyük.
Kesim Noktası, β_{00}	519.98	43.75	11.89	119	
Engel, β_{01}	-14.15*	5.85	-2.42	119	-0.25
Nmo, β_{02}	0.69	2.35	0.29	119	---
Terk, β_{03}	-1.83	2.33	-0.79	119	---
Bagıs, β_{04}	0.33*	0.13	2.51	119	0.006
Uzmanlık, β_{05}	0.03	0.12	0.27	119	---
OranMo, β_{06}	326.37	189.37	1.72	119	---
OgrtMorali, β_{07}	3.47	6.31	0.55	119	---
Eksiklik, β_{08}	-12.51	8.56	-1.46	119	---
MatematikYarisi, β_{09}	-13.29	9.98	- 1.33	119	---
SanatKulubu, β_{10}	-12.79	12.55	-1.02	119	---
Aktiviteler,β_{11}	-19.97	11.20	-1.78	119	---
EkstraAktiviteler, β_{12}	3.63	8.36	0.43	119	---
Randum Etkiler	Standart sapma	Varyans bileşenleri	s.d	χ^2	
Düzey2, u_0	49.58248	2458.42	119	1832.167	
Düzey1, r	70.03086	4904.32			

*p<0.05

Tablo 4 incelendiğinde Engel değişkeni katsayısı (β_{01}) yaklaşık 5.85 standart hata ile yaklaşık -14.15 olarak kestirildiği görülmüştür. Bu katsayının p değeri istatistiksel olarak manidar bulunduğu (p<0.05, sd=119) ve negatif bir değer aldığı için okullardaki öğrenmeyi engelleyici durumların artması ile öğrencilerin problem çözme yeterliği puanlarının düştüğü söylenebilir. Başka bir ifade ile öğrenmeyi engelleyici durumların fazla olduğu okuldaki bir öğrencinin problem çözme yeterliği puanı, öğrenmeyi engelleyici herhangi bir durumun olmadığı okuldaki öğrencinin problem çözme okuryazarlığı puanından 14.15 birim daha düşüktür. Engel değişkeni katsayısının (β_{03}) %95 güven aralığı oluşturulduğunda, gerçek değerinin -25.62 ile -2.68 aralığında olması beklenir. Engel değişkenin etki büyüklüğü hesaplandığında (-0.25) bu değişkene ait etkinin günlük hayatta çok az hissedildiği söylenebilir. Çok fazla öğrenme engeline sahip okullardaki öğrencilerin, öğrenme engeli olmayan okullardaki öğrencilere göre problem çözme okuryazarlığı 0.25 standart sapma daha düşüktür.

Tablo 4 incelendiğinde Ailebagisi değişkeni katsayısının (β_{04}) istatiksel olarak manidar kestirildiği görülmektedir (p<0.05, sd=119). Ailebagisi değişkeni katsayısı yaklaşık 0.13 standart hata ile yaklaşık 0.33 olarak kestirilmiştir ve bu katsayının gerçek değerinin 0.08 ile 0.58 aralığında olması beklenir. Bu katsayının istatistiksel olarak manidar bulunması aileler tarafından desteklenen okullardaki öğrencilerin problem çözme yeterliğinin, aileler tarafından desteklenmeyen okullardaki öğrencilere göre 0.33 birim daha fazla olduğu seklinde yorumlanır. Etki büyüklüğü incelendiğinde, bu değişken etkisinin günlük hayatta hissedilemeyecek kadar küçük olduğu söylenebilir.

3. Araştırma Problemi ile İlgili Bulgular

Türkiye ve Sırbistan'da etkisi manidar bulunan okul değişkenleri, PISA 2012 problem çözme yeterliğindeki varyansın ne kadarını açıkladığına dair sorunun cevaplanması için ortalamaların bağımlı olduğu model rastgele etkiler analiz sonucu Tablo 5'te raporlanmıştır.

Tablo 5. Türkiye için Ortalamaların Bağımlı Olduğu Modeli Rastgele Etkiler Analizi Sonucu

Ülkeler	Rastgele Etkiler	Standart sapma	Varyans bileşenleri	s.d	χ^2
Türkiye	Düzey2 hata, u_0	47.95*	2299.61	134	3170.36
	Düzey1 hata, r	54.59	2980.25		
Sırbistan	Düzey2, u0	49.58248*	2458.42	119	1832.167
	Düzey1, r	70.03086	4904.32		

*p<0.05

Tablo 5, ikinci düzeye okul değişkenleri eklendikten sonra okul düzeyindeki varyansın her iki ülke için ne olduğunu göstermektedir. Okul değişkenlerinin ikinci düzey varyansının ne kadarını açıkladıklarının belirlenmesi için Tablo 2'deki varyanstan Tablo 5'teki varyanslar çıkarılarak, Tablo 5'teki varyansa bölünür. Türkiye için okul düzeyinde etkili değişkenlerin (Terk, Ogrtmorali ve Matyarisi), okul düzeyinde açıkladıkları varyans oranı 0.31 olarak hesaplanmıştır $((3335.68 - 2299.61)/3335.68)$. Okul düzeyinin problem çözme yeterliği puanlarındaki değişkenliğin 0.53'ünü açıkladığı göz önüne alınırsa Terk, Ogrtmorali ve Matyarisi değişkenlerinin problem çözme yeterliği puanlarındaki değişkenliğin 0.16'sını açıkladığı belirlenmiştir. Sırbistan için okul düzeyinde etkili değişkenlerin (Engel ve Ailebagisi), okul düzeyinde açıkladıkları varyans oranı 0.21 olarak hesaplanmıştır $((3107.07 - 2458.42)/3107.07)$. Okul düzeyinin problem çözme yeterliği puanlarındaki değişkenliğin 0.39'unu açıkladığı göz önüne alınırsa Engel ve Ailebagisi değişkenlerinin problem çözme yeterliği puanlarındaki değişkenliğin 0.08'ini açıkladığı belirlenmiştir.

Türkiye ve Sırbistan'da PISA 2012 problem çözme yeterliğine etki eden okul değişkenlerinin benzerliklerini incelemek için HLM analizine her iki ülke için ortak 12 değişken dahil edilmiştir. Bu değişkenler: Engel, Matogrts, Terk, Ailebagi, Ogrtuzmanligi, Matogrtorani, Ogrtmorali, Ogrteksikligi, Matyarisi, Sanatklubu, Ulkeyeozguetkinlik ve Ekstraaktivite'dir. Her iki ülke için yapılan HLM analiz sonuçları Tablo 3 ve Tablo 4'te verilmiştir. İlgili tablolar incelendiğinde Türkiye örnekleminde Terk, Ogrtmorali ve Matyarisi değişkenlerinin Sırbistan örnekleminde ise Engel ve Ailebagi değişkenlerinin problem çözme yeterliği üzerindeki etkilerinin manidar olduğu görülmektedir. Sonuç olarak iki ülkede problem çözme yeterliğini yordayan ortak bir değişken olmadığı sonucuna ulaşılmıştır.

TARTIŞMA ve SONUÇLAR

Bu çalışma Türkiye ve Sırbistan'ın PISA 2012 problem çözme yeterliğine etki eden okul değişkenlerinin belirlenmesi ve karşılaştırılması amacıyla yapılmıştır. Bu amaç doğrultusunda her iki ülkeden PISA 2012 uygulamasına katılmış okullardan elde edilen veriler üzerinde ayrı ayrı iki düzeyli HLM analizi yapılmıştır. HLM analizi üç düzeye de izin vermektedir. Eğer bu çalışma üç düzeyli HLM analizi ile yapılacak olsaydı üçüncü düzeyi ülkelere ait değişkenler oluşturacaktı. Fakat araştırmada ilgilenilen sadece iki ülke vardır ve HLM analizinin yansız kestirimde bulunması için son düzeydeki birim sayısının 30 ve üzeri olması gerekir (Maas ve Hox, 2005). Bu nedenle bu araştırma iki düzeyli HLM analizi ile sınırlıdır.

Problem çözme yeterliğine etki eden değişkenleri belirlemek için PISA 2012 okul anketinden yararlanılmıştır. Araştırmanın başında okul anketinde bulunan tüm değişkenler HLM analizine alınmak istenmiştir. Fakat yapılan ön analizler sonucunda Türkiye için 47 değişken, Sırbistan için 21 değişkenin HLM analizine alınabileceği görülmüştür. Her iki ülkede problem çözme yeterliğine etki eden okul değişkenleri belirlenip karşılaştırılmak istendiği için her iki ülkede ortak olan değişkenler

HLM analizine dahil edilmiştir. Bu doğrultuda bu araştırmada sadece 12 değişkenin etkisi incelenebilmiştir ve bu sınırlılıkla bulgular tartışılmıştır.

12 değişken ile sınırlı olan bu çalışmada öncelikle HLM analizinin avantajından faydalanılarak düzeylerin problem çözme yeterliği puanlarındaki varyansı açıklama oranları incelenmiştir. Öğrenci başarısındaki varyansın büyük bir kısmının öğrenci düzeyi tarafından açıklanması beklenir (Scheerens ve Bosker, 1997; Teodorovic, 2005). Nitekim Ryoo'nun 2001 yılında HLM analizi ile yaptığı uluslararası karşılaştırma çalışmasında başarıdaki varyansın büyük çoğunluğunun öğrenci düzeyi tarafından, kalan kısmının okul düzeyi tarafından açıklanması bu durumu desteklemektedir. Benzer şekilde Scheerens ve Bosker'ın (1997) çalışmaları da bu durumu desteklemektedir. Fakat varyansın düzeyler tarafından açıklanma oranı her zaman bu şekilde olmamaktadır. Mohammadpour ve Abdul Ghafar'ın (2014) yaptığı uluslararası karşılaştırma çalışması ise bu duruma örnek verilebilir. Mohammadpour ve Abdul Ghafar (2014) başarıdaki varyansın küçük bir kısmının öğrenci düzeyi tarafından, kalan kısmının ise okul düzeyi tarafından açıklandığını belirlemişlerdir. Bu çalışmada problem çözme yeterliği puanlarındaki varyansın büyük bir kısmı Türkiye'de okul düzeyi tarafından açıklanırken, Sırbistan'da öğrenci düzeyi tarafından açıklanmaktadır. Bu doğrultuda Türkiye'deki okulların, öğrenci başarıları üzerinde daha fazla etkiye sahip olduğu söylenebilir.

Türkiye ve Sırbistan'da problem çözme yeterliğini etkileyen okul değişkenlerinin incelendiği bu çalışmada Satıcı'nın (2008) çalışmasına benzer şekilde ülkelerde farklı değişkenlerin başarıyı etkilediği görülmüştür. Sırbistan için "Engel ve Bagıs" değişken etkileri, Türkiye için "Terk, OgrtMorali ve MatematikYarisi" değişken etkileri istatistiksel olarak manidar bulunmuştur. Sırbistan'daki hemen hemen her okulun ısıtması, aydınlatması, masası, sandalyesi, tahtası endüstrileşmiş ülkelere benzemektedir. Bu nedenle temel okul tesisleri ve teknolojik araç-gereç imkanlarının öğrenci öğrenmelerindeki farklılıklar üzerinde düşük etkilerinin olması beklenilen bir durumdur (Teodorovic, 2005). Ayrıca öğrenci gelişimlerini değerlendirme sıklığı, velilere ulaşılabilirlik ve aile katılım politikası daha çok sınıf düzeyinde incelenen ve etkileri düşük olan değişkenlerdir (Kavgacı, 2010; Teodorovic, 2005). PISA uygulaması gereği her okuldan bir sınıf uygulamaya katıldığı için bu çalışmada sınıf düzeyi okul düzeyi ile birleştirilmiştir ve Teodorovic'in (2005) belirttiği üzere "Bagıs" değişkeninin etkisinin düşük olduğu tespit edilmiştir. Benzer şekilde "Terk" değişkeni de Türk öğrencilerinin başarılarını çok düşük düzeyde etkileyen bir değişkendir. Öğrenim gördüğü okuldan, yöneticilerden ve öğretmenlerden memnun olmayan, gerekli yardımı alamayan, kuralsız davranışların çok olduğu bir okulda öğrenciler, okulu bırakma eğilimindedirler (Şimşek ve Şahin, 2012). Böyle bir ortamda verilen eğitimin kalitesi tartışılır ve öğrenci başarısının düşük olması beklenir. Ayrıca öğrencinin başarısının yanında öğretmenlerin motivasyonu, morali ve okula duydukları güven de düşecektir. Çalıştığı ortamda güvenli hissetmeyen öğretmenlerin işlerinde verimli olması beklenemez. Nitekim Mihyap (2011) yapmış olduğu çalışmasında çalışma koşullarının uygun, güvenli bir okul ortamında çalıştığını düşünen öğretmenlerin öğrencilerinin daha başarılı olduğunu tespit etmiştir.

Literatürde okulların öğrenci başarıları üzerindeki etkilerini araştıran birçok çalışma (Çalık ve Kurt, 2010; Kavgacı, 2010; Mihyap, 2011, vb.) mevcuttur. Bu çalışmalar incelendiğinde etkisi incelenen değişkenleri okul iklimi kavramı altında toplamak mümkündür. Okul iklimi kavramı pek çok farklı şekilde tanımlansa da genel olarak öğrencilerin, öğretmenlerin, yöneticilerin ve velilerin etkilediği ve etkilendiği örgütsel bir özellik olduğu söylenebilir (Çalık ve Kurt, 2010). Bu bağlamda bir okulun eğitimi etkileyen fiziki koşulları, aile-okul ilişkisi, öğretmenlerin ve öğrencilerin okula ilişkin algıları okul iklimi çalışmalarında incelenmektedir. Bu çalışma sonucunda etkisi manidar bulunan değişkenlerin PISA okul anketinde ilgili oldukları kategoriler incelendiğinde "MatematikYarisi" değişkeni dışındaki tüm değişkenlerin "okul iklimi" kategorisinin bileşenleri olduğu görülmüştür. Matematik yarışı değişkeni ise okulun eğitimi, müfredatı ve değerlendirmesi kategorisine aittir.

HLM analizi sonucu iki ülkede farklı değişkenlerin problem çözme yeterliği üzerinde manidar etkileri olduğu görülse de bu değişkenlerin okul iklimi kavramının birer bileşeni olması oldukça dikkate değerdir. Mortimore, Sammons, Stoll, Lewis ve Ecob'un 1988 yılında yaptıkları çalışmada "okul iklimi" en önemli okul düzeyi değişkenlerinden biri olarak belirlenmiştir. 2000'li yıllara gelindiğinde Scheerens (2000) kitabında, okul iklimi gibi değişkenlerin başarıya etkilerini inceleyen çalışmalar

üzerindeki meta-analiz çalışmalarını özetlemiştir. Scheerens (2000) incelemesi sonucunda okul iklimi gibi değişkenlerin az da olsa öğrenci başarısını etkiledikleri sonucuna ulaşmıştır. 2001 yılında ise OECD'nin PISA verilerini kullanarak yaptığı çalışmanın sonucunda benzer şekilde okul anatomisi, öğretmen morali gibi okul iklimi değişkenlerinin öğrencilerin başarılarını etkilediği tespit edilmiştir. Sonuç olarak Mihyap'ın (2011) çalışmasında belirttiği gibi uygun okul iklimi ve çalışma koşullarına sahip okullara giden öğrencilerin başarıları diğer öğrencilere göre anlamlı derecede yüksektir.

HLM analizi sonucunda yukarıda bahsedildiği gibi Sırbistan için "engel ve bağış" değişken etkileri, Türkiye için "terk, öğretmen morali ve matematik yarışı" değişkenlerinin problem çözme yeterliği üzerinde istatistiksel olarak anlamlı etkisi olduğu görülmüştür. Etki büyüklükleri dikkate alındığında ise Sırbistan'da "engel" değişkeninin, Türkiye'de "öğretmen morali ve matematik yarışı" değişkenlerinin problem çözme yeterliği üzerindeki etkilerinin istatistiksel olarak anlamlı olmasının yanı sıra pratikte de önemli oldukları belirlenmiştir. Okul değişkenlerinin problem çözme yeterliği puanlarındaki varyansı açıklama yüzdeleri incelendiğinde Sırbistan için 0.08, Türkiye için 0.16 olduğu görülmüştür. OECD'nin (2001) yaptığı çalışmada bu oran 5.8 olarak hesaplanmıştır. Ancak OECD (2001)'in çalışmasında bu çalışmaya kıyasla okul düzeyinde daha fazla değişken olduğu için başarıdaki varyansın daha fazla açıklanması beklenilen bir durumdur. Teodorovic (2005, s. 74) çalışmasında okul düzeyinde bulunan okul kültürü ve okul iklimi değişkenlerinin Sırp öğrencileri üzerinde küçük etkiye sahip olduklarını ifade ettikten sonra bunun nedenlerini sıralamıştır. Bu nedenlerden birincisi, Farrell ve Oliveira (1993) ve Scheerens'in (2000) çalışmalarında belirttiği üzere endüstriyel ülkelerde etkili okul değişkenlerinin küçük etki büyüklüklerine sahip olmalarıdır. Sırbistan'da okul büyüklüğü, öğrencilerin teknolojik araç-gereçlere ulaşma imkanı gibi konularda endüstriyel ülkelere benzediği için bu değişkenlerin başarıda açıkladıkları varyans düşüktür. İkinci olarak da Sırp öğrenciler vakitlerini daha çok sınıfta geçirdikleri için okul özelliklerinden pek etkilenmiyor olmalarıdır.

ÖNERİLER

Türkiye ve Sırbistan örnekleminde PISA 2012 problem çözme yeterliğini etkileyen okul değişkenlerinin belirlenmesi ve karşılaştırılması amaçlanan bu çalışmada Türkiye'de okul düzeyinin öğrenci başarısında Sırbistan'a göre daha etkili olduğu belirlenmiştir. Bu durum Türkiye'de yetkililerin doğru kararlar alarak okullarda yapacakları değişikliklerle öğrenci başarılarını daha hızlı arttırabileceklerini göstermektedir. Örneğin bu çalışmada öğretmen morali öğrenci başarısını etkileyen diğer bir değişken olarak tespit edilmiştir. Öğrencilerin okulda bulundukları sürede öğretmenlerle olan ilişkileri onların ders başarısını etkilemektedir. Yetkililer, yapacakları iyileştirmeler ve alacakları önlemlerle birlikte öğretmen morallerini yükselterek öğrenci başarılarını arttırabilirler. Ayrıca okullarda yapılan denemelerle öğrenciler sıralanarak karşılaştırmalar yapılmaktadır. Kendini bir yarışın içinde hisseden öğrenci başarısız olmaktadır. Öğrenciler bir yarışın içine sürüklenmeden, bireysel olarak değerlendirilerek başarılarının artması sağlanabilir. Sırbistan'da okul engelinin öğrenci başarısını olumsuz engellediği belirlenmiştir. Sırp yetkililerin okul kaynaklarında yapacakları iyileştirmeler ile öğrenci öğrenmeleri arttırılabilir ve dolayısıyla öğrenci başarıları arttırılabilir.

Bu çalışmada her iki ülke için ortak olan 12 değişkenin etkisi incelenmiştir. Başka bir çalışmada her iki ülke için problem çözme yeterliği ile ilişkili diğer değişkenler analize dahil edilerek ülkeler bireysel değerlendirilebilir. HLM analizi sonucunda Sırbistan için "engel ve bağış" değişken etkileri, Türkiye için "terk, öğretmen morali ve matematik yarışı" değişken etkileri istatistiksel olarak anlamlı bulunmuştur. Bu değişkenler için yürütülecek nitel çalışmalar ile değişken etkileri daha detaylı incelenebilir. Bu çalışmada iki düzeyli HLM analizi kullanılmıştır. Çalışmaya başka ülkeler dahil edilerek üç düzeyli HLM analizi ile çalışma tekrarlanabilir. Bu çalışmadaki değişkenler çok düzeyli yapısal eşitlik modeli ile tekrar incelenebilir.

KAYNAKÇA

Aydın A., Sarıer Y. ve Uysal Ş. (2012). Sosyoekonomik ve sosyokültürel değişkenler açısından PISA matematik sonuçlarının karşılaştırılması. *Eğitim ve Bilim, 37*(164), 20-30.

Baucal, A., Pavlovic-Babic, D., & Willms, J. D. (2007). Differential selection into secondary schools in Serbia. *Prospects, 37*(4), 539-546.

Baucal, A., & Pavlovic-Babic, D. (2009). *Quality and equity of education in Serbia: Educational opportunities of the vulnerable PISA assessment 2003 and 2006 data*. Ministry of Education of the Republic of Serbia.

Bryk, A., & Raudenbush, S. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education, 97*(1), 65-108.

Büyüköztürk, Ş., Çakmak, E. K., Akgün, Ö. E., Karadeniz, Ş. ve Demirel, F. (2008). *Bilimsel araştırma yöntemleri* (14. baskı). Ankara: Pegem Akademi.

Cadiz, J. (2001). *Differences in 4th grade mathematics achievement among chilean elementary schools an application of hierarchical linear models (HLMS)* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database.

Çalık, T. ve Kurt, T. (2010). Okul iklimi ölçeği'nin (OİÖ) geliştirilmesi. *Eğitim ve Bilim Dergisi*, *35*(157), 167-180.

Çelen F. K., Çelik A. ve Seferoğlu S. S. (2011, Şubat). *Türk eğitim sistemi ve PISA sonuçları*. Akademik Bilişim'11-XIII. Akademik Bilişim Konferansı, İnönü Üniversitesi, Malatya.

Demir, E. (2010). *Uluslararası öğrenci değerlendirme programı (PISA) bilişsel alan testlerinde yer alan soru tiplerine göre Türkiye'de öğrenci başarıları* (Yüksek lisans tezi, Hacettepe Üniversitesi, Ankara.) https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp adresinden edinilmiştir.

EARGED. (2005). *OECD PISA 2003 araştırmasının Türkiye ile ilgili sonuçları-PISA 2003 projesi ulusal nihai rapor.* http://pisa.meb.gov.tr/wp-content/uploads/2013/07/PISA-2003-Ulusal-Nihai-Rapor.pdf adresinden erişildi.

Farrell, J. P., & Oliveira, J. B. (1993). *Teachers in developing countries: Improving effectiveness and managing costs.* Washington, DC: The World Bank.

Fuller, B., & Clarke, P. (1994). Raising school effects while ignoring culture? Local conditions and the influence of classroom tools, rules, and pedagogy. *Review of Educational Research, 64(1),* 119-157.

Hox, J. J. (2010), *Multilevel analysis: Techniques and applications*. (2nd ed.) Great Britain: Routledge.

İş, Ç. (2003). *Uluslararası öğrenci başarı belirleme programına göre (PISA) matematik okur yazarlığını belirleyen faktörlerin kültürler arası karşılaştırılması* (Yüksek lisans tezi, Ortadoğu Teknik Üniversitesi, Ankara). https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp adresinden edinilmiştir.

Kavgacı, H. (2010). *İlköğretimde örgütsel iklim ve okul-aile ilişkileri* (Yüksek lisans tezi, Gazi Üniversitesi, Ankara). https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp adresinden edinilmiştir.

Kılıç, S., Çene, E. ve Demir, İ. (2012). Comparison of learning strategies for mathematics achievement in Turkey with eight countries. *Educational Sciences: Theory and Practice, 12*(4), 2594-2598.

Lesh, R., & Zawojewski, J. S. (2007). Problem solving and modeling. In F. Lester (Ed.), *The handbook of research on mathematics teaching and learning* (2nd ed., pp. 763–804). Reston, VA: National Council of Teachers of Mathematics; Charlotte, NC: Information Age Publishing (joint publication).

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86-92.

Mihyap, K. (2011). *Uluslararası matematik ve fen eğilimleri araştırması (2007) tarafından belirlenen öğretmen endekslerinin incelenmesi ve bu değişenlerin sekizinci sınıf Türk öğrencilerinin başarısı ile ilişkisinin analizi* (Yüksek lisans tezi, Ortadoğu Teknik Üniversitesi, Ankara). https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp adresinden edinilmiştir.

Milli Eğitim Bakanlığı (2015). *PISA 2012 araştırması ulusal nihai rapor.* https://drive.google.com/file/d/0B2wxMX5xMcnhaGtnV2x6YWsyY2c/view adresinden erişildi.

Mohammadpour, E., & Abdul Ghafar, M. N. (2014). Mathematics achievement as a function of within-and between-school differences. *Scandinavian Journal of Educational Research, 58*(2), 189-221.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Ecob, R. (1988). *School matters: The junior years.* England: Somerset. Retrieved from https://books.google.com.tr/books

OECD. (2001). *Knowledge and skills for life: First results from PISA 2000.* Paris, France: OECD. Retrieved from https://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/33691620.pdf

OECD. (2004a). *Learning for tomorrow's world: Firsts results from PISA 2003.* Paris, France: OECD. Retrieved from http://www.oecd.org/education/school/programmeforinternationalstudentassessmentpisa/34002216.pdf

OECD. (2004b). *Problem solving for tomorrow's world first measures of cross-curricular competencies from PISA 2003.* Paris, France: OECD. Retrieved from https://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/34009000.pdf

OECD. (2005). *PISA 2003 technical report*. Paris, France: OECD. Retrieved from https://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/35188570.pdf

OECD. (2009). *PISA 2009 assessment framework: Key competencies in reading, mathematics and science.* Paris, France: OECD. Retrieved from https://www.oecd.org/pisa/pisaproducts/44455820.pdf

OECD. (2013a). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris, France: OECD. Retrieved from *http://dx.doi.org/10.1787/9789264190511-en*

OECD. (2013b). *PISA result from PISA 2012 problem solving*. Paris, France: OECD. Retrieved from https://www.oecd.org/pisa/.../PISA-2012-results-turkey.pdf

OECD. (2014). *PISA 2012 technical report*. Paris, France: OECD. Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Newbury Park, CA: Sage.

Ryoo, H. (2001). *Multilevel influences on student achievement: An international comparative study* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database.

Satıcı, K. (2008). *PISA 2003 sonuçlarına göre matematik okuryazarlığını belirleyen faktörler: Türkiye ve Hong Kong-Çin* (Yüksek lisans tezi, Balıkesir Üniversitesi, Balıkesir). https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp adresinden edinilmiştir.

Scheerens, J. (2000). *Improving school effectiveness* (Fundamentals of Educational Planning No.68). http://doc.utwente.nl/92592/1/Improving-122424e.pdf adresinden erişildi.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness.* Oxford, England: Pergamon.

Schleicher, A. (2007). Can competencies assessed by PISA be considered the fundamental school knowledge 15 year olds should possess? *Journal of Educational Change*, *8*, 349-357. doi:10.1007/s10833-007-9042-x

Şimşek, H. ve Şahin, S. (2012). İlköğretim ikinci kademe öğrencilerinde okulu bırakma eğilimi ve nedenleri (Şanlıurfa ili örneği). *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 12*(2), 41-72.

Teodorovic, J. (2005). *Factors related to student achievemenet: What works for children in Serbia?* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database.

Yavuz, E., & Atar, H. Y. (2016). Examining the effects of students and school variables on PISA 2012 problem solving achievement in Turkey. *New Trends and Issues Proceedings on Humanities and Social Sciences*, *2*(5), 24-30. Retrieved from http://sproc.org/ojs/index.php/pntsbs/article/view/1124.

EXTENDED ABSTRACT

Introduction

The weights given to the areas that are measured in different years in PISA vary. In other words, the selected area is examined in more detail. Problem-solving competency was given in 2003 and 2012. In 2003 and 2012 the average of? problem solving competency of Turkey was approximately 408 and 454; and the average of problem solving competency of Serbia is approximately 420 and 473, respectively (OECD, 2005, 2014). According to the OECD's PISA 2012 Turkey problem-solving report, Turkey and Serbia were at the same mathematical literacy level. But Serbia's problem-solving competency was higher than Turkey's. In this study, school variables that affect problem-solving competency of these two countries were examined and compared.

Method

In this study, causal comparison design was used in the quantitative research methods since it is aimed to determine and compare the variables that affect the problem solving competency of Turkey and Serbia in 2012 PISA exam. HLM analysis was performed on the data of 4494 students sampled from 147 schools in Turkey and 4059 students sampled from 132 schools in Serbia separately. All variables in the study were derived from the OECD school survey in order to obtain data related to the schools of pupils participating in PISA applications.

It was not possible to incorporate the school data of the two countries directly into the HLM analysis. For this reason, the school variables of the two countries have been examined in different contexts such as multiple connections, missing data, outliers, and they have been taken from the data when it

is necessary. As a result of primarily analysis, there were twelve variables which were the same for two countries in the HLM analysis.

Results and Discussion

At the beginning of the study, all variables in the school survey were aimed to be included in the HLM analysis. However, as a result of the preliminary analysis, 47 variables for Turkey and 21 variables for Serbia could be taken into HLM analysis. The common variables for both countries were included in the HLM analysis, because school variables that affect problem-solving competency in both countries would be identified and compared. In this respect, only the effects of 12 variables were investigated in this study and the findings were discussed with this limitation.

In this study, which was limited to 12 variables, the explanatory ratios of the variance of problem solving competency of the levels were examined by taking advantage of the HLM analysis. It was expected that a large part of student variance would be explained by student level (Scheerens and Bosker, 1997; Teodorovic, 2005). However, the rate of disclosure by the levels of variance is not always that way, which can be exemplified by an international comparative study by Mohammadpour and Abdul Ghafar (2014). In this study, a large part of the variance in problem solving competency scores was explained by the school level in Turkey, whereas it was explained by the student level in Serbia. In this respect, it can be said that the schools in Turkey had more influence on student achievement.

As a result of HLM analysis, "obstacle and family donation" variable for Serbia and "abandon, teacher morale and mathematics competition" variable for Turkey were statistically significant. Almost every school in Serbia has a warming, lighting, table, chairs and board, which is similar to industrialized countries. For this reason, it was expected that basic school facilities and technological equipment facilities would have a low impact on the differences in student learning (Teodorovic, 2005). The frequency of evaluating student development, accessibility to parents, and family participation policy are variables that were examined at the class level and had little effect (Kavgacı, 2010; Teodorovic, 2005). This study was combined the class level and school level because PISA practice participated in a classroom practice from each school and it was determined that the effect of the "Fundraising" variable was low as indicated by Teodorovic (2005). Similarly, the "abandon" variant was found to be a variable that affects the success of Turkish students at a very low level. According to Şimşek ve Şahin (2012) students tend to leave the school at a school where there is a lot of unqualified behavior. Also they tend to leave when they are not satisfied with the school, the administrators and the teachers. The quality of education given in such an environment was discussed and the student's success was expected to be low. In such an environment, as well as the success of the student, the motivation, morale and confidence of the teacher would decrease. Teachers who do not feel safe in the working environment cannot be expected to be productive in their work. As a matter of fact, Mihyap (2011) found that the students who thought that their working conditions were in a safe and secure school environment were more successful.

In this study of school variables affecting problem-solving competency in Turkey and Serbia, it was seen that different variables in the countries affected problem solving success of students in the same way as in the study of Satıcı (2008). When we examined the categories in which the variables with influence were relevant in the PISA school survey, it was seen that all variables except "math competition" were the components of the "school climate" category. Mathematics competition variable belongs to school education, curriculum and evaluation category. This finding is the result of the study of the OECD in 2001 that used PISA data. It has been determined that school climate variables such as school anatomy and teacher morale affect the success of the students. As a result, as stated in the study of Mihyap (2011), the success of students who go to schools with appropriate school climate and working conditions is significantly higher than the other students. Although it was found that for each countries different variables influence the problem-solving competency, it was quite remarkable that these variables are in common in that they are components of the school climate concept.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(4);453-468

Eğitim Kurumlarında Kullanılan Psikolojik Testlerin Ölçme Standartlarına Göre İncelenmesi*

The Analysis of the Psychological Tests Using In Educational Institutions According To the Testing Standards*

Ezgi MOR DİRLİK** Nizamettin KOÇ***

Öz

Bu araştırmada Rehberlik ve Araştırma Merkezleri ile okulların rehberlik servislerinde sık kullanılan psikolojik testlerden dördünün Amerikan Psikologlar Birliği (APA-American Psychological Association) tarafından yayınlanan Eğitim ve Psikolojide Ölçme Standartları ile Uluslararası Test Komisyonu (ITC-International Testing Commission) tarafından yayınlanan test uyarlama standartlarına uygunluğunu incelemesi amaçlanmıştır. Araştırmaya alınan psikolojik testler, amaçlı örneklem belirleme yöntemiyle belirlenmiştir ve Rehberlik ve Araştırma Merkezleri ile okulların rehberlik servislerinde sıklıkla kullanılan psikolojik testlerden seçilmiştir. Bu testler; Akademik Benlik Kavramı Ölçeği (ABKÖ), Gazi Erken Çocukluk Gelişimi Değerlendirme Aracı (GEÇDA), Temel Kabiliyetler Testi 7-11 Yaş Formu (TKT 7-11) ve Wechsler Çocuklar İçin Zekâ Ölçeği Revize Edilmiş Formu (WISC-R)'dur. Araştırma kapsamında, APA tarafından 1999'da İngilizce olarak yayınlanmış Eğitim ve Psikolojide Ölçme Standartları kitabının geçerlik, güvenirlik ve test geliştirme ve revizyon bölümlerindeki standartlar ile ITC tarafından geliştirilen test uyarlama standartları Türkçeye çevrilmiş ve kontrol listesi oluşturulmuştur. Söz konusu standartlar temel alınarak hazırlanan kontrol listesi kısa ve uzun olmak üzere iki form halinde oluşturulmuştur. Çalışmaya alınan psikolojik testler hazırlanan kontrol listesine göre araştırmacı tarafından incelenmiştir. İncelemelerin güvenirliğini sağlama amacıyla her psikolojik test için yapılan incelemeler üç hafta arayla tekrarlanmıştır. İncelemeler sonucu elde edilen veriler SPSS 20.0 programında betimsel istatistikler kullanılarak incelenmiştir. Araştırma sonucunda incelenen psikolojik testlerin kontrol listesindeki standartları karşılama düzeyleri ve bu standartlar bakımından geliştirilmesi gereken yönleri belirlenmiştir. Elde edilen bulgulara göre, ABKÖ ve GEÇDA için incelemeye alınan standartların yeterli düzeyde karşılanmadığı belirlenmiş ve ABKÖ için yapılan güvenirlik incelemelerinin geliştirilmesi gerektiği sonucuna ulaşılmıştır. Uyarlama çalışmaları olan WISC-R ve TKT 7-11 Yaş Formu'nun test uyarlama standartlarını karşılama düzeyleri incelendiğinde ise, WISC-R için standartların büyük çoğunluğunun karşılandığı, TKT 7-11 Yaş Formu için ise karşılanmayan birden çok standart olduğu sonucuna ulaşılmıştır.

Anahtar Kelimeler: Psikolojik testler, APA eğitimde ve psikolojide ölçme standartları, ITC test uyarlama standartları.

Abstract

The purpose of this research is to analyze four psychological tests which are frequently used in the Guidance and Research Centers and in the guidance services of the schools according to the standards for educational and psychological testing of APA (American Psychological Association) and test adaption standards of ITC (International Testing Commission). The tests were determined based on the goal- oriented sample selecting method and were selected from the most frequently used psychological tests in Guidance and Research Centers and school's guidance centers. These tests are: Scale of Academic Self-Concept (Akademik Benlik Kavramı Ölçeği-ABKÖ), Evaluation of Early Childhood Development Tool (Gazi Erken Çocukluk Gelişimi Değerlendirme Aracı-GEÇDA), Primary Mental Abilities 7-11 (TKT 7-11), and Wechsler

*Bu çalışma Prof. Dr. Nizamettin Koç danışmanlığında ilk yazar tarafından 2013 yılında tamamlanan "Eğitim Kurumlarında Kullanılan Psikolojik Testlerin Ölçme Standartlarına Göre İncelenmesi" isimli yüksek lisans tez çalışmasından üretilmişir.

**Yrd.Doç.Dr, Kastamonu Üniversitesi, Kastamonu-Türkiye, emor@kastamonu.edu.tr, ORCID ID: https://orcid.org/0000-0003-0250-327X

** Prof. Dr.Ankara Üniversitesi, Ankara-Türkiye, nkoc@ankara.edu.tr, ORCID ID: https://orcid.org/0000-0002-3308-7849

DOI: 10.21031/epod.365141
Geliş Tarihi: 15.10.2016
Kabul Tarihi: 28.09.2017

Intelligence Scale for Children Revised Form (WISC-R). In this research, the chapters related to the validity, reliability and test development and revision of "Standards For Educational And Psychological Testing" (APA, 1999) and the adaptation standards developed by ITC were translated into Turkish and a checklist was created by using these documents. The checklist has got two forms as short and long form. The tests were analyzed according to the short form of the checklist by researcher. In order to examine the reliability of these analyses, the analyses were repeated in three weeks' time. Data of these analyses were exported to the Statistical Package for Social Sciences (SPSS) 20.0 and descriptive analysis was perfomed. As a result of this research, the meeting levels of the psychological tests to the test standards in the checklist and the features of the tests which should be improved according to the validity, reliability, test development and revision and test adaptation were determined. In conclusion, the standards analyzed have not been met satisfactorily by ABKÖ and GEÇDA, and according to the analyses of the realibility of ABKÖ, it has been found that more relibility analyses should be done for ABKÖ. As for the results that obtained for the adapted tests, it has been found that most of the adaptation standards have been met for WISC-R and there are many standards that have not been met by for Primary Mental Abilities 7-11.

Keywords: Psychological tests, APA standards for educational and psychological testing, ITC test adaptation standards.

GİRİŞ

Eğitimin evrensel amaçları arasında, bireylerin kendilerini tanımasını sağlayarak doğru karar verme güçlerini geliştirmek ve bu sayede kendilerini gerçekleştirmelerine yardımcı olmak bulunmaktadır (Koç, 1993). Rehberlik ve psikolojik danışma hizmetlerinin amacıyla da paralel olan bu amacı gerçekleştirmek için, bireyin kendisini tanımasına, kontrol etmesine, gelişimine uygun karar vermesine ve ilgi ve yeteneği doğrultusunda eğitim almasına yardımcı olma ve yönlendirme amaçlarıyla çeşitli bilgi edinme yollarına başvurulmaktadır. Önceden belirlenmiş amaçlar doğrultusunda, bireylerin psikolojik özelliklerinin ölçülmesi söz konusu olduğunda psikometrik ve izlenimci yaklaşım olmak üzere iki temel yaklaşım ortaya çıkmaktadır. Psikometrik yaklaşım izlenimci yaklaşıma göre daha fazla tercih edilmekte ve bu yaklaşım kapsamında, bireylere psikolojik bir test uygulanmakta ve izlenimci yaklaşıma göre daha kapsamlı ve derinlemesine bilgiler elde edilmektedir (Anastasi, 1988; Özgüven, 1994). Psikolojik testlerin diğer bilgi toplama yöntemlerine göre daha kolay uygulanabilmesi, objektif olarak puanlanabilmesi, geçerli ve güvenilir gözlemler yapmaya olanak sağlaması, bu ölçme araçlarının daha fazla tercih edilmesinin nedenlerini oluşturmaktadır (Cronbach, 1990).

Ölçekler, doğrudan gözlem yapmanın mümkün olmadığı ancak teorik olarak var olduğu kabul edilen değişkenlerin bireylerdeki düzeylerini ortaya çıkarmayı amaçlayan maddelerden oluşan ölçme araçlarıdır ve söz konusu soyut yapıların bireylerde ne düzeyde olduğunu belirlemek için geliştirilmekte ve bireylere uygulanmaktadır (DeVellis, 2012). Ölçek hazırlamak için iki temel yola başvurulabilir. Bunlardan ilki bireyin ait olduğu kültüre ilişkin özgün bir ölçek geliştirmek, bir diğer yol ise farklı bir kültür için geliştirilmiş, geçerlik ve güvenirliğine ilişkin kanıtlar toplanmış bir ölçeği hedef kültüre uyarlamaktır (Erkuş, 2007). Gerek geliştirilmiş gerekse uyarlanmış ölçekler kullanılarak, bireyler hakkında bilgi toplanırken, geçerli ve güvenilir ölçmelere dayanarak değerlendirmeler yapmak hedeflenmekte, bundan dolayı da söz konusu ölçeklerin mevcut psikometrik niteliklerinin sorgulanması, teknik özelliklerinin belli standartlara uygunluğunun incelenmesi gerekliliği ortaya çıkmaktadır.

Son yıllarda Türkiye'de ulusal kongrelerde sunulan bildiriler ve dergilerdeki yayınlar incelendiğinde, geliştirilen ve uyarlanan ölçek sayısında büyük artış olduğu görülmektedir (Acar Güvendir ve Özer Özkan, 2015; Çüm ve Koç, 2013). Bu durum ölçme ve değerlendirme alanına ilişkin ilginin artmış olması nedeniyle sevindirici bir gelişme olarak nitelendirilebilir, ancak farklı alanlarda çalışan akademisyenler tarafından geliştirilen ve uyarlanan bu ölçeklerin birçoğu pek çok hata ve eksiğe sahiptir (Erkuş, 2007). Ölçme ve değerlendirmeye ilişkin bilgi ve uygulama eksikliğinden kaynaklanan sorunlara ek olarak, geliştirilen ölçeklere yönelik gereken izleme çalışmalarının yeterince yapılmıyor olması da söz konusu ölçekler kullanılarak toplanan bilgilerin geçerlik ve

güvenirliğine gölge düşüren bir gelişmedir. Araştırmacılar tarafından çeşitli çalışmalar kapsamında ölçekler geliştirilmekte, deneme uygulamaları yapılmakta ve yürütülen çalışmanın amacına yönelik olarak kullanılmakta, fakat sonrasında testi geliştiriciler dahi testle ilgili gereken izleme çalışmalarını yapmakta yetersiz kalmaktadırlar. Geliştirilen psikolojik testlerin standardizasyon çalışmaları yapılmamakta, dolayısıyla testler ülke çapında yaygınlık kazanamamakta ve diğer araştırmacılar tarafından bilinmemektedir (Gülgöz, 1994; Öner, 1994). Söz konusu psikolojik testlerle geçerli ve güvenilir ölçmeler yapmak için gerek bu ölçme aracını geliştirenler tarafından gerekse farklı araştırmacılar tarafından izleme çalışmaları gerçekleştirilmeli ve bu testlerin teknik nitelikleri yeni çalışmalarla sorgulanmalıdır.

Ölçek geliştirme çalışmalarındaki sıkıntıların benzerleri test uyarlama çalışmaları için de geçerlidir. Farklı kültürler için geliştirilmiş olan bir ölçeğin Türk kültürüne uyarlanması sırasında ortaya çıkan sorunlar birçok çalışmada incelenmiş ve genel olarak ölçülen psikolojik yapının kültürel eşdeğerliği incelenmeden testlerin kullanıma sunulduğu, çeviri işlemlerinin uyarlama olarak kabul gördüğü ve testlerin psikometrik nitelikleri bakımından takibinin ve güncelleştirilmesinin yapılmadığı sonuçlarına ulaşılmıştır. Ayrıca uyarlanan testlerden elde edilen puanların orijinal kültürden elde edilen normlar temel alınarak yorumlandığı da elde edilen sonuçlar arasındadır (Çıkrıkçı-Demirtaşlı, 2007; Çukur, 1999; Gülgöz, 1994; Savaşır, 1994). Bu durum da uyarlanmış psikolojik testler ile yapılan değerlendirmelere kuşku ile yaklaşılmasına neden olmaktadır. Uyarlama çalışmalarıyla ilgili yaşanan bu sorunlara çözüm üretmek, uyarlama çalışmalarının hedef kültüre uygunluğunu sağlamak, uyarlanan testlerden daha anlamlı ve isabetli sonuçlar elde etmek ve yeni testi hedef kültüre kazandırmak için uyarlama çalışmaları sırasında bir takım standartların gözetilmesini gerekmektedir (Cansever, 1982).

Gerek geliştirilmiş gerekse uyarlanmış psikolojik testlere ilişkin ülke çapında yapılan araştırmaların sayısının da ölçek sayısıyla paralel olarak son dönemlerde artış gösterdiği görülmektedir. Bu alandaki ilk çalışmalardan biri olan Özge (1981) tarafından yapılan araştırmada, o tarihlerde ülke çapında kullanılmakta olan testler belirlenmiş ve mevcut özellikleri incelenmiştir. Bu çalışmaya göre söz konusu tarihte kullanımda olan 126 psikolojik testin yalnızca beşi Türkiye'de geliştirilmiş özgün psikolojik testler iken, 121 testin ise uyarlama ve çeviri olarak kullanılmakta olan testler olduğu belirlenmiştir. Ayrıca araştırmacı geçerlik, güvenirlik ve norm çalışmalarının birlikte yapıldığı test sayısının beş olduğunu ve yalnızca altı psikolojik testin el kitabına sahip olduğunu belirlemiştir. Ramazan (1988) tarafından yapılan çalışmada ise psikolojik testlerle ilgili yayınlar, tez ve klinik çalışmalar ile araştırma merkezlerindeki kaynaklar taranmış ve Türkiye'de o tarihte kullanılmakta olan psikolojik test sayısı 164 olarak belirlenmiştir. Bu psikolojik testlerin 36'sının ölçek geliştirme, 60'ının uyarlama çalışması olduğu ve 68'inin ise yalnızca çeviri işlemine tabii tutulduğu saptanmıştır. Kağıtçıbaşı ve Savaşır (1988)'ın Doğu Akdeniz ülkelerinde insan yeteneklerinin ölçülmesi ile ilgili yapılan çalışmaları inceledikleri çalışmanın Türkiye ile ilgili bölümünde ise ülke çapında yapılan araştırmaların çoğunun batı kültürlerinde geliştirilen ölçeklerin uyarlanması ile ilgili olduğu ve kültüre özgü yeni ölçeklerin geliştirildiği çalışamalara çok az rastlandığını belirtilmiştir (akt. Savaşır, Sezgin ve Erol, 1992).

Psikolojik testlere ilişkin yapılan çalışmalar incelendiğinde, dikkat çeken diğer bir çalışmanın ise 1990 yılında Ölçme ve Değerlendirme Sistemi Özel İhtisas Komisyonu tarafından hazırlanan "Ölçme ve Değerlendirme Sistemi Geliştirme Çalışmaları 1" adlı rapor olduğu görülmektedir. Söz konusu rapor kapsamında, Rehberlik ve Araştırma Merkezleri'nin fiziki ortam ve koşullarının elverişsiz olduğu, psikolojik ölçme araçlarının yetersizliği, hem nicelik hem de nitelik olarak personelin ve bu uygulamalar için ayrılan bütçenin yetersiz olduğu belirtilmiştir. Ayrıca yönlendirme hizmetleri sunan elemanların mesleklerinin gerektirdiği psikolojik ölçme araçları konusunda merkezi ve ulusal düzeyde hizmet sunan bir kuruluşça desteklenmeleri, onların mesleklerini evrensel düzeydeki uygulamalarla bütünleşebilecek bir düzeyde yapabilmelerinin temel koşulu olarak kabul edilmiştir. Böyle bir kuruluşun oluşturulmasıyla hem yönlendirme hizmetlerinde çalışan bireylerin eğitim alabileceği, hem de psikolojik ölçme araçlarıyla ilgili yapılması gereken işlemlerin (uyarlama, geliştirme, dağıtım ve uygulama gibi) tek bir merkezde toplanması gerektiği belirtilmiştir. Bu rapora göre kurulması öngörülen Psikolojik Ölçme Araçları Geliştirme Merkezi'nde Psikolojik Ölçme

Araçları alanında formasyonu olan özellikle üniversitelerden olmak üzere yeterli sayıda uzmandan, yararlanılması gerektiği belirtilmiştir. Ayrıca ÖSYM ve TUBİTAK gibi kurumlarla işbirliği yapılarak bilgisayar destekli programlardan yararlanılabileceği ve bir proje grubunun kurulmasıyla geliştirilmesi hedeflenen psikolojik testlerin eğitim-öğretim etkinlikleriyle bütünleşik bir şekilde deneme uygulamalarının okullarda yapılabileceği rapor kapsamında belirtilmiştir (MEB, 1990). Özetle söz konusu raporda psikolojik testlere ilişkin ülke çapında yaşanan tüm aksaklıklara dikkat çekilmiş ve çözüm önerisi olarak ise, kurumsallaşmaya gidilmesi gerekliliği belirtilmiştir.

Son yıllarda yapılan çalışmalarda ise genellikle ölçek geliştirme ve uyarlama çalışmalarını temel alan makaleler incelemeye alınmakta ve bu çalışmalar kapsamında geliştirilen ya da uyarlanan ölçeklerin ölçek geliştirme ve uyarlama ilkelerine uygunluğu incelenmektedir. Alanda önde gelen dergilerde yayınlanan makalelerin tarandığı geniş çaplı bir araştırma olan çalışmada, Doğan(2009) tarafından 958 makale incelemeye alınmış ve bu çalışmalarda 389 farklı psikolojik testin kullanıldığı belirlenmiştir. Bu testlerin %75'inin uyarlama, %13'ünün telif ve %12'sinin çeviri psikolojik testler olduğu ve uyarlama çalışmalarının büyük oranda dilsel veya kültürel denklik konusuyla ilgili ölçütleri sağlamadıkları belirlenmiştir. Ayrıca farklı çalışmalarda, ölçek geliştirme ve uyarlama makaleleri incelenmiş (Acar Güvendir ve Özer Özkan, 2015; Boztunç Öztürk, Eroğlu ve Kelecioğlu, 2014; Çüm ve Koç, 2013; Tavşancıl, Güler ve Ayan, 2014) ve geliştirilen ya da uyarlanan ölçeklerin geçerlik ve güvenirlik gibi temel psikometrik nitelikleri belirlenmeye çalışılmıştır. Araştırmacı tarafından yapılan bir başka çalışmada ise temel amacı ölçek geliştirme olan doktora tez çalışmaları kapsamında geliştirilmiş olan ölçeklerin, test geliştirme ve revizyon stadartlarına uyumu incelenmiştir. Beş ölçek geliştirme çalışmasının incelendiği bu çalışma sonucunda, incelemeye alınan tezlerden ikisinin standartları oldukça yüksek düzeyde, birinin orta düzeyde karşıladığı belirlenirken, diğer iki tez çalışması kapsamında geliştirilen ölçeklerin standartları karşılama düzeyinin oldukça düşük olduğu belirlenmiştir (Mor Dirlik, 2014).

Psikolojik testlere ilişkin yapılan inceleme çalışmaların sayısındaki artış da ölçek geliştirme ve uyarlama konusunda sıkıntılar yaşandığının göstergelerinden biri olarak kabul edilmekte (Acar Güvendir ve Özer Özkan, 2015) ve ölçeklere ilişkin yapılacak izleme çalışmalarının geçerli ve güvenilir ölçümler elde edebilmek için büyük önem arz ettiğini göstermektedir. Psikolojik testlerin geniş kullanım alanları göz önünde bulundurulduğunda, eğitim kurumlarının da psikolojik testlerden sıkça faydalanılan kurumlardan biri olduğu ve bu kurumlarda bireyleri tanımak, tehşis etmek ve yönlendirme yapma gibi amaçlarla kullanılmakta olan psikolojik testlerin de niteliklerinin sorgulanması gerekli olduğu sonucuna ulaşılmaktadır. Söz konusu inceleme için uluslararası alanda kabul görmüş, geçerlik, güvenirlik ve ölçek geliştirme ve uyarlama standartlarının bir kontrol listesi formunda hazırlanıp uygulayıcılara ve araştırmacılara sunulması, ölçek geliştirme ve uyarlama aşamasındaki sorunlara çözüm olabilecği gibi hali hazırda kullanımda olan ölçeklerin de psikometrik niteliklerinin belirlenmesini sağlayabileceği düşünülmektedir.

Araştırmanın Amacı

Ülke çapında bu alanda yapılan çalışmaların sınırlılığı gözetilerek, hem psikolojik testlerin temel psikometrik nitelikleri olan geçerlik ve güvenirliğini sorgulamak hem de bir kontrol listesi oluşturarak alana katkıda bulunmak için, bu çalışmada Türkiye'de eğitim kurumlarından sıklıkla kullanılan bir grup psikolojik test belirlenerek, bu testlerin American Psychological Association(APA) ve International Testing Commission (ITC) tarafından belirlenmiş olan temel test standartlarına ne derece uygunluk gösterdiğinin araştırılması, mevcut durumlarının betimlenmesi ve varsa geliştirilmesi gereken niteliklerinin belirlenmesi amaçlanmıştır. Bu sayede dünya çapında geçerliliği kabul edilen standartların sıklıkla kullanılan psikolojik testlere uygulanması sağlanarak bu çalışmanın daha sonradan yapılacak çalışmalara örnek oluşturması ve psikolojik testlerin gelişimine katkıda bulunması beklenmektedir.

YÖNTEM

Bu araştırma, tarama modeli kapsamında yer alan betimsel bir araştırmadır. Betimsel araştırmalarda, var olan bir durumun, herhangi bir müdahele olmaksızın, olduğu gibi tanımlanması esastır ve bu tanımlamalar nitel ya da nicel yaklaşıma göre yapılmaktadır (Karasar, 2010). Bu çalışmada nitel yaklaşım tercih edilmiş, çalışmanın verileri doküman inceleme yöntemi ile elde edilmiştir. Elde edilen veriler ise betimsel ve içerik analizi yöntemlerinden yararlanılarak analiz edilmiştir (Yıldırım ve Şimşek, 2011). Bu bağlamda çalışma kapsamında incelemeye alınan psikolojik testlerin belirlenen standartlar gözetilerek derinlemesine incelenmesi ve mevcut durumlarının betimlenmesi amaçlanmıştır.

İncelemeye Alınan Psikolojik Testler

Araştırmanın amacı doğrultusunda standartları karşılama düzeyleri incelenecek psikolojik testlerin belirlenmesi için seçkisiz olmayan örnekleme yöntemlerinden amaçlı örnekleme yöntemi kullanılmış ve uzman görüşleri temel alınarak incelenecek psikolojik testler belirlenmiştir. Rehberlik ve Araştırma Merkezleri ile okulların rehberlik servislerinde sıklıkla kullanıldığı belirlenen testlerden oluşturulan bir anket Ankara ve Bursa'da devlet okullarında görev yapmakta olan psikolojik danışmanlardan oluşan 79 kişilik bir gruba uygulanmış ve katılımcılardan, ankette yer alan psikolojik testlerden geliştirilmesi gereken yönleri bulunanları belirtmeleri istenmiştir. Anket uygulaması sonrasında, uzmanların görüşlerine göre üzerinde çalışma yapılması gerekli görülen psikolojik testler belirlenmiştir. Bu psikolojik testler uzmanlar tarafından belirtilen sıklığa göre şu şekildedir; Wechsler Çocuklar İçin Zekâ Ölçeği Revize Edilmiş Form (WISC-R), Temel Kabiliyetler Testi 7-11 Yaş Formu (TKT 7-11), Akademik Benlik Kavramı Ölçeği (ABKÖ), Gazi Erken Çocukluk Gelişimi Değerlendirme Aracı (GEÇDA) ve Standford Binet Zekâ Testi'dir. Uzmanlar tarafından en çok belirtilen bu beş psikolojik testin, ikisi (Stanford Binet Zekâ Testi ve Wechsler Çocuklar İçin Zekâ Testi) aynı değişkeni, zekâyı, ölçtüğü için bu çalışma kapsamına yalnızca birinin alınması kararlaştırılmış ve WISC-R araştırma kapsamına alınmıştır. Böylece uzmanların görüşü ölçüt olarak alınarak araştırma çerçevesinde incelenecek olan dört psikolojik test belirlenmiştir. Bu çalışma kapsamında incelemeye alınan dört psikolojik testin temel özellikleri aşağıda verilmektedir.

Akademik benlik kavramı ölçeği (ABKÖ): Bireylerin dört yetenek ve 12 ilgi alanına ilişkin benlik kavramını ölçmeyi amaçlayan bir psikolojik testtir. Bu psikolojik test ile ölçülen yetenekler; sözel yetenek, sayısal yetenek, şekil uzay yeteneği ile göz el koordinasyonu iken ölçülen ilgi alanlar ise, fen bilimleri ilgisi, sosyal bilimler ilgisi, ziraat ilgisi, mekanik ilgi, ikna ilgisi, ticaret ilgisi, ayrıntı ilgisi, edebiyat ilgisi, güzel sanatlar ilgisi, müzik ilgisi ve sosyal yardım ilgisidir. Söz konusu psikolojik test dört seçenekli 170 maddeden oluşmaktadır. Testin uygulaması ve puanlaması tamamen testi alan birey tarafından yapılmaktadır. Ölçeğin el kitabında puanlama ve profil oluşturmanın testi alan birey tarafından yapılması gerektiği özellikle vurgulanmış ve bunun dışında yapılan herhangi bir uygulamada ölçekten beklenen faydanın sağlanamayacağı belirtilmiştir. Söz konusu psikolojik testin, testi alandan farklı bir birey tarafından puanlanması durumunda, testi alan bireyin kendini tanıma sorumluluğunun üzerinden alındığına, bu nedenle cevapların gelişigüzel verilebileceğine dair uyarılarda bulunulmuştur (Kuzgun, 2011).

Gazi erken çocukluk gelişimi değerlendirme aracı (GEÇDA): Bu psikolojik test, 0-72 ay çocuklarının gelişimlerinin ayrıntılı olarak değerlendirilmesini, eğitim yaşantılarının düzenlenmesini ve çocuklardaki olası gelişimsel geriliklerin erken tanılanması amaçlanarak geliştirilmiş bir gelişim değerlendirme aracıdır. Psikomotor, bilişsel, dil gelişimi ve sosyal-duygusal gelişimi ölçmeyi amaçlayan dört alt test ve toplam 249 maddeden oluşan test, gelişimsel oyunlar sırasında çocuğun gözlenmesi yoluyla uygulanmakta ve standart bir materyal seti ve el kitabı ile kullanılmaktadır. Ayrıca testi uygulayacak bireylerin "kullanıcı sertifikası" na sahip olması gerekmektedir (Temel, Ersoy, Avcı ve Turla, 2005).

Temel kabiliyetler testi 7-11 yaş formu (TKT): Çok faktörlü zeka kuramının önde gelen savunucularından olan Thurstone tarafından geliştirilen bu test, çoklu faktör kuramına

dayanmaktadır. Thorndike çoklu faktör kuramı çerçevesinde, kelime anlamı, sayısal akıl yürütme, kavrama, ilişkileri görsel algılama gibi faktörlerden söz ederek, zekâyı soyut zekâ, sosyal zekâ ve mekanik zekâ olarak sınıflandırmıştır (Anastasi,1988). Thurstone tarafından ilk olarak 1941 yılında 11-17 yaşları arasındaki çocuklar için geliştirilen bu test daha sonra, 5-7, 7-11 ve 11-17 olarak üç ayrı form halinde düzenlenmiştir. Bu çalışma kapsmında incelenen TKT 7-11 formu, yedi alt teste sahip bir grup testidir. Bu alt testler; "kelimeler", "yer kavramı", "resimler", "kelime gruplaması", "şekil gruplaması", "ayırt etme",ve "hesap" testleridir. Türkçeye uyarlaması ilk olarak 1953 yılında Test Araştırma Bürosu tarafından yapılan test, 2001 yılında maddeler üzerinde çeşitli redaksiyonlar yapılarak tekrar güncelleştirilmiştir. Testin ikinci uyarlama çalışması, Özel Eğitim ve Rehberlik Genel Müdürlüğü tarafından 1998 yılında başlatılmıştır ve 2001 yılında son bulmuştur. Çalışma dâhilinde, maddeler üzerinde yapılan araştırmalarla, madde güçlük ve ayırt edicilik indeksleri hesaplanmış ve bu özellikler bakımında istenilen düzeyde olmayan maddeler ölçekten çıkarılmış, 256 maddeyle başlanan çalışma 181 madde ile sonlanmıştır. Testin tümü ve alt ölçekler için güvenirlik analizleri yapılmıştır. Geçerlik çalışmalarında ise yordama geçerliği hesaplanmış ve ölçüt olarak ders ve sınıf geçme notları kullanılmıştır. Norm çalışmaları da yapıldıktan sonra uyarlanan testin beş özel bir genel yetenek puanı veren bir test olarak Rehberlik ve Araştırma Merkezleri ile okulların rehberlik servislerinde kullanılması uygun bulunmuştur (MEB, 2001a). Testin genel özellikleri incelendiğinde, testi uygulayacak bireylerin psikoloji, rehberlik ve psikolojik danışma veya eğitimde psikolojik hizmetler lisans programı mezunu olmaları ve bununla birlikte; TKT (7 – 11) ile ilgili uygulama sertifikasının verildiği eğitim yaşantılarına katılmış olmaları gerekmektedir. Uygulama yapılacak grup aynı yaştaki bireylerden oluşturulabileceği gibi, farklı yaşlardaki bireylerden de oluşturulabilir. Test uygulanan grup 7 – 11 yaş grubuna dahil ve 10 kişi olmalıdır. Test toplam 80 dakikada uygulanabilmektedir ve her alt test için farklı süreler ayrılmıştır(MEB, 2001b)

Wechsler çocuklar için zekâ ölçeği revize edilmiş form (WISC-R): WISC-R, 6-16 yaş arasındaki bireylere, bireysel olarak uygulanan bir zekâ testidir. Ölçek sözel ve performans bölümlerinden ve 12 alt testten oluşmaktadır. Orijinal formu İngilizce olan bu testin, Türkçeye uyarlama çalışmasında, testin 1949 yılında geliştirilmiş formunun Türkçe çevirisi ele alınmış, fakat yalnızca çevirisi yapılan ölçeğin uygulama için yeterli olmadığı ve bir takım değişikliklerin gerekli olduğu konusunda karara varılarak yeni bir standardizasyon çalışması başlatılmıştır. Bu çalışma kapsamında ilkokul kitapları taranarak , "Genel bilgi", "Benzerlikler" ve " Yargılama" alt testleri için sorular seçilmiş ve denenmiştir. Elde edilen verilere göre, bu alt testlerdeki soruların güçlük sıralaması yeniden yapılmıştır. "Sözcük dağarcığı" altesti için ilkokul kitapları taranmış 36.000 sözcük elde edilmiş ve bunların 36 tanesi seçilerek, bu alt test yeniden oluşturulmuştur. Performans bölümüne dahil olan alt testlerin üçü, resim tamamlama, resim düzenleme ve küplerle desen, 64 kişilik bir çocuk grubuna uygulanmış ve madde analizleri yapılmıştır. Her alt testteki maddeler için madde analizleri yapılmış, maddeler güçlük değerlerine göre yeniden sıralanmıştır. Tüm bu çalışmalardan sonra, bu psikolojik testin Türkiye normlarının elde edilmesi için daha geniş bir örneklem üzerinde uygulama yapılmış ve test için normlar oluşturulmuştur (Savaşır ve Şahin, 1995).

Veri Toplama Yöntemleri

Çalışma kapsamında incelemeye alınacak standart testlerin belirlenmesinden sonra, söz konusu psikolojik testlerin el kitaplarına ulaşılmış ve bu testler APA tarafından geliştirilen Eğitimde ve Psikolojide Ölçme Standartları ile ITC tarafından geliştirilen Test Uyarlama Standartları temel alınarak hazırlanan kontrol listesine göre incelenmiştir.

Kontrol listesinin hazırlanması aşamasında, APA tarafından 1999 yılında İngilizce olarak yayınlanan "Standarts for Educational and Psychological Testing (Eğitim ve Psikolojide Ölçme Standartları)" kitabı incelenmiş ve bu kaynakta yer alan standartlardan geçerlik, güvenirlik ve test geliştirme ve revizyon ile ilgili standartlar araştırmacı tarafından Türkçeye çevrilmiştir. Belirlenen standartlar yalnızca çeviri işleminden farklı olarak, dilsel anlamlılık da göz önünde bulundurularak yalınlaştırılmış ve kullanıcıların daha rahat anlayabileceği şekilde düzenlenmiştir. Bu standartlara ek olarak ITC tarafından belirlenen test uyarlama standartları üzerinde çalışılmış ve bu standartlar da

yalınlaştırılarak Türkçeye çevrilmiştir. Tüm bu standartlar kullanılarak bir kontrol listesi oluşturulmuştur. Uyarlama yaklaşımı ile Türkçeye çevrilen ve kontrol listesi haline getirilen bu form, beş ölçme değerlendirme uzmanı, iki türk dili uzmanı tarafından incelenmiş ve uzman görüşleri temel alınarak oluşturulan kontrol listesi üzerinde çeşitli revizyonlar yapılmıştır. İlgili tüm standartları içeren uzun form ve APA tarafından tüm psikolojik testler tarafından karşılanması gereken temel test standartları olarak belirtilen temel standartlardan oluşan kısa form olmak üzere kontrol listesinin iki ayrı formu oluşturulmuştur.

Araştırma kapsamına alınan psikolojik testler, el kitaplarında yer alan bilgiler doğrultusunda hazırlanan kısa form kontrol listesi çerçevesinde incelenmiştir. Araştırmaya dahil edilen dört psikolojik testten uyarlanmış olan TKT 7-11 Yaş Formu ve WISC-R sadece uyarlama standartlarına göre, diğer iki test, GEÇDA ve ABKÖ, ise geçerlik, güvenirlik ve test geliştirme ve revizyon standartlarına göre incelenmiştir. Psikolojik testlerin belirlenen standartları karşılama düzeyleri 1'den 3'e kadar puanlandırılarak belirlenmiş ve söz konusu standardın incelenen psikolojik test için hiç karşılanmadığı durumda 1 puan, kısmen karşılandığı durumda, 2 puan ve tamamen karşılandığı durumda ise 3 puan verilmiştir. Yapılan test incelemelerinin güvenirliğini saptamak amacıyla incelemeler araştırmacı tarafından üç hafta arayla tekrarlanmış ve iki inceleme arasındaki tutarlılık uyum yüzdesi indeksi kullanılarak hesaplanmıştır. İki inceleme arasındaki sürenin ne kadar olacağı konusunda alan yazında net bir bilgi bulunmazken, bu zaman aralığının çok uzun tutulması araştırmacının konuya tamamen yabancılaşmasına, çok kısa tutulması ise önceki incelemenin etkisi altında kalmasına neden olabileceği belirtilmektedir (Tavşancıl ve Arslan, 2001). Bu gibi olumsuz etkilerin araştırma sürecine dâhil olmasını önlemek için, iki inceleme arasındaki zaman üç hafta olarak belirlenmiştir.

Verilerin Analizi

Araştırmanın amaçları çerçevesinde elde edilen verilerin analiz edilmesi için içerik analizi tekniği kullanılmıştır. Elde edilen ham verilerin sayısallaştırılması için frekans dağılımından yararlanılmıştır. Hazırlanan kontrol listesindeki her bir psikolojik test standart için frekans tablosu oluşturulmuş ve böylece her psikolojik testin, kontrol listesindeki standartları karşılama durumu belirlenmiştir. Araştırmacı tarafından iki farklı zamanda yapılan incelemeler arasındaki tutarlılığın belirlenmesi için ise uyum yüzdesi indeksi değeri hesaplanmıştır. Uyum yüzdesi indeksi, aynı kodlamanın yapıldığı durumların mevcut tüm durumlara oranı hesaplanarak bulunan bir değerdir. Puanlayıcılar arası ya da puanlayıcı içi uyum yüzdesinin %70'den daha yüksek olması beklenmektedir (Tavşancıl ve Arslan, 2001).

BULGULAR

İncelemeye alınan psikolojik testlerin hazırlanan kontrol listesine uygunluğunun belirlenmesi için yapılan tekrarlı incelemelere ilişkin hesaplanan uyum yüzdesi indeksleri Tablo 1'de verilmiştir.

Tablo 1. İncelemeler Arası Uyum Yüzdesi İndeksi

Psikolojik Testler	Uyum Yüzdesi İndeksi
WISC-R	0.90
TKT 7-11	0.83
GEÇDA	0.78
ABKÖ	0.94

Tablo 1'de verilen değerler incelendiğinde, tüm testler için hesaplanan uyum yüzdesi indeksi değerlerinin 0.70'den yüksek olduğu görülmektedir. Uyum yüzdesi indeksinin 0.70 değeri ve

üstünde olması kabul edilebilir bir uyumu gösterdiğinden (Tavşancıl ve Arslan, 2001), her psikolojik test için yapılan incelemeler arasındaki uyumun kabul edilebilir olduğu sonucuna ulaşılmaktadır.

Geçerlik Standartlarının Karşılanma Durumuna Yönelik Bulgular

Kontrol listesinin ilk bölümünde yer alan 10 geçerlik standardına göre incelenen GEÇDA ve ABKÖ'nin bu standartları karşılama durumu ve geçerlik standartları Tablo 2'de sunulmuştur.

Tablo 2. Geçerlik Standartlarının GEÇDA ve ABKÖ için Karşılanma Durumuna İlişkin Özet Tablo

Geçerlik Standartları	GEÇDA	ABKÖ
Test puanlarının tüm yorumlarına ilişkin geçerlik kanıtları sunulmuştur.	Hiç karşılanmadı	Hiç karşılanmadı
Testin uygulanacağı grup/gruplar sınırlandırılmış ve testle ölçülecek yapı betimlenmiştir.	Tamamen karşılandı.	Tamamen karşılandı.
Testin amacı ve kullanımı ile tanımlanan evrenin bağlantısı kurulmuştur. Test kapsamı belirlenirken uygulanan işlemler açıklanmıştır.	Kısmen karşılandı.	Kısmen karşılandı.
Geçerlilik kanıtlarının elde edildiği grup ayrıntılı bir şekilde tanıtılmıştır.	Kısmen karşılandı.	Kısmen karşılandı.
Test maddelerinin kapsam geçerliği ile ilgili uzman yargısına başvurulmuşsa uzmanların eğitimi, deneyimi ve nitelikleri betimlenmiştir.	Kısmen karşılandı.	Hiç karşılanmadı
Testin belli maddeleri ya da bazı alt testleri için özel yorumlar önerilmişse, bununla ilgili kanıtlar ve gerekçeler sunulmuştur.	Hiç karşılanmadı	Hiç karşılanmadı
Bazı test maddelerinin olası fakat kabul edilmeyecek yorumları varsa, puanlayıcı buna karşı uyarılmıştır.	Hiç karşılanmadı	Hiç karşılanmadı
Testin geçerliği bir ya da birden fazla ölçüte dayanarak belirlendiyse, bu ölçütlerin uygunluğu ve nitelikleri belirtilmiştir.	Kısmen karşılandı.	Tamamen karşılandı.
Teste dayalı olarak bileşik puanlar geliştirilmişse, alt puanların ağırlığına ilişkin temeller ve mantık açıklanmıştır.	Hiç karşılanmadı	Hiç karşılanmadı
Alt test puanlarının yorumlanmasında ve bu testler arasındaki puan farklarının yorumlanmasında kullanılacak bilgiler açıkça verilmiştir.	Hiç karşılanmadı	Tamamen karşılandı.

Türkiye'de geliştirilen GEÇDA ve ABKÖ için incelemeye alınan geçerlik standartları ve bu standartların testlerce karşılanma durumu Tablo 2'de görülmektedir. GEÇDA için yapılan incelemelerde 10 geçerlik standardının yalnızca birinin bu test için tamamen karşılandığı, dördünün kısmen karşılandığı ve beşinin hiç karşılanmadığı belirlenmiştir. İncelemeye alınan diğer psikolojik test olan ABKÖ için de benzer bir durum söz konusudur. Geçerlik standartlarından beşi bu test için hiç karşılanmazken, ikisi kısmen, üçü ise tamamen karşılanmıştır.

Güvenirlik Standartlarının Karşılanma Durumuna Yönelik Bulgular

Güvenirlik standartlarına uygunluğu incelenen ABKÖ ve GEÇDA için kontrol listesinin güvenirlik standartları bölümünde yer alan 10 standardın tümü testlerin yapısı gereği incelemeye alınamamıştır. ABKÖ için sekiz standart incelemeye alınırken GEÇDA için dokuz güvenirlik standardı incelenebilmiştir. İncelemelerde kullanılan standartlar ve testlerin bu standartları karşılama durumu Tablo 3'te verilmiştir.

Tablo 3. Güvenirlik Standartlarının GEÇDA ve ABKÖ için Karşılanma Durumuna İlişkin Özet Tablo

Güvenirlik Standartları	GEÇDA	ABKÖ
Yorumlanacak her toplam puan, alt test puanı ya da puanların kombinasyonu için güvenirliğe ilişkin tahminler sunulmuştur.	Hiç karşılanmadı	Hiç karşılanmadı
Test puanlarının güvenirliğinin hesaplanmasında kullanılan her yöntem açıkça tanımlanmış ve bu yöntemlerde kullanılan istatistiksel teknikler sunulmuştur.	Tamamen karşılandı.	Hiç karşılanmadı.
Güvenirlik analizlerinde ve betimsel analizlerde kullanılan örneklem grubunun seçilme süreçleri rapor edilmiştir.	Tamamen karşılandı.	Hiç karşılanmadı.
Testin uygulanma sürecinde önemli değişikliklere izin veriliyorsa, örneklem sayısı yeterli oldukça her uygulamada elde edilen puanlar için ayrı güvenirlik analizleri sunulmuştur.	Hiç karşılanmadı	Hiç karşılanmadı
Testin puanlanmasına öznel yargıların karışması olasılığı söz konusu ise, puanlayıcılar arası tutarlılık ve puanlayıcının birden fazla uygulanan ölçme uygulamalarındaki tutarlılığına dair kanıtlar sağlanmıştır.	Tamamen karşılandı.	İncelenmeye alınmadı.
Testin birkaç farklı sınıf ya da yaş seviyesinde gruplarda uygulanması söz konusu ise ve her sınıf ya da yaş grubu için ayrı normlar sunulmuşsa, her grup ya da yaş seviyesi için ayrı güvenirlik bilgileri sunulmuştur.	Tamamen karşılandı.	Hiç karşılanmadı
Farklı alt gruplarda beklenen ölçmenin standart hatasının ve güvenirlik katsayısının büyük oranda değişebileceğine dair genel olarak deneysel ve teorik sebepler varsa, her grupla ilgili güvenirlik bilgisi sağlanmıştır.	Hiç karşılanmadı	Hiç karşılanmadı
Test maddelerinin bir kısmı ya da alt testler kısmen farklı bir eğilimi ya da yeteneği ölçmek üzere hazırlanmışsa, güvenirlik tahmini için uygulanan işlemler testin çok faktörlü yapısına uygun olarak uygulanmıştır.	Hiç karşılanmadı.	Hiç karşılanmadı..
Ölçmenin standart hatası testin önerilen tüm yorumlamalarında kullanılacak puanlar için rapor edilmiştir.	Hiç karşılanmadı	Hiç karşılanmadı

Tablo 3 incelendiğinde, yapılan incelemeler sonucunda, ABKÖ için incelemeye alınan güvenirlik standartların hiçbirinin karşılanmadığı, GEÇDA için incelenen dokuz güvenirlik standardının beşinin hiç karşılanmadığı, dört standardın ise bu test için tamamen karşılandığı belirlenmiştir. Bu durumda her iki test için yeni güvenirlik kanıtlarına ihtiyaç duyulduğu fakat özellikle ABKÖ için bu durumun daha gerekli olduğu sonucuna ulaşılmıştır.

Test Geliştirme ve Revizyon Standartlarının Karşılanma Durumuna Yönelik Bulgular

Çalışma kapsamında geliştirilen kontrol listesinin üçüncü bölümünde, toplam 15 standarttan oluşan "Psikolojik testlerin sahip olması gereken test geliştirme ve revizyon standartları" yer almaktadır ve Türkiye'de geliştirilmiş olan ve araştırmaya alınan GEÇDA ve ABKÖ için bu standartlar da incelemeye alınmıştır. Test geliştirme ve revizyon bölümünde yer alan 15 standarttan GEÇDA için 14'ü incelemeye alınırken, ABKÖ için 12 standart incelemeye alınmıştır. Her iki test için de bazı standartlar inceleme dışında tutulmuştur. Bunlardan GEÇDA için inceleme dışında tutulan standart "farklı ağırlıklandırılmış maddelerden oluşan testler için incelenmesi gereken bir standarttır" ve bu testin tamamı eşit ağırlıklandırılmış maddelerden oluştuğu için, bu standart GEÇDA için inceleme dışında bırakılmıştır.

İncelemeye alınan diğer psikolojik test olan ABKÖ için bu bölümde incelemeye alınmayan üç standart bulunmaktadır. Bunlardan ilki testte açık uçlu maddeler bulunması durumunda puanlayıcıya sunulması gereken puanlama kriterleri ile ilgili standarttır. Söz konusu testte açık uçlu bir madde bulunmadığı için bu standardın bu test için inceleme dışı tutulmuştur. Bu test için inceleme dışında tutulan diğer bir standart ise GEÇDA için de incelemeye alınmayan ve test maddelerinin farklı ağırlıklandırılmış olması durumunda işe koşulan standarttır. ABKÖ eşit ağırlıklandırılmış maddelerden oluştuğu için bu standart da incelenmeye alınmamıştır. Bu test için inceleme dışı tutulan son standart ise puanlayıcıların seçimi ve eğitimi ile ilgilidir ve ABKÖ bireyin kendi

puanlamasına dayandığı için tek puanlayıcı bireydir ve farklı puanlayıcılara ihtiyaç duyulmuyor olması bu standardın inceleme dışında tutulmasının gerektirmiştir. Her iki psikolojik test için de incelen test geliştirme ve revizyon standartlarını ve standartların bu testler için karşılanma durumu özet olarak Tablo 4'te sunulmuştur.

Tablo 4. Test Geliştirme ve Revizyon Standartlarının GEÇDA ve ABKÖ için Karşılanma Durumuna İlişkin Özet Tablo

Test Geliştirme ve Revizyon Standartları	GEÇDA	ABKÖ
Test geliştirme sürecine dayanan yeterli sayıda kanıt toplanmış ve rapor edilmiştir.	Hiç karşılanmadı	Hiç karşılanmadı.
Testin amacı / amaçları, test edilen özelliğin tanımı ve testin özellikleri ve maddeleri açıkça belirtilmiştir.	Kısmen karşılandı.	Tamamen karşılandı.
Test yönergeleri testin kapsamını, testte bulunan madde sayısını, madde formatını ve madde ve alt test düzenlemelerini tanımlamıştır.	Tamamen karşılandı.	Kısmen karşılandı.
Madde türleri, cevaplama biçimleri, puanlama ve test uygulama işlemleri, testin amacına, ölçülecek özelliğe ve testi alması hedeflenen grubunun özelliklerine dayanılarak seçilmiştir.	Tamamen karşılandı.	Tamamen karşılandı.
Test yönergeleri alanla ve ölçme ile ilişkili, test geliştirenler dışında kalan uzmanlar tarafından gözden geçirilmiştir.	Tamamen karşılandı.	Hiç karşılanmadı.
Maddelerin geliştirilmesi, gözden geçirilmesi, denenmesi ve madde havuzundan seçilmesinde uygulanan işlemler belirtilmiştir.	Kısmen karşılandı.	Hiç karşılanmadı.
Madde seçiminin hangi süreçle yapıldığı, madde seçiminde hangi verilerin kullanıldığı, madde güçlüğü, madde ayırt ediciliği gibi, rapor edilmiştir.	Kısmen karşılandı.	Kısmen karşılandı.
Testin gözden geçirilme süreci, deneysel analizleri ve maddelere ve cevaplama biçimlerine yönelik uzman yargılarına başvurulmuştur.	Kısmen karşılandı.	Hiç karşılanmadı.
Test maddeleri test yönergelerine göre farklı kategorilere ya da alt testlere ayrılmışsa sınıflama işlemleri ve bu sınıflamanın netliği ile uygunluğu belirtilmiştir.	Tamamen karşılandı.	Hiç karşılanmadı.
Testi cevaplayanların açık uçlu maddelerdeki performansını puanlarken kullanılacak kriterler belirtilmiştir.	Tamamen karşılandı.	İncelemeye alınmadı.
Uygulayıcı için hazırlanan yönergeler yeterince açıktır ve vurgulanarak sunulmuştur.	Tamamen karşılandı.	Tamamen karşılandı.
Testi cevaplayacaklar için hazırlanan yönergeler yeterince ayrıntılıdır. Uygun durumlarda materyal ve alıştırma örnekleri ya da soru örnekleri verilmiştir.	Tamamen karşılandı.	Tamamen karşılandı.
Test geliştirici testin uygulama koşullarının kabul edilebilir değişiklikler belirtilmiş ve bu farklı koşulların kabul edilebilir olma sebepleri belgelendirilmiştir.	Hiç karşılanmadı.	Tamamen karşılandı.
Puanlayıcıların seçimi, eğitimi ve puanları tanımlama/betimleme niteleme süreçleri test geliştirici tarafından rapor edilmiştir.	Kısmen karşılandı.	İncelemeye alınmadı.

Tablo 4'te GEÇDA ve ABKÖ için incelenen test geliştirme ve revizyon standartları yer almakta ve testlerin söz konusu standartları karşılama durumu görülmektedir. İncelemeler sonucu yapılan değerlendirmelere göre GEÇDA için incelenen iki stdanrt hiç karşılanmazken, beş standart kısmen ve yedi satdnart ise tamamen kaşılanmaktadır. ABKÖ için yapılan incelemelerde ise incelemeye alınan on iki standardın beşinin hiç karşılanmadığı, ikisinin kısmen ve beşinin ise tamamen karşılandığı belirlenmiştir.

Test Uyarlama Standartlarının Karşılanma Durumuna Yönelik Bulgular

Kontrol listesinin son bölümünde uyarlanmış olan testler için incelenmesi gereken standartlar yer almaktadır. Bu bölümde 15 standart bulunmakta ve araştırma kapsamında incelenen testlerden Türkçeye uyarlamış olan psikolojik testler, Wechsler Çocuklar İçin Zeka Ölçeği Revize Edilmiş Form ve Temel Kabiliyetler Testi 7-11 Yaş Formu bu standartlara göre incelenmiştir. Her iki

psikolojik test için incelemeye alınan test uyarlama standartları ve tetlerin bu standartları karşılama durumu Tablo 5'te sunulmuştur.

Tablo 5**.** WISC-R ve TKT 7-11 Yaş Formu için Uyarlama Standartlarının Karşılanma Durumuna İlişkin Özet Tablo

Test Uyarlama Standartları	WISC-R	TKT 7-11
Çalışmanın ana amaçları göz önünde bulundurularak testin orijinal formu ile hedef dildeki formu arasındaki kültürel farklılıkların etkileri olabildiğince azaltılmıştır.	Tamamen karşılandı.	Hiç karşılanmadı.
Çalışma yapılan evrendeki test veya ölçekler tarafından ölçülen özelliğin örtüşme düzeyi belirlenmiştir.	Tamamen karşılandı.	Hiç karşılanmadı.
Testi uyarlayanlar araştırma yapılan evrende bütün dilsel ve kültürel farklılıkları göz önünde bulundurarak, testin uygulanacağı evrenin kültürel ve dil yapısına uygun test yönergeleri, test puanları ve test maddeleri oluşturmuştur.	Tamamen karşılandı.	Hiç karşılanmadı.
Testi uyarlayanlar test tekniklerinin seçimi, madde formatı, test düzeni ve diğer yapılması gereken işlemleri araştırmanın yapılacağı evrene uygun olmasını sağlamıştır.	Tamamen karşılandı.	Hiç karşılanmadı.
Testi uyarlayanlar uyarlanan ölçeğin geçerliliğine dair bilgileri toplamıştır.	Tamamen karşılandı.	Tamamen karşılandı.
Testi uyarlayanlar, uyarlama sürecinin doğru işlemesi ve diller arası eşitlik sağlanması için hem dilsel hem de psikolojik açıdan sistematik bir yol izlemiştir.	Tamamen karşılandı.	Hiç karşılanmadı.
Uyarlama işlemleri sırasında, uyarlanan her maddenin orijinal testteki maddelere denk olması gözetilmiştir.	Tamamen karşılandı.	Hiç karşılanmadı
Testin uygulanmasını etkileyebilecek çevresel faktörler hedef gruplarda orijinal gruplara olabildiğince denk olacak şekilde düzenlenmiştir.	Tamamen karşılandı.	Tamamen karşılandı.
Testi uyarlayanlar olası problemleri öngörüp bu problemleri çözebilmek için uygun materyal ve yönergeleri hazırlamışlardır.	Tamamen karşılandı.	Tamamen karşılandı.
Testin el kitabı testle ilgili bütün özellikleri ve testin yeni bir kültüre uyarlanması durumundaki tüm gereksinimleri içermektedir.	Tamamen karşılandı.	Tamamen karşılandı.
Testin uygulama eğitimleri hedef dilde gerçekleştirilmiştir.	Tamamen karşılandı.	Tamamen karşılandı.
Testin el kitabında testle ilgili uyulması gereken kesin kurallar bulunmaktadır.	Tamamen karşılandı.	Tamamen karşılandı.
Bir test başka bir kültüre uyarlandığı zaman, testin uyarlanmış hali ve orijinal halinin eşitliğini desteklemek açısından yapılan değişimler raporlaştırılmıştır.	Hiç karşılanmadı	Hiç karşılanmadı
Testi uyarlayanlar testin uygulanacağı hedef evrendeki bireylerin testteki performansları etkileyebilecek sosyo-kültürel ve ekolojik özellikler ile ilgili spesifik bilgilere sahiptir ve bu özelliklerin test sonuçlarının yorumlanması ile ilgili etkilerini belirlemişlerdir.	Tamamen karşılandı.	Tamamen karşılandı.
Testin uygulandığı evrenlerin örneklemleri arasındaki puan farkları görünen değer olarak kabul edilmemiş, ampirik kanıtlarla değişkenlerin anlamlılığı kanıtlanmıştır.	Tamamen karşılandı.	Tamamen karşılandı.

Tablo 5 incelendiğinde çalışmaya alınan uyarlanmış testler olan WISC-R ve TKT 7-11 Yaş Formu için ITC tarafından geliştirilen test uyarlama standartlarının karşılanma durumu görülmektedir. İncelenen testlerden WISC-R için bu standartların karşılanma durumuna bakıldığında, yalnızca bir standardın hiç karşılanmadığı, kalan 14 standardın ise bu test için tamamen karşılandığı görülmekdir. Uyarlama çalışması olarak Türk kültürüne kazandırılan ve bu çalışma kapsamında incelenen bir diğer psikolojik test olan TKT 7-11 Yaş Formu için ise WISC-R'dan farklı bir durum söz konusudur. Bu test için incelemeye alınan yedi standart hiç karşılanmazken, sekiz standart tamamen karşılanmaktadır. Tablo 5'te yer alan bilgilere göre WISC-R'ın uyarlama çalışmasının oldukça nitelikli olduğu fakat aynı durumun TKT 7-11 Yaş Formu için geçerli olmadığı sonucuna ulaşılabilir.

İncelemeye Alınan Psikolojik Testlerin Geliştirilmesi Gereken Yönlerine İlişkin Bulgular

APA tarafından geliştirilen Psikolojide ve Eğitimde Ölçme Standartları ile ITC atarfından geliştirilen Test Uyarlama Standartları temel alınarak hazırlanan kontrol listesine göre dört psikolojik test incelenmiş ve bu testlerin söz konusu standartları karşılama durumları belirlenmiştir. Yapılan incelemeler sonucunda her psikolojik testin incelemeye alınan standartlara göre eksik yönleri tespit edilmiş ve yapılması gereken güncellemeler belirlenmiştir. İncelenen psikolojik testlerden, Türkiye'de geliştirilmiş olan Gazi Erken Çocukluk Gelişimi Değerlendirme Aracı'nın geçerlik standartlarını karşılama düzeyinin düşük olduğunu ve GEÇDA' nın geçerlik standartları temel alınarak tekrar incelenmesi gerekliliğini ortaya çıkmıştır. GEÇDA' nın psikolojik testlerin sahip olması gereken geçerlik standartlarını karşılama düzeyinin arttırılması için yapılması gerekenler şunlardır:

- Testle ölçülen yapı detaylı olarak betimlenmelidir.
- Test puanlarının ilişkin yapılan yorumlamaların dayandığı kanıt ve teoriler sunulmalıdır.
- Testin amacı ile testle temsil edilen evrenin bağlantısı kurulmalıdır.
- Geçerlik kanıtlarının elde edildiği gruplar detaylı bir biçimde tanıtılmalıdır.
- Test maddelerini inceleyen uzmanların deneyimleri sunulmalıdır.
- Test yanıtlarını değerlendirirken puanlayıcıların yapabileceği olası fakat kabul edilmeyecek yorumlar için puanlayıcılar uyarılmalıdır.
- Geçerlik kanıtlarının toplandığı koşullar rapor edilmelidir.

GEÇDA için yapılan güvenirlik incelemesinde ise bu testin güvenirlik standartlarını karşılamada konusunda da yetersiz olduğu belirlenmiştir ve güvenirlik standartlarını karşılama durumunun iyileştirilmesi için yapılması gerekenler belirlenmiştir. Bunlar;

- Yorumlanan tüm puan türleri için güvenirlik bilgileri sağlanmalıdır.
- Farklı gruplar için ölçmenin standart hata değerleri belirlenmeli ve güvenirlik kestirim işlemleri yapılmalıdır.

GEÇDA' nın test geliştirme ve revizyon standartlarımı karşılama durumunun geliştirilmesi için ise yapılması gerekenler aşağıda sıralanmıştır.

- Test geliştirme sürecine dayanan yeterli kanıt toplanmalıdır.
- Uygulama koşullarına ya da testi alan bireylere göre değişiklik gösterebilecek ancak kabul edilebilir olan koşullar belirtilmelidir.
- Testin amaçları, test edilen alanın özellikleri ve test maddeleri gereken açıklıkta belirtilmelidir.
- Maddelerin geliştirilmesi, gözden geçirilmesi ve seçilmesi sürecindeki işlemler daha detaylı olarak betimlenmelidir.
- Uzman yargılarına başvurulmalı ve puanlayıcıların seçimi, eğitimi ve puanları tanımlama düzeyleri detaylı olarak rapor edilmelidir.

Çalışma kapsamında incelemeye alınan bir diğer psikolojik test, Akademik Benlik Kavramı Ölçeği'dir. İncelenen standartlar göz önünde bulundurulduğunda, bu psikolojik teste ilişkin geliştirilmesi gereken nitelikler belirlenmiş ve bu testin özellikle güvenirlik standartları bakımından güncellenmesi ve güvenirliğine ilişkin yeni kanıtların toplanması gereklidir. Söz konusu testin güvenirlik standartlarını karşılama düzeyinin iyileştirilmesi için, yapılması gerekenler şunlardır:

- Yorumlanacak puan türlerine yönelik ölçmenin standart hatası hesaplanmalıdır.
- Güvenirlik analizlerinin yapıldığı yöntem ve kullanılan istatistikler açıklanmalıdır.
- Güvenirlik analizlerinin gerçekleştirildiği örneklem grubu ayrıntılı olarak betimlenmelidir.
- Farklı uygulamalar için farklı güvenirlik analizleri yapılmalıdır.
- Norm gruplarına ilişkin güvenirlik kestirimleri gerçekleştirilmelidir.
- Kullanılan güvenirlik belirleme yöntemleri testin çok faktörlü yapısına uygun olarak seçilmelidir.

İncelemelerden elde edilen sonuçlara göre, Akademik Benlik Kavramı Ölçeği'nin geçerlik standartları bakımından geliştirilmesi ve güncelleştirilmesi gereken noktalar şunlardır:

- Bu teste dair yorumlanacak puanların dayandığı teori ve kanıtlar sunulmalıdır.
- Kapsam geçerliğinin belirlenmesi için uzman yargılarına başvurulmalıdır.
- Testin alt testlerine dair yapılan yorumlarla ilgili kanıtlar sunulmalıdır.
- Madde yanıtlarının değerlendirilmesi sırasında puanlayıcıların yapabileceği olası fakat kabul edilmeyecek yorumlar için puanlayıcılar uyarılmalıdır.
- Geliştirilen bileşik puanlardaki alt puanların mantığı sunulmalıdır.

Test geliştirme ve revizyon standartları temel alındığında ise Akademik Benlik Kavramı Ölçeği'nin geliştirilmesi gereken yönleri şunlardır:

- Test geliştirme sürecine dayanan yeterli sayıda kanıt toplanmalıdır.
- Test yönergeleri daha detaylı olarak sunulmalıdır.
- Yönergeler için uzmanlardan görüş alınmalıdır.
- Maddelerin seçilme ve gözden geçirilme süreçleri betimlenmelidir.
- Testi oluşturan maddeler için uzman görüşüne başvurulmalıdır.
- Madde seçim sürecinde hangi verilerin kullanılmış olduğu da detaylı olarak sunulmalıdır.

Araştırma çerçevesinde Türkiye'de geliştirilen testlerin yanı sıra, uyarlama çalışması olarak kullanılmakta olan Temel Kabiliyetler Testi 7-11 Formu ve Wechsler Çocuklar İçin Zeka Ölçeği Revize Edilmiş Form da yer almaktadır. Bu psikolojik testler uyarlanmış psikolojik testler oldukları için hazırlanan kontrol listesinde yer alan test uyarlama standartlarına göre incelenmiş ve bu standartlara göre geliştirilmesi ve güncelleştirilmesi gereken noktalar belirlenmiştir. Yapılan incelemelere göre, TKT 7-11 Yaş Formu için uyarlama standartları bakımından gözden geçirilmesi gereken noktalar şunlardır:

- Testin orijinal formu ile uyarlanmış formu arasındaki kültürel farkların etkileri olabildiğince azaltılmalıdır.
- Çalışma evrenindeki test ve ölçekler ile ölçülen özelliğin örtüşme düzeyi belirlenmelidir.
- Test yönergeleri ve maddeleri dilsel ve kültürel farklardan arındırılmalı ve madde formatı ve test düzeni hedef evrene uygun olarak seçilmelidir.
- Uyarlanan her maddenin orijinal testteki maddelere denk olup olmadığı araştırılmalıdır.
- Testin uyarlanmış hali ile orijinal halinin eşitliğini desteklemek için yapılan değişimler raporlaştırılmalıdır.

Bu güncelleştirme çalışmaları sonucunda söz konusu psikolojik test uyarlama standartlarının bütününü tamamen karşılayacak bir konuma kavuşabilecektir.
Araştırma kapsamında incelenen son psikolojik test ise bir diğer uyarlanmış test olan Wechsler Çocuklar İçin Zeka Ölçeği Revize Edilmiş Form'dur. Bu psikolojik test için yapılan incelemeler sonucunda, uyarlama standartlarının yüksek düzeyde karşılandığı belirlenmiştir. Söz konusu test için yalnızca bir uyarlama standardı hiç karşılanmamaktadır ve testin uyarlanmış hali ile orijinal halinin eşdeğerliğini desteklemek için yapılan değişimlerin raporlaştırılmasıyla bu test için belirlenen tüm uyarlama standartları karşılanabilecektir.

SONUÇLAR ve TARTIŞMA

Bu araştırmada kapsamında Rehberlik ve Araştırma Merkezleri ile okulların rehberlik servislerinde ortak olarak sıkça kullanılmakta olan dört test belirlenerek, bu testlerin eğitimde ve psikolojide ölçme standartları ile test uyarlama standartlarını karşılama durumu değerlendirilmiştir. Araştırma kapsamına alınan testlerden Türkiye'de geliştirilmiş olanlar, kontrol listesinde yer alan geçerlik, güvenirlik ve test geliştirme ve revizyon standartlarına göre incelenmiş, Türkçeye uyarlanmış olanlarsa uyarlama standartları gözetilerek incelenmiştir.

Çalışma kapsamında elde edilen sonuçlar incelendiğinde, eğitim kurumlarında sıklıkla kullanılan psikolojik testlerden ABKÖ ve GEÇDA'nın özellikle geçerlik ve güvenirlik gibi kritik psikometrik nitelikleri gereken düzeyde karşılamaktan uzak olduğu belirlenmiştir. Söz konusu ölçeklerin test geliştirme ve revizyon standartları bakımından durumları incelendiğinde ise, Acar Güvendir ve Özer Özkan (2015) tarafından yapılan çalışmada da belirtildiği üzere, ölçeklerin geliştirilme süreçleri arasında birliktelik bulunmadığı görülmektedir. GEÇDA ve ABKÖ için temel test geliştirme ve revizyon standartları tamamen karşılanmamakta ve farklı standartlar farklı düzeylerde karşılanmaktadır. Söz konusu standartlara uyumlu olarak sistematik bir yaklaşım izlenmeli ve tüm araştırmacıların standartları takip etmesi ile oluşabilecek hatalar en aza indirgenerek, tekrar eden hataların önüne geçilmelidir.

Araştırma çerçevesinde incelenen diğer iki psikolojik test yabancı kültürlerde geliştirilmiş ve Türkçeye uyarlanarak kazandırılmış olan testlerdir. Dolayısıyla bu testler uyarlama standartlarına göre incelenmiş ve mevcut durumları bu standartlar çerçevesinde betimlenmiştir. Uyarlanmış testlerden ilki Temel Kabiliyetler Testi 7-11 Yaş Formu'dur. Yapılan incelemeler sonucunda, 15 uyarlama standardının sekizinin bu test için tamamen karşılanmakta olduğu, yedisinin ise hiç karşılanmadığı belirlenmiştir. Çüm ve Koç (2013) tarafından yapılan çalışmada incelenen testlerde olduğu gibi, bu psikolojik test için de uyarlama çalışmasının gereklerinden olan eşdeğerlik incelemelerinin ihmal edildiği görülmektedir. Son olarak TKT 7-11 Yaş Formu için yapılan uyarlama çalışması sırasında çok sayıda maddenin test kapsamı dışında bırakılmış olduğu ve bu durumun da daha sonra gerçekleşecek uyarlama çalışmaları için uygunsuz bir örnek oluşturduğu sonucuna ulaşılmıştır.

Bu araştırma kapsamında test uyarlama standartlarına uygunluğu incelenen diğer psikolojik test olan WISC-R' ın uyarlama standartlarını karşılama durumunun TKT 7-11 Yaş Formu'na göre daha yüksek olduğu belirlenmiştir. Bu test için incelenen 15 uyarlama standardının 14'ünün tamamen karşılanmakta olduğu, yalnızca bir standardın karşılanmadığı belirlenmiştir. Bu durumda söz konusu testin uyarlama standartlarına uygun bir şekilde Türk kültürüne uyarlandığı sonucuna varılmıştır.

Çalışma kapsamında elde edilen bulgulara göre, psikolojik testlere ilişkin gereken izleme çalışmalarının yapılması ve testlerin niteliklerinin yeniden sorgulanmasına ve testlerin psikometrik yönlerinin gelişimine katkıda bulunacağı sonucuna ulaşılmıştır. Tavşancıl, Güler ve Ayan (2014) tarafından da belirtildiği üzere, ülke çapında kullanılmakta olan psikolojik testlerin hem denetimini yapan hem de koordinasyon sağlayan bir test geliştirme merkezi kurularak, bu merkez tarafından psikolojik testler için gerekli olan izleme çalışmaları gerçekleştirilebilir. Kurulacak olan test merkezi ile psikolojik testlerin sistematik olarak bilimsel olarak incelenmesi, bu testlerin psikometrik niteliklerinin zamana göre değişiminin belirlenmesi sağlanarak, söz konusu testler kullanılarak verilen kararların isabetli olma olasılığı artırılacaktır. Son olarak bu çalışma kapsamında geliştirilen kontrol listesi Acar Güvendir ve Özer Özkan (2015) tarafından belirtilen ölçek geliştirme ve uyarlama için bir el kitabı mahiyetinde kullanılabilir.

KAYNAKÇA

Acar Güvendir, M. ve Özer Özkan Y. (2015). Türkiye'deki eğitim alanında yayımlanan bilimsel dergilerde ölçek geliştirme ve uyarlama konulu makalelerinin incelenmesi. *Elektronik Sosyal Bilimler Dergisi, 14*(52), 23-23.

American Psycholocigal Assosiation, American Educational Reserach Association & National Council on Measurement in Education. (1999). *Standarts for educational and psychological testing.* Washington: American National Research Education.

Anastasi, A.(1988). *Psychological testing*.(6th Edition). New York: Macmillan).

Boztunç Öztürk, N., Eroğlu, M. G. ve Kelecioğlu, H. (2014, Haziran). Eğitim bilimleri alanında yapılan ölçek uyarlama makalelerinin incelenmesi. IV. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresinde sunulan bildiri, Hacettepe Üniversitesi, Ankara.

Cansever, G. (1982). *Klinik psikolojide değerlendirme yöntemleri*. İstanbul: Boğaziçi Üniversitesi Yayınları Gözlem Matbaacılık Koll. Sti.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th Edition). New York: Harper Collins.

Çıkrıkçı-Demirtaşlı, R. N. (2017). Psikolojik ölçmelere ilişkin doğru bilinen yanlışlar. *Türk Psikoloji Bülteni, 13*(41), 65-68.

Çukur, C. S. (1999). Kültürün psikoloji bilimindeki yeri üzerine görüşler. *Türk Psikoloji Yazıları, 1*(3), 1-16.

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

DeVellis, R. F. (2012). *Scale development theory and application* (3rd Editon). SAGE Publications, Inc.

Erkuş, A. (2007). Ölçek geliştirme ve uyarlama çalışmalarında karşılşaılan sorunlar. *Türk Psikoloji Bülteni, 13*(40), 17-25.

Gülgöz, S. (1994). Test kullanımında temel konular. *Türk Psikoloji Dergisi, 9*(33), 1-8.

Karasar, N. (2010). *Bilimsel arastırma yöntemi-kavramlar, ilkeler, teknikler*.(21. Baskı) Ankara: Nobel. Yayınevi.

Koç, N. (1993). Eğitim sistemimizde ölçme ve değerlendirme alanındaki gelişmeler. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 25*(2), 387-407.

Kuzgun, Y. (2011). *Akademik Benlik Kavramı Ölçeği el kitabı.* Ankara: Nobel.

Milli Eğitim Bakanlığı, Ölçme ve Değerlendirme Sistemi Komisyonu. (1990). *Ölçme ve değerlendirme sistemi özel ihtisas komisyonu raporu.* Ankara: Milli Eğitim Bakanlığı

Milli Egitim Bakanligi. (2001a). *Temel kabiliyetler testi 7-11 uyarlama çalışması el kitabı*. Ankara: Milli Egitim Basimevi.

Milli Egitim Bakanliği. (2001b). *Temel kabiliyetler testi 7-11 yönergesi*. Ankara: Milli Egitim Basimevi.

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.

Öner, N. (1994). Güvenirliği ve/veya geçerliği sınanmış psikolojik testler. *Türk Psikoloji Dergisi , 9*(33), 9-18.

Özge, S. (1981). *Psychological tests used in Turkey: A preliminary survey* (Yayınlanmamış yüksek lisans tezi). Bogaziçi Üniversitesi, İstanbul.

Özgüven, İ. E. (1999). *Psikolojik testler*. Ankara: PDREM.

Ramazan, O. (1988). *Survey for a source book on psychological tests used in Turkey* (Yayınlanmamış yüksek lisans tezi). Bogaziçi Üniversitesi, İstanbul.

Savaşır, I. (1981). *Psikolojik testler* (Rapor No. 7). Ruh Sağlığı ve Hastalıkları, Türkiye Sinir ve Ruh Sağlığı Derneği Yayınları, Ankara.

Savaşır, I. (1994). Ölçek Uyarlamasındaki Sorunlar ve Bazı Çözüm Yolları. *Türk Psikoloji Dergisi* , 9(33), 27-32.

Savaşır, I., Sezgin, N. ve Erol, N. (1992). 0-6 yaş çocukları için gelişim tarama envanteri geliştirilmesi: Ön çalışmalar. *Türk Psikiyatri Dergisi, 3*(2), 33-42.

Savaşır, I. ve Şahin, N. (1995). *Wechsler Çocuklar İçin Zeka Ölçeği.* Ankara: Türk Psikologlar Derneği.

Tavşancıl, E. ve Arslan, E. (2001). *İçerik analizi ve uygulama örnekleri.* İstanbul: Epsilon.

Tavşancıl, E., Güler G. ve Ayan C. (2014, Haziran). *2002-2012 yılları arasında Türkiye'de geliştirilen bazı tutum ölçeği geliştirme çalışmalarının ölçek geliştirme süreci açısından değerlendirilmesi*. IV. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresinde sunulan bildiri, Hacettepe Üniversitesi, Ankara.

Temel, F., Ersoy, Ö., Avcı, N. ve Turla, A. (2005). *Gazi Erken Çocukluk Gelişimi Değerlendirme Aracı.* Ankara: Gazi Üniversitesi.

Yıldırım, A., Şimşek, H. (2011). *Sosyal Bilimlerde Nitel Araştırma Yöntemleri.* (8. Baskı). Ankara: Seçkin Yayıncılık.

EXTENDED ABSTRACT

Introduction

Psychological tests are the source of knowledge used in many fields such as education, personnel selection and journalism. Besides the wide range of usage, by using the results of the tests, thereare o many crucial decisions are made for people. In order to make right decisions about people by using these tests, it is essential to analyze and investigate the quality of psychometric features of the tests. For this reason, the goal of this study is to analyze the psychometric qualities of the psychological tests which are frequently used in the Guidance and Research Centers and the guidance services of the public schools. Analyses of the tests used in this area is much more important because these tests are used to make decisions about students' educational life and according to the results of the tests,

students prefer their occupations, or they are diagnosed about their pace of learning and something like that. Thanks to the analyzing of the tests by using the standards for educational and psychological testing of APA (American Psychological Association) and test adaption standards of ITC (International Testing Commission), it is possible to get more valid inferences about the people and the psychological traits.

Method

The four psychological tests were analyzed in this research and the tests were determined based on the goal- oriented sample selecting method. In order to decide which tests should be analyzed, a questionnaire was prepared which was composed of the list of tests that are in common usage in these institutions. The questionnaire was administered 79 counselors and their opinions were asked to state which tests require revisions. According to results of questionnaire, four tests were determined as requiring revisions. These tests are: Akademik Benlik Kavramı Ölçeği, Gazi Erken Çocukluk Gelişimi Değerlendirme Aracı, Primary Mental Abilities 7-11, and Wechsler Intellegence Scale for Children Revised Form. After determining the tests that were analyzed, a checklist was prepared by researcher to investigate the psychometric properties of the tests. In order to prepare the checklist, the chapters related to the validity, reliability and test development and revision of "Standards For Educational and Psychological Testing" that was published by APA in 1999 were translated into Turkish. Actually, the process was not just a translation, it was considered to make more clear and easy to understand, so the standards were also simplified. In addition to the APA standards, the adaptation standards developed by ITC were translated into Turkish and the checklist was composed by using these standards. The checklist has got two forms as short and long form. While the long form of the checklist is composed of all of the standards included in the relevant chapters, the short form includes only the basic standards which should be investigated for all kinds of psychological tests. The tests that were chosen in this study were analyzed according to the short form of this checklist by researcher, because in the long form, there are so many standards which can be analyzed only the special kind of psychological test. In order to question the reliability of these analyses, the analyses were repeated in three weeks' time by the researcher and the fit index between the analyses was computed. Data of these analyses were exported to the Statistical Package for Social Sciences (SPSS) 20.0 and descriptive analysis was perfomed.

Results and Discussion

As a result of this study, the standards that have been met and have not met satisfactorily by the tests were determined. By this way, the psychometric qualitites of the tests, such as validity and realibility were defined. As for the GEÇDA, most of the validity, reliability and test revision standards were met partially. So it has been found that, for this test, it is a necessity to gather some new proves for validity, realibility and test dveleopment process. As for ABKÖ, none of the reliability standards were met, so this test should be revised especially for reliability. The adapted tests, WISC-R and TKT 7-11, were analyzed according the adaptation standards and it was found that while all of the standards except one were met for WISC-R, only the half of the adaptation standards were met for TKT 7-11. It is clear that adapted test have met more standandards than the others, but especially for TKT 7-11, some revisions should be done in order to get more valid and realiable results. To conclude, all of the psychological test should be revised and their psychometric properties should be investigated to make valid and reliable decision for people's lives.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(4); 469-489

Akran Değerlendirmesinde Puanlayıcı Katılığı Kayması

Rater Severity Drift in Peer Assessment

Bengü BÖRKAN*

Öz

Akran değerlendirmesinde elde edilen puanların geçerliği ve güvenirliği hakkında sağlam psikometrik dayanağı olan ve özellikle puanlayıcı etkisine değinen yeteri kadar çalışma bulunmamaktadır. Bu çalışmada puanlayıcı etkilerinden olan puanlayıcı katılık kaymasının (rater severity drift), akran değerlendirmede ne derece görüldüğü araştırılmıştır. Eğitim fakültesindeki bir ders kapsamında öğrenciler tarafından gerçekleştirilen sözlü sunum performansları aynı dersi alan 29 akran tarafından dereceli puanlama anahtarı kullanılarak puanlanmıştır. İlk üç gün iki sunum, dördüncü gün üç sunum olmak üzere toplam dokuz sunum dört ayrı günde gerçekleştirilmiştir. Puanlayıcı kayması iki farklı çok yüzeyli Rasch ölçme modeli (ayrı modeller ve kukla zaman) yardımıyla incelenmiştir. Her gün için hesaplanan puanlayıcı kestirimlerinden standartlaştırılmış farklar indeksi ve kukla zaman modelinden etkileşim terimleri hesaplanmıştır. Puanlayıcı kayması analizinde, Gün-1 temel gün alınmış, Gün-1'den diğer günlere (Gün-2, 3 ve 4) değişimler incelenmiştir. Analizler genel olarak akran puanlayıcıların arkadaşlarını oldukça cömert bir biçimde puanladıklarını göstermiştir. Puanlayıcılar kendi aralarında kıyaslandığında ise katılık/cömertlik seviyelerinin birbirlerinden farklı olduğu görülmüştür. Sunumlar puanlayıcılar tarafından tutarlı bir şekilde niteliklerine göre sıralandırılmıştır. Puanlayıcı kaymasını incelemek için kullanılan iki yöntem benzer sonuçlar vermiştir. Gün-1 ve 2 arasında puanlayıcı kestirimlerinde bir farklılık görülmemektedir. Her ne kadar ortalamada puanlayıcılar daha cömert puanlama yapsa da, kaymalar istatistiksel olarak anlamlı değildir. Gün-1 ve 3 arasında puanlayıcıların kestirimlerinde önemli kaymaların olduğu puanlayıcıların oranı %38,10'dur. İki yönteme göre de puanlayıcılar ortalamada yaklaşık 0,14 logit kayma gösterip daha katı puanlama davranışı sergilemiştir. Gün-1 ve 4 arasında puanlayıcıların kestirimlerinde önemli kaymaların olduğu puanlayıcıların sayısı standartlaştırılmış farklar yöntemiyle üçgen, etkileşim terimi yöntemiyle birdir. Ortalamada iki yöntemle de puanlayıcılar daha katılaşmıştır. Ortalamada kaymanın en yüksek olduğu Gün-4'tür.

Anahtar Kelimeler: Akran değerlendirme, geçerlik, puanlayıcı kayması, çok yüzeyli Rasch ölçme modeli

Abstract

There are not enough psychometrically sound studies about the validity and reliability of the scores obtained from peer assessment. This study examined degree of rater severity drift, a rater effect, in peer assessment. The college students' presentations were scored by 29 peers in the class using a rating scale. Nine presentations were held on four separate days, two presentations on each of the first three days and three presentations on the fourth day. Drift was investigated with two many-facet Rasch measurement models (separate models and dummy time MFRM). Standardized differences were calculated from the estimates obtained with separate models and interaction terms were calculated with the dummy time MFRM. In drift analysis, shifts in estimations were examined from Day-1 which is a baseline to other three days. Results showed that peer raters varied according to their level of severity and they tend to be lenient. Statistics showed that the quality of the scale was acceptable and its items behaved as expected. In drift analysis, standardized differences and interaction term provided very similar results. Between Day-1 and 2, there was no statistically significant difference in the estimates of the rater severity. Between Day-1 and 3, the percentage of scorers with significant drift in the estimates was 38.10%. The raters' severity shifts on the average of about 0.14 logit and they displayed more severe scoring behavior. Between Day-1 and 4, the number of raters who had significant shifts in their estimates was three according to the standardized difference method, while one according to interaction method. On the average, the raters became more severe. Among three comparison, Day-4 had the

*Yard. Doç. Dr., Boğaziçi Üniversitesi, Istanbul-Turkiye, e-posta: bengu.borkan@boun.edu.tr, ORCID ID: https://orcid.org/0000-0003-1414-1528

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, Cilt 8, Sayı 4, Kış 2017, 469-488.
Journal of Measurement and Evaluation in Education and Psychology, Vol. 8, Issue 4, Winter 2017, 469-488.
DOI: 10.21031/epod.328119
Geliş Tarihi: 12.07.2017
Kabul Tarihi: 21.11.2017

largest rater severity drift on the average; Day 3, however, has the highest number of raters with rater severity drift.

Keywords: Peer assessment, validity, rater drift, Many-Facets Rasch Measurement

GİRİŞ

1774-1826 yılları arasında Glasgow Üniversitesi'nde profesör olan George Jardine, yazma becerisini geliştirmede akran değerlendirme yöntemini ve avantajlarını anlatmıştır (Gaillet, 1992 aktaran Topping, 2009). Okul bağlamında akran değerlendirme, öğrencilerin akranlarının performanslarını ya da oluşturdukları bir ürünün seviyesini, değerini veya niteliğini değerlendirmesi olarak tanımlanır (Falchikov, 1995; Topping, 2009). Akran değerlendirmesi, farklı ortamlarda uygulanabilmektedir. Örneğin yüz yüze yapılabildiği gibi son zamanlarda teknolojinin sağladığı olanaklarla çevrimiçi ortamlarda (ör. Demirbilek, 2015, Tseng & Tsai, 2007, Yang & Tsai, 2010) da kullanılabilmektedir (Topping, 2009). Akran değerlendirme tek taraflı olduğu gibi değerlendiriciler arasında karşılıklı yapılabilmekte, sınıf içinde ya da dışında gerçekleştirilebilmektedir (Topping, 2009).

Son yıllarda, özellikle de biçimlendirici değerlendirme açısından akran değerlendirmesine olan ilgi artmıştır. Akran değerlendirmesi, özel eğitime ihtiyaç duyan öğrenciler de dâhil olmak üzere her yaş grubunda ve tüm eğitim kademelerinde başarıyla kullanılmaktadır (Scruggs & Mastropieri, 1998). Akran değerlendirmesi niteliksel ve/veya niceliksel değerlendirme sonuçları sağlayabilmekte ve bu durum bir sürekliliğin iki ucu olarak düşünülebilmektedir. Bir uçta akran değerlendirme sadece geri bildirim amaçlı niteliksel veri/bilgi sağlarken diğer uçta niceliksel bir puanlama yer alır. Akran değerlendirme bu sürekliliğin herhangi bir noktasında yer alabilir. Benzer şekilde, akran değerlendirmesinin amacı biçimlendirici ve/veya düzey belirleme amaçlı olabilir (Somervell, 1993). Akran değerlendirmenin öğrenme etkinliğinde ve niteliğinde iyileşmeler sağlayabileceğine dair önemli kanıtlar bulunmaktadır (Weaver & Cotrell, 1986). Özellikle yazma becerisinin geliştirilmesinde, öğretmen değerlendirmesi kadar etkili olabilmektedir (Topping, 2009).

Akran değerlendirmesinde farklı yöntemler bulunmaktadır. Bunlardan en az kullanılanı akran sıralamadır (*peer ranking*). Bu yöntemde her bir grup üyesi gruptaki bütün akranlarını ölçülen niteliğe göre en başarılıdan en başarısıza doğru sıralar. Bir diğer yöntemde ise her bir grup üyesi, gruptaki en iyi performansı gösteren akranı ya da akranları aday gösterir (*peer nomination*) (Docy, Segers, & Sluijsmans, 1999; Kane & Lawyer, 1978; Love, 1981). Benzer bir şekilde en başarısız performansı gösteren akran ya da akranlar da aday gösterilebilir (Heyman & Sailors, 2011). Üzerine en çok akademik çalışma yapılan yöntem ise akran derecelendirmesidir (*peer rating*). Bu yöntemde kontrol listesi, derecelendirme anahtarı ya da rubrik yardımıyla her bir grup üyesi gruptaki bütün akranları, belirlenmiş bir dizi ölçüte göre puanlar (Docy, Segers, & Sluijsmans, 1999; Kane & Lawyer, 1978; Love, 1981).

Akran değerlendirme, değerlendirme sürecinde bulunan her iki taraf için de kazanç sağlamaktadır (Topping, 2005; Topping & Ehly, 1998). Bu değerlendirme türüyle farklı öğretim programlarındaki yazma becerisi ve sözlü sunum gibi çeşitli ürün veya davranışlar değerlendirilebilmektedir. Akran değerlendirme biçimlendirici değerlendirme olarak kullanıldığında akılcı sorgulamayı, farkındalığı (sels-disclosure), dolayısıyla anlamanın değerlendirmesini sağlar. Buna ek olarak hataların ve kavram yanılgılarının fark edilmesine olanak sağlayarak bilgilerdeki boşlukların kapatılmasını mümkün kılar ve değerlendirme öncesi, sırası ve sonrasında bilişsel ve üstbilişsel yararlar oluşturur (Topping, 2009)

Güvenirlik ve Geçerlik

Her ölçme işleminde olduğu gibi akran değerlendirme sürecinde elde edilen ölçme sonuçlarının geçerliği ve güvenirliğinin tartışılması gerekmektedir. Alan yazınında kullanılan akran değerlendirmesinin güvenirlik ve geçerliği tabiriyle aslında akran değerlendirmeden elde edilen ölçümlerin geçerliliği ve güvenirliğinden bahsedilmektedir. Alan yazınında akran değerlendirmede

ulaşılan ölçümlerin güvenirlik ve geçerliğini belirlemek için, genel olarak öğrencilerin değerlendirme sonuçlarıyla öğretmenin değerlendirme sonuçları arasındaki uyuma bakılmış ancak puanlayıcılar arasındaki uyum ya da aynı puanlayıcının zaman içindeki kendisiyle olan tutarlılığına (puanlayıcı içi güvenirlik) değinilmemiştir. Bu tür çalışmalar öğretmen/uzman değerlendirmen puanlarını kendi başına oldukça güvenilir ve geçerli olduğu varsayılarak yapılmıştır. Bu durum, bazı bağlamlarda şüpheli bir varsayım olduğundan söz konusu çalışmaların güvenirlik veya geçerlik ya da her ikisi için yapılmış çalışmalar olup olmadığı tartışmalıdır (Topping, 2003; Topping, 2009). Bu çalışmalarda genellikle doğruluk (accuracy), geçerlik ve güvenirlik terimleri birbirleri yerine kullanılmıştır.

Akran değerlendirmenin güvenirliği ve geçerliği üzerine yapılan araştırmalar, çoğunlukla yükseköğretim düzeyinde yapılmıştır. (Falchikov, 2001; Topping, 2003). Çalışmaların çoğunluğu güvenirlik ve geçerlik derecesini yeterli bulmaktadır (Sadler & Good, 2006); bazı çalışmalarda ise farklı sonuçlar raporlanmıştır (Falchikov & Goldfinch, 2000; Topping, 1998). İlköğretim ve ortaöğretim düzeyinde yapılan çalışma sonuçları da yükseköğretimde yapılan çalışmalara benzer sonuçlar vermiştir (Toppings, 2003).

Topping (1998) üniversite öğrencileriyle yapılmış 31 çalışmada öğrencilerin akran değerlendirme sonuçlarıyla öğretim elemanı gibi uzman kişiler tarafından yapılan değerlendirme sonuçlarının uyumunu incelemiş ve bu değerlendirme türünün güvenirlik ve geçerlik açısından yüksek ölçümler verdiği sonucuna varmıştır. Benzer şekilde Falchikov ve Goldfinch (2000) tarafından gerçekleştirilen meta analiz çalışmasında 56 deneysel çalışma incelenmiştir. Tüm çalışmalarda ortalama korelasyon katsayısı 0,69 olarak bulunmuş, ortalama olarak akran ve öğretmen puanları arasında uzlaşma sağlandığına dair kanıt sunulmuştur. İstatistiksel olarak anlamlı olmayan etki değeri de öğretmenlerin puanlamasıyla öğrencilerin puanlaması arasında uyum olduğunu göstermektedir. Bu iki meta analiz çalışmasına göre daha yeni olan bir araştırma (Hafner & Hafner, 2003) akran değerlendirmesinde kullanılan rubriğin güvenirliği ve geçerliği üzerine odaklanmıştır. Üç yıllık bir periyotta yürütülen çalışma, biyoloji bölümüne kayıtlı 107 lisans öğrencisinin katılımıyla gerçekleşmiştir. Çalışma kapsamında toplam 1577 akran-grup sözlü sunum puanlaması yapılmıştır. Araştırma sonucunda öğretim elemanı puanlaması ile öğrencilerin puanlaması arasında mükemmel bir ilişki raporlanmıştır (r=1,0). Genellenebilirlik çalışmasına göre de yıllar boyunca akranlar arası güvenirlik (inter-rater reliability) orta seviyede bulunmuştur. Hafner ve Hafner çalışmanın sonucunda akran değerlendirilmesinin üç yıl süresinde tutarlı bir şekilde kullanıldığı sonucuna varmıştır. Sadler ve Good (2006) ortaöğretim fizik dersinde gerçekleştirdikleri çalışmada öğretmen ile akran puanlanması karşılaştırılmış ve öğrencilerin rubrik kullanarak yaptıkları puanlamalar ile öğretmen tarafından verilen puanlar arasında 0,90'ın üzerinde güçlü bir ilişki bulunmuştur. Fakat öğrencilerin akranlarını değerlendirirken yanlı davranış sergiledikleri saptanmış, en iyi performans gösteren akranlara verilen puanların, öğretmen tarafından verilen puanlardan daha düşük olduğu gözlenmiştir.

Puanlayıcı Kayması

Açık uçlu sınavlar ve sunum gibi karmaşık performans görevleri puanlayıcılar tarafından puanlanırken puanlayıcının kendi yargısı ölçme sonuçlarını etkileyerek ölçümlerin geçerliğini düşürebilmektedir. Alan yazınında bu durum puanlayıcı etkisi ve puanlayıcı hatası gibi farklı terminolojilerle ele alınmaktadır; 'puanlayıcı etkisi', 'puanlayıcı yanlılığı' ya da 'puanlayıcı hatası'. Bu terminolojilerin tam olarak tanımları yapılamadığından terimler birbirlerinden ayrılamamaktadır ve birbirleri yerine yanlış kullanılmaktadır (Myford & Wolfe, 2003). Bu çalışmada, Scullen, Mount ve Goff (2000) tanımından yola çıkarak 'puanlayıcı etkisi' terimi kullanılacaktır. Puanlayıcı etkisi, puanlanan bireyin performans puanında bu bireyin gerçek performasından ziyade puanlayıcıdan kaynaklanan, sistematik farklılığa yol açan geniş bir etki kategorisidir. Performans değerlendirmede ölçme sonuçlarında hataya yol açan en büyük etki puanlayıcının kendisidir (Engelhard,1994; Gabrielson, Gordon & Engelhard, 1995; McNamara, 1996). Puanlayıcıdan kaynaklanan herhangi bir hata ölçmek istediğimiz yapının dışındaki nedenlerden dolayı puanların varyansına yansıyabilir (Messick, 1994).

Geleneksel olarak puanlayıcı hataları puanlayıcı katılığı ve cömertliği, halo etkisi, merkezi eğilim ve ranj sınırlaması olarak dört başlıkta incelenir. Fakat bunların yanı sıra daha az bahsedilen hatalılık (inaccuracy), yanlılık (değişen puanlayıcı katılığı/cömertliği), sıralama, puanlayıcı kayması gibi diğer hata türleri de mevcuttur (Myford & Wolfe, 2003). Yaygın olarak tanımlanan puanlayıcı etkileri çok sayıda çalışmada belgelenmiş ve bunun üzerine bu etkilerin ortadan kaldırılabilmesi için farklı işlemler tartışılmıştır (ör. Braun & Wainer, 1989, Engelhard, 1996, Myford & Wolfe, 2003, 2004). Son zamanlarda uygulaması ve puanlaması zamana yayılmış geniş ölçekli testlerde puanlayıcı davranışlarının ölçme hatasına önemli ölçüde yol açabileceği vurgulanmaktadır (Harik ve diğerleri., 2009).

Araştırmacılar geçmişte puanlayıcı etkisini statik etki (puanlayıcı etkisi her öğrencinin performans puanını tam olarak aynı şekilde etkiler) olarak tanımlarken daha yeni çalışmalarda her bir puanlayıcının davranışının zamanla değişebildiği görülmüştür (Myford & Wolfe, 2009). Puanlayıcı kayması (DRIFT—differential rater functioning over time) bir performansın farklı zamanlarda yapılan puanlamalarında, puanlayıcı davranışlarındaki değişikliklerin meydana gelmesi olarak tanımlanmaktadır (Park, 2011; Wolfe, Moulder, & Myford 2001). Puanlayıcı kaymasını inceleyen araştırmalar çoğunlukla İngilizce yeterlilik sınavı puanlamalarıyla yapılan çalışmalarla karşımıza çıkmaktadır (ör., Englehard & Myford, 2003, Yang, 2010). Bunun yanı sıra similasyon verisi kullanılarak gerçekleştirilmiş çalışmalar da bulunmaktadır (ör., Park, 2011; Wolfe, Moulder, & Myford, 2001). Bu çalışmaların tamamında, zamana bağlı olarak puanlayıcı kayması görülmüştür. (ör., Braun, 1988, Casabianca, Lockwood & McCaffrey, 2015, Congdon & McQueen, 2000, Englehard & Myford, 2003, Harik vd.., 2009, Hoskens & Wilson, 2001, Lunz & Stahl, 1990, Lumley & McNamara, 1995, McQueen & Congdon, 1997, Myford, 1991, Myford & Wolfe, 2009, Wolfe, Moulder & Myford, 2001, Wilson & Case, 2000). Puanlayıcı kaymasının sebebi, zamanla puanlayıcıların tecrübe edinmesi veya yorulmalarından kaynaklanabilir. Puanlayıcı kaymasını engellemek için sürekli puanlayıcı eğitimleri önerilmektedir fakat bu eğitimler tam olarak puanlayıcıların davranışlarındaki değişimleri engelleyememektedir (Congdon & McQueen, 2000; McKinley & Boulet, 2004).

Ölçme sonuçlarıyla bireyleri birbirleriyle kıyaslayabilmek için madde kalibrasyonlarının gruptan gruba ve zamandan zamana sabit kalması gerekmektedir (Wright ve Masters, 1982). Benzer bir şekilde birden fazla puanlayıcının kullanıldığı puanlamalarda puanlayıcı kalibrasyonunun puanlanan bireyden bireye ve zamandan zamana değişmemesi beklenir. Ölçmede puanlayıcı kayması üç ayrı şekilde gözlenebilir; 1) zamanla puanlayıcı daha katı ya da daha cömert davranış sergileyebilir (differential severity), 2) puanlayıcının puanlamada yaptığı hata miktarı farklı zamanlarda farklı olabilir (differential accuracy) ve 3) puanlayıcı zamanla ölçekteki kategori kullanımda farklı eğilimler gösterebilir (differential category use). Alan yazınında puanlayıcı kayması çalışan araştırmacıların büyük çoğunluğu ilk kayma türünü incelemiştir. Bunlardan bazıları kronolojik sıraya göre aşağıda özetlenmiştir.

Lunz ve Stahl (1990) çok yüzeyli Rasch ölçme modeli (ÇYRÖM) kullanarak üç farklı sınavda (İngiliz Edebiyatı kompozisyon sınavı, klinik sınavı ve sağlık meslek sözlü sınavı) puanlayıcıların katılık/cömertlik seviyelerinin üç ya da dört gün süren puanlamalarda sabit kalıp kalmadığını incelemişler ve iki sınav türünde (kompozisyon ve klinik) puanlayıcıların katılık düzeylerinde belirgin farklılaşmalar olduğunu bulmuşlardır. Myford (1991) da ÇYRÖM kullanarak gerçekleştirdiği çalışmasında drama performanslarını değerlendiren farklı tecrübeye sahip hakemlerin bir ay süresince yaptıkları puanlamalarda açık bir puanlayıcı katılığı kayması olduğunu bulmuştur. Bir başka çalışmada Lumley and McNamara (1995) puanlaması 20 ay süren İngilizce konuşma testinde puanlayıcı davranışlarını incelemişlerdir. Puanlayıcı ana terimi ve puanlayıcı-zaman etkileşim terimleri için puanlayıcı katılığında önemli değişiklikler bulmuşlardır. Wilson ve Case (1997) ise sekizinci sınıf matematik sınavı için iki oturumda gerçekleşen puanlamada puanlayıcı katılığı kaymasını araştırmışlardır. Puanlayıcıların katılık seviyelerinde bir zaman diliminden diğerine istatiksel olarak anlamlı kaymalar olduğunu bulmuşlardır. Congdon ve McQueen (2000), yedi iş gününe (arada bir hafta sonu olmak üzere) yayılan ilköğretimde yazma performanslarının bütüncül rubrik kullanılarak yapılan değerlendirilmelerinde puanlayıcı kaymasını

araştırmışlardır. Araştırmacılar ÇYRÖM kullanarak her bir gün için 16 puanlayıcının göreceli katılık kestirimlerini hesaplamışlardır. Analiz sonuçları, puanlayıcı katılığının günden güne çoğu puanlayıcı için değiştiğini fakat bu değişimin genel bir deseni olmadığını göstermiştir.

Yukarıdaki çalışmalardan farklı olarak Wolfe, Moulder ve Myford (2001) simülasyon verisi kullanarak puanlayıcı katılığı kaymasını çalışmışlardır. ÇYRÖM kullanan araştırmacılar, çeşitli puanlayıcı kayması türlerini tespit etmişlerdir. Fakat Harik ve diğerleri (2009) bu çalışmanın sonuçlarının sadece simülasyon verisine dayalı olduğu için genellenebilirliğinin sorgulanması gerektiğini, çalışmadaki simüle edilmiş koşulların, her bir denemede farklı bir kayma türünü temsil ettiğini, oysa gerçek verilerde, zamana bağlı puanlayıcı kayma türlerinin farklı kombinasyonlarda ortaya çıkabileceği belirtmişlerdir.

McLaughlin, Ainslie , Coderre, Wright ve Violato (2009) 10-12 dakikalık istasyonlarda farklı zaman dilimlerinde yapılan tıp sınavında puanlayıcı kaymasını incelemişlerdir. Zamanla puanlayıcıların daha katı puanlama davranışı gösterdiklerini bulmuşlar ve bunun sebebini yorgunluk olarak tanımlamışlardır. Puanlayıcı kayması gibi sistematik yanlılığın, testin geçerliğini tehlikeye attığını belirtmişlerdir. Az da olsa çalışmaların bir kısmında da genellenebilirlik kuramı kullanılmıştır. G kuramı kullanılarak gerçekleştirilen çalışmaların birinde Casabianca, Lockwood ve McCaffrey (2015) puanlayıcı kaymasını çalışmışlardır. Çalışmalarında 458 matematik ve İngilizce ortaöğretim öğretmeninin öğrencileri tarafından değerlendirilmelerinde puanlayıcı davranışlarının zaman içinde değişip değişmediğini araştırmışlardır. Toplanan veriye göre eğitim kalitesindeki değişimin çok küçük olduğunu fakat gözlemlerin başında puanlayıcı kaymasının çok büyük olduğunu ve bu kaymanın iki yıl boyunca devam ettiğini raporlamışlardır.

Araştırmanın Önemi ve Amacı

Giriş kısmında alan yazını incelemesinde ele alındığı gibi akran değerlendirmesiyle elde edilen puanların geçerliliğine dair kanıtlar sınırlıdır. Bu nedenle psikometrik ağırlığı olan çalışmaların yapılması oldukça gereklidir. Bu çalışmada olduğu gibi özellikle akran değerlendirmesi sürece yayılıyorsa, performans puanlamada puanlayıcı etkisinin görülme olasılığı daha da artmaktadır. Üniversite öğrencileriyle yapılan bu çalışmada dört ayrı güne yayılan dereceli puanlama anahtarı kullanılarak sunum performansları puanlanma sürecinde, akran puanlayıcıların davranışları ve ölçme aracının niteliği araştırılmıştır ve özel olarak aşağıdaki iki ana araştırma sorusuna cevap aranmıştır.

1. Dört değerlendirme günü boyunca göreceli katılık seviyesinde ilk puanlama gününe göre değişen herhangi bir puanlayıcı var mıdır? Bir başka deyişle değişen katılık ya da değişen cömertlik gösteren puanlayıcı var mıdır?

2. Değişen katılığı/esnekliği tespit etmek için kullanılan farklı yaklaşımlar benzer sonuçlar veriyor mu? Farklı yaklaşımlar değişen katılık/esnekliğe sahip benzer puanlayıcıları ortaya çıkarıyor mu?

YÖNTEM

Bu çalışma, akran değerlendirmesinde puanlayıcı katılığında zamana bağlı kaymayı inceleyen betimsel bir çalışmadır. Dereceli puanlama anahtarı kullanılarak dört ayrı günde puanlanan öğrenci sunum performanslarında akran puanlayıcı katılığı kayması çok yüzeyli Rasch ölçme modeli (ÇYRÖM) (Linecra, 1989) yardımıyla incelenmiştir.

Çalışma Grubu

Bu çalışma orta büyüklükte olan bir üniversitenin eğitim fakültesinde 2017 Bahar döneminde verilen eğitimdeki ölçme ve değerlendirme dersinde gerçekleştirilmiştir. Öğrenciler dersin öğretim elemanının danışmanlığında belirlenen PISA, test yanlılığı gibi eğitimde ölçme ve değerlendirme konularını kapsayan kendi hazırladıkları performans görevini dönem sonunda sınıf arkadaşlarına sözlü olarak sunmuşlardır. Grup çalışması olarak gerçekleştirilen bu çalışmada grup büyüklükleri iki ila dört kişi arasında değişiklik göstermektedir. Sunumlar yaklaşık 45 dakika sürmüştür. Sunumun

sonunda sunumu yapan grubun sunum performansı, sınıftaki akranları tarafından dereceli puanlama anahtarıyla değerlendirilmiştir. Derse devam sağlayan 29 öğrenci bu akran değerlendirme çalışmasına katılmıştır. Fakat değerlendirmenin yapıldığı her bir günde sınıfta tam katılım sağlanamamıştır. Dolayısıyla bu durum az da olsa veri kaybına yol açmıştır. Puanlama yapılmadan önce sınıfta puanlama anahtarı üzerinden gidilmiştir ve açıklama yapılmıştır. Fakat olması gerektiği gibi gerçek bir sunum üzerinden öğrenciler anahtar kullanımı üzerine eğitilmemiştir.

İlk üç gün iki sunum, dördüncü gün üç sunum olmak üzere toplam dokuz sunum dört ayrı günde gerçekleştirilmiştir. Her bir sunum puanlamasına katılan akran sayısı 19 ila 25 arasında değişmektedir. Her sunum grubu, sunumu yapan grup ve derse katılım sağlamamış öğrenciler dışında her akran tarafından puanlanmıştır. Fakat sunum performansları günlerin içinde yuvalanmış olduğu için, örneğin ilk iki sunum ilk gün, sonraki iki sunum ikinci gün gibi, günler arasında bağlantı yoktur. Her sunum, bir puanlayıcı tarafından bir kere sunumun yapıldığı gün puanlanmıştır. Bu nedenle ÇYRÖM'ye gün değişkeni kukla değişken olarak eklenebilmektedir.

Veri Toplama Aracı

Öğrenci sunumlarının akranlar tarafından değerlendirilmesi için araştırmacı, dereceli puanlama anahtarı geliştirmiştir. Bu puanlama anahtarı son beş yıldır aynı derste kullanılmaktır. Geçmiş yıllarda derse kayıtlı ve akran değerlendirmesini kullanan öğrencilerin geri bildirimleriyle son halini almıştır. Anahtarda toplam 10 ölçüt (madde) bulunmaktadır ve her bir ölçüt üç dereceli ölçekle puanlanmıştır. Anahtar örnek maddeler şunlardır: "Sunum sonunda konuyu anladım," "Örnekler vererek fikirlerini açığa kavuşturdu," "Sunum ilginçti[1]."

Ölçme Modeli

Bu çalışmada toplanan verilerin analizinde ÇYRÖM kullanılmış ve analizler FACET v3.71.4 (1987-2014) yazılımıyla gerçekleştirilmiştir. Tüm terimlerin ve öğelerin kestirimleri ortak bir metrikte (logit) ve yaygın olarak kullanılan çeşitli uyum istatistikleri bu yazılım yardımıyla hesaplanmıştır. ÇYRÖM, Rasch ailesinin üyesi olduğu için, temel Rasch modelinin tüm özelliklerine sahip olmakla birlikte daha kullanışlı bir modeldir. ÇYRÖM, birden fazla puanlayıcı olan performans ölçümlerinde çeşitli değişkenlik kaynaklarının incelenmesini sağlar. Örneğin değerlendirilen bireyin becerisi, görev zorluğu, puanlayıcı katılığı ve bu değişkenlerin birbirleriyle olan etkileşimi. Bu çalışmada değişen katılık/esneklik ölçüsü olarak ÇYRÖM kullanılarak hesaplanan iki indeks kullanılmıştır – ***standartlaştırılmış farklar*** ve ***etkileşim terimi***.

Standartlaştırılmış Farklar

Her bir gün için puanlayıcı katılık/cömertlik kestirimleri standart puana çevrilerek puanlayıcı kayması araştırılmıştır. Kestirimlerini elde etmek için üç yüzeyli ÇYRÖM (Denklem 1) kullanılmıştır; yüzeyler sunum grubu, puanlayıcı ve ölçüttür. Tipik olarak böyle bir modelde β_n terimi, ortalaması sıfır olacak şekilde eklenir. Fakat bu tür modelleme, dört gün için ortalaması sıfır olmayan (non-centered) dört ayrı puanlayıcı katılığı kestirimini sağlar. Bu kestirimler her gün değerlendirilen sunum performanslarının nitelik seviyelerine göre değişebilir. Farz edelim ki birinci gün puanlanan sunum, ikinci gün puanlanan sunuma kıyasla niteliği daha yüksektir. Her bir gün için sunum niteliğinin kestirimlerinin ortalaması 0,00 logite sabitlendiğinden puanlayıcı kestirimlerinde, ikinci gün kestirimleri ilk güne göre daha düşük olmasına yol açacaktır. Bir başka değişle puanlayıcılar ilk gün daha cömert gibi algılanacaktır. Bu etkiyi ortadan kaldırmak için, her bir kalibrasyonda puanlayıcı kestirimlerinin ortalaması 0,00 logit olarak alınmalıdır. Bu düzenleme,

[1]Çalışmanın yapıldığı kurumun öğretim dili İngilizce olduğundan maddelerin orijinal dili İngilizcedir. Bu makalede çevirisi kullanılmıştır.

farklı zamanlarda gerçekleştirilen puanlamalarda puanlayıcıların *göreceli katılık seviyelerindeki* değişimi takip etmemizi sağlayacaktır.

$$\ln\left(\frac{\text{Pnrik}}{\text{Pnri(k-1)}}\right) = \beta_n - \gamma_r - \delta_i - \tau_k \qquad \text{(1, Ayrı Model (Seperate Model))}$$

P_{nrik}, = puanlanan sunum n'nin 'i' ölçütünde gösterdiği performansın 'r' puanlayıcı tarafından 'k' kategorisinde puanlanma olasılığı

$P_{nri(k-1)}$ = puanlanan sunum 'n'nin 'i'ölçütünde gösterdiği performansın 'r' puanlayıcısı tarafından 'k-1' kategorisinde puanlanma olasılığı

β_n = Sunum n'nin niteliği

γ_r = Puanlayıcı r'nin katılığı

δ_i = Ölçüt i'nin güçlüğü

Gün 1 temel zaman alınarak sırasıyla diğer günler için sapmalar (SAI_{rc}) hesaplanır.

$$SAI_{rc} = S_{rc} - S_{rb}, \quad (2)$$

SAI_{rc} = Zaman c'yi temel zamanla (baseline) karşılaştıran puanlayıcı SAI (Signed Area Index) indeksi

c = karşılaştırılan zaman

b = temel zaman

S_{rj} = Okuyucu r'nin zaman j'de katılık ölçüsü.

Raju (1988, 1990 Wolfe, Myford, Engelhard and Manalo'da alıntılandığı gibi, 2007) farkların nasıl standardize edileceğini aşağıdaki gibi açıklamıştır.

$$Z_{SAIrc} = \frac{SAI_{rc}}{\sqrt{SE^2_{Src} + SE^2_{Srb}}} \quad (3)$$

Z_{SAIrc}= Okuyu *r* için standartlaştırılmış fark indeksi, temel katılık ölçümüne kıyasla zaman c'deki katılığı

SE^2_{Srj} = *j* zamanda okuyucu *r*'nin katılık kestirimindeki standart hatası

Okucunun temel zamandaki katılık seviyesiyle c zamanındaki katılık seviyesi arasında fark olmadığını belirten sıfır (null) hipotezini test etmek için yukarıdaki formülle hesaplanan z puanı standart normal dağılımla karşılaştırılır. Bununla birlikte SAI değeri bir etki değeri ölçütü olarak da kullanılabilir; 0,50'den büyük değerler anlamlı bir farkın olduğunu göstermektedir (Draba, 1977; Swaminathan & Rogers, 1990). Bu çalışmada pozitif Z_{SAIrc} değerleri, puanlayıcının zamanla daha katılaştığını gösteriken negatif Z_{SAIrc} değeri puanlayıcının zamanla daha az katı hale geldiğini göstermektedir.

Etkileşim Terimi

İkinci değişen katılık/esneklik ölçüsü olarak Denklem 4'de verilen zaman kukla etkileşim modelinden elde edilen etkileşim indeksi (I_{rc}) kullanılmıştır (Linacre, M. kişisel iletişim, Haziran 2017; Wolfe ve diğerleri, 2007). Bu modelde zaman terimi kukla değişken olarak eklenmiştir.

$$\ln\left(\frac{\text{Pnritk}}{\text{Pnrit(k-1)}}\right) = \beta_n - \gamma_r - \delta_i - \pi_t - \tau_k \qquad \text{[4, Zaman Kukla Etkileşim Facet Modeli (Time Dummy Interaction Facet Model)]}$$

P_{nritk}, = puanlanan sunum *n*'nin 'i' ölçütünde gösterdiği performansın 'r' puanlayıcı tarafından 't' zamanında, 'k' kategorisinde puanlanma olasılığı

$P_{nrit(k-1)}$ = puanlanan sunum 'n'in 'i'ölçütünde gösterdiği performansın 'r' puanlayıcısı tarafından 't' zamanında, 'k-1' kategorisinde puanlanma olasılığı

β_n = Sunum n'nin niteliği

γ_r = Puanlayıcı r'nin katılığı

δ_i = Ölçüt i'nin güçlüğü

π_t= t zamanında gözlenen performans azalması

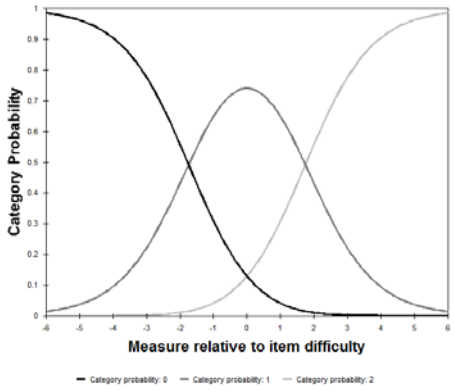
τ_k = Kategori 'k-1'den kategori k'ya geçiş güçlüğü

I_{rc}, standart hataya bölünerek Welch t-test'de kullanılan istatistik hesaplanmış ve bu istatistik I_{rc}'nin sıfırdan farklı olup olmaadığını test etmek için kullanılmıştır.

BULGULAR

Ön Analizler

Bu çalışmada kullanılacak verinin ÇYRÖM modeli ile uyumu ve ölçme aracının tek boyutlu olması beklenmektedir (Bond & Fox, 2015; Eckes, 2011). Model-veri uyumunu değerlendirmek için standartlaştırılmış artık değerler (standardized residuals) kullanılmıştır. Veri-model uyumunun kabul edilebilmesi için bu değerlerin en fazla %1'nin mutlak değerce üçe eşit ya da daha büyük olması ve gene en fazla %3'nün mutlak değerce ikiye eşit ya da ikiden büyük olması gerekmektedir (Linacre, 2002). Analizde kullanılan 1964 veriye ait 20 (%1,02) standartlaştırılmış artık değer |3|'ten büyükken 103 (%5.24) değer ise |2|'den büyüktür. Rasch Modelini kullanabilmek için diğer gereklilik ise kullanılan ölçme aracının tek boyutlu olmasıdır. Ölçekte bulunan maddelere ait uygunluk içi ve dışı indeksler en düşük 0,8 değerini alırken en yüksek 1,12 değerini almaktadır. Bu indeks değerleri ideal olarak 0,6 ile 1,4 aralığında beklenmektedir (Wright & Linacre, 1994). Maddelere ait uyum indeksleri tek boyutluluğa kanıt sağlarken standartlaştırılmış artık değerler ise model-veri uyumunun kabul edilebilir olduğunu göstermektedir. Verinin uygunluğuna karar verildikten sonra kontrol edilmesi gereken bir diğer nokta ise ölçme aracında kullanılan üç puanlı dereceli puanlama ölçeğinin beklendiği gibi kullanılıp kullanılmadığıdır. Şekil 1'de ölçekteki kategorilerin olasılık eğrilerini birbirlerinden net bir şekilde ayrıldığı, beklenen thresholdların beklenen sırada ve yönde arttığı görülmektedir. Threshold değerleri -1.74 ve 1.74'tür. Her ne kadar öğrenciler ikinci ve üçüncü kategoriyi kullanma eğitimde olsalar da ölçme aracında kullanılan üç puanlı dereceli puanlama ölçeğinin beklendiği gibi kullanıldığı sonucuna varabiliriz.



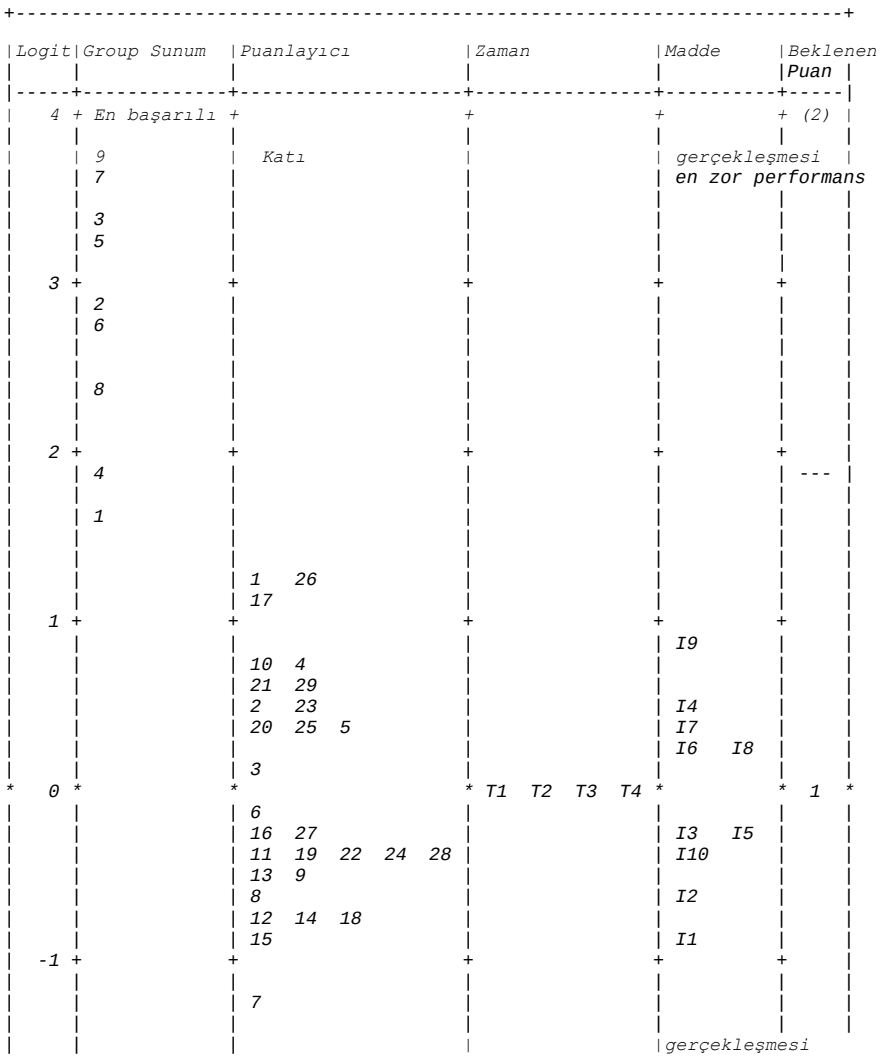
Şekil 1. Kategori Olasılık Eğrisi

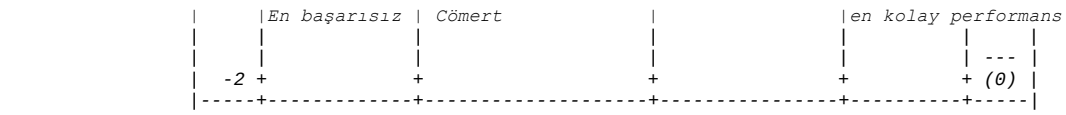
Kukla Zaman FACET Modeli

Bu iki araştırmanın sorularına cevap aranmadan önce kukla zaman Facet modelinin sonuçları bu bölümde verilmiştir. Şekil 2, ÇYRÖM analizinde bulunan farklı yüzeylere ait parametre kestirimlerinin sonuçlarını görsel olarak sunmaktadır. Şekildeki birinci sütun bütün yüzeylere ait kestirimlerin gösterildiği ortak logit ölçüsüdür. İkinci sütunda 9 numaralı grubun sunumunun en başarılı, 1 numaralı grubun sunumunun en başarısız bulunduğu görülmektedir. Sunumların logit değerleri 1,56 ile 3,75 arasında değişiklik göstermektedir (daha fazla bilgi için Ek-Tablo 1). Ayırma indeksi ve güvenirlik katsayısı sırasıyla 4,62 ve 0,96'dır. 'Sunumların öğrenciler tarafından belirlenen nitelikleri arasında bir fark yoktur' hipotezi χ^2 ile test edilmiş ve yokluk hipotezi

reddedilmiştir [$\chi^2_{(8)}$=197.0, p<.000]. Tüm bu istatistikler, Şekil 2'de görüldüğü gibi sunumların performans niteliklerine göre anlamlı olarak birbirlerinden ayrıldığı gösterilmektedir.

Üçüncü sütunda puanlayıcı olarak akran puanlayıcılara ait değerler yer almaktadır. Akran puanlayıcılar katılık seviyelerine göre farklılık göstermektedir. En katı olan puanlayıcının logit değeri 1,30'ken en cömert puanlayıcının puanı -1,27'dir. Ayırma indeksi ve güvenirlik katsayısı sırasıyla 1,88 ve 0,78'dir. Yokluk hipotezi 'Akran puanlayıcıların katılık dereceleri arasında anlamlı bir fark yoktur' hipotezi χ^2 ile test edilmiş ve hipotez reddedilmiştir [χ^2(28)=160,9, p<.000]. Bu sonuçlar öğrencilerin sunumları değerlendirirken birbirlerinden farklı katılık/cömertlik davranışları sergilediğini göstermektedir. Bu yüzey için düşük ayırma indeksi veya güvenirlik katsayısı beklenirken yokluk hipotezinin de reddedilmemesi gerekmektedir. Geçerli bir puanlama için puanlayıcıların benzer davranışlar sergilenmesi beklenmektedir. Bir başka deyişle bir performansın kim tarafından puanlandığından bağımsız olarak aynı ölçütler göz önüne alındığında her puanlamada aynı değer alınmalıdır. Her bir puanlayıcıyı ayrı ayrı incelemek için de uygunluk indeksleri kullanılmaktadır. Bu indeksler değerlendirilirken Linacre and Wright'ın (2002) önerdiği gibi 0,5-1,5 aralığı kullanılmıştır. Buna göre 29 puanlayıcı içinde sadece dokuz numaralı puanlayıcının 'uygunluk dışı' indeks değeri, istenilen sınırların biraz dışındadır (Ek-Tablo 2). Uygunluk içi indekslerin işaret ettiği probleme kıyasla uygunluk dışı indekslerin işaret ettiği problem, ölçme işlemi için daha az ciddi bir problemdir ve dolayısıyla başa çıkılması daha kolaydır (Linacre, 2002). Dokuz numaralı puanlayıcının uygunluk içi indeks değerinde bir problem olmadığından uygunluk dışı indeksinin kabul edilebilir aralığı biraz aşmasında çok önemli bir sakınca bulunmamıştır. Sunum ve puanlayıcı kestirimleri karşılaştırıldığında puanlayıcıların cömert davrandıkları görülmektedir.





Şekil 2. Kalibrasyon Haritası

Puanlamada kullanılan dereceli değerlendirme ölçeğin on maddeden oluşmaktadır. Madde 1, puanlayıcıların kolaylıkla yüksek puan verdikleri, bir başka deyişle sunum sırasında en kolay yerine getirilebilen maddedir. Bunun tam tersi Madde 9 sunumu yapmak gruplar için gerçekleştirilmesi en zor davranışı göstermektedir. Madde logit değerleri -0,84 ile 0,92 arasında değişmektedir. Kestirimde ortalama standart hata 0,17 (SS=0,02)'dir. Puanlamada kullanılan ölçeğin güvenirliğinin (0,90) iyi olduğu söylenebilir. Maddelerin uyum değerleri incelendiğinde hepsinin modele uygun davrandığı, çoğunun değerinin istenen değer olan '1' etrafında olduğu görülmektedir (Ek-Tablo 3). Madde ayırma indeksi 3.01, güvenirlik katsayısı 0,90'dır. Sonuç olarak elde edilen istatistikler, ölçeğin niteliğinin kabul edilebilir olduğunu ve ölçekteki her bir maddenin de beklendiği gibi davrandığını, puanlayıcıların katılık/cömertlik seviyelerinde farklılık olduğunu göstermektedir.

Puanlayıcı Kayması Analizi

Standartlaştırılmış Farklar, Z_{SAIrc}

Puanlayıcı katılığındaki kaymayı belirlemede kullanılan yaklaşım, makalenin yöntem kısmında anlatılmıştır. Bunlardan ilki olan Ayrı Model kullanılarak elde edilen puanlayıcı kestirimine ait betimsel istatistikler aşağıdaki tabloda verilmiştir.

Tablo 1: Ayrı Model Puanlayıcı Katılığı Kestirimleri

	Gün 1	Gün 2	Gün 3	Gün 4	Gün 4*
Ortalama	0,00	0,00	0,00	0,00	0,26
S.S.	0,89	0,99	1,22	1,31	1,44

*3 maximum puanlama yapan puanlayıcı dahil

Ayrı Model kullanılarak elde edilen puanlayıcı kestirimleriyle hesaplanan SAI_{rc} indeksine[2] ait istatistikler Tablo 2'de verilmiştir. SAI_{rc} indeksi yani temel alınan zamandan (Gün 1'den diğer bir güne) bir okuyucunun göreceli katılığındaki değişimi göstermektedir. Toplam 21 puanlayıcıya ait katılık seviyesinde Gün 1'den Gün 2'ye ortalama 0,048 puan düşme gözlenmiştir. Ortalamada puanlayıcıların daha cömert hale geldiği söylenebilir. Puanlayıcı katılık seviyesinde artış olan puanlayıcılarda (%38) gözlenen en yüksek artış 1,48 puanken ikinci günde daha cömert olan puanlayıcılar arasında en fazla kaymayı gösteren puanlayıcı 1,28 logit puan daha cömert olmuştur. Fakat kaymalardaki bu artış ve düşüşlerden hiçbiri istatistiksel olarak anlamlı değildir.

Tablo 2: SAI_{rc} Katılık Seviyesi Karşılaştırması

	Gün 1 vs. 2	Gün 1 vs. 3	Gün 1 vs. 4
SAI_{rc} Ortalama	-0,048	0,150	0,339
SAI_{rc} S.S.	0,744	1,571	1,299
SAI_{rc} En düşük	-1,28	-2,99	-1,49
SAI_{rc} En yüksek	1,48	2,60	3,79
Daha çok katı olanların %	38,10	52,38	42,86
Z_{SAIrc} > |1,96| %	0	38,10	15,00

[2] Ayrı modele dayalı puanlayıcı kestirimleri ve standartlaştırılmış farklar Ek-Tablo 4' de verilmiştir.

Gün 1'den Gün 3'e 21 puanlayıcının göreceli katılık seviyesinde ortalama 0,15 logit puan artış gözlenmiştir. Bunun sebebi, ortalamada puanlayıcıların daha katı puanlama eğilimde olmalarıdır. Göreceli katılık düzeyinde artış olan puanlayıcılarda (%52,38) gözlenen en yuksek artış 2,60 puanken üçüncü gün daha cömert olan puanlayıclar arasında göreceli katılık seviyesinde en fazla düşüş 2,99 logit puandır. Bu katılık düzeyindeki kaymalardan yaklaşık %38'i istatistiksel olarak anlamlıdır. Beş puanlayıcı daha katı puanlama yaparken üç puanlayıcı daha cömert puanlama yapmıştır. Son karşılaştırma da Gün 1 ve Gün 4 arasında yapılmıştır. Sonuçlar bir önceki karşılaştırmaya benzemekle birlikte burada puanlayıcı katılığındaki artış daha yüksektir. Bu katılık düzeyindeki kaymalardan %15'i istatistiksel olarak anlamlıdır. İki puanlayıcı daha katı puanlama yaparken bir puanlayıcı daha cömert puanlama yapmıştır. Her zaman dilimi için ve kaymayı gösteren puanlayıcı logit değerleri ve kaymayı gösteren Z_{SAIrc}değerleri ayrıntılı olarak Ek-Tablo 4'te verilmiştir.

Etkileşim Terimi, I_{rc}

Puanlayıcı kaymasını tespit etmek için kukla değişkenli zaman Facet modeli (Dummy Time-Facet Model) kullanılarak yapılan Rasch analizi, puanlayı ve gün yüzeyi arasında etkileşim olduğunu göstermektedir [Chi-square (99)=146.6, p<.00]. Bu analizden elde edilen etkileşim terimi (I_{rc}) indeksin istatistikleri Tablo 3'de verilmiştir[3]. Tabloda ilk sütun ilgili istatistiğin açıklamasını vermektedir. Son üç sütun ise ikinci, üçüncü ve dördündü gün için elde edilen kestirimlerin temel alınan birinci gün ile karşılaştırılmalarıyla elde edilen indekslerin istatistiklerini vermektedir. İkinci sütunda iki logit arasındaki farkı gösteren I_{rc} indeksinin ortalaması -0,058'dir. Bu durum puanlayıcıların gün 1'e kıyasla gün 2'de daha cömert olarak sunumları puanladıklarını gösterir. Bazı puanlayıcılar daha katı olurken diğerleri daha cömert puanlama davranışı göstermiştir. Üçüncü ve dördüncü sütundaki ortalama I_{rc} ise ilk güne kıyasla puanlayıcıların üçüncü ve dördüncü günde daha katı olma eğiliminde olduklarını göstermektedir. Özellikle son günde ortalama katılık seviyesindeki artış yaklaşık 0,3 logittir. Bu indeksin standart sapmasında yaşanan en düşük ve en yüksek değeri sırasıyla ikinci, üçüncü ve dördündü satırda verilmiştir.

Tablo3: Puanlayıcı-Zaman Arasındaki Etkileşime Ait Özet Değerler

	Gün 1 vs. 2	Gün 1 vs. 3	Gün 1 vs. 4
Ortalama I_{rc}	-0,058	0,126	0,289
SS I_{rc}	0,721	1,538	1,221
En düşük I_{rc}	-1,16	-2,88	-1,23
En yüksek I_{rc}	1,41	2,53	3,44
Anlamlı Welch t-test p<0,05 %	0	38,10	5,00
Etki büyüklüğü |Irc| > 0,50 %	66,7	85,7	65,0

Beşinci sıra, istatistiksel olarak sıfırdan farklı olan olan I_{rc} indeksinin yüzdesini vermektedir. Welch t-test'e göre ilk karşılaştırmada (ikinci sütün) katılık seviyesinde gözlenen bu kaymalar, istatistiksel olarak anlamlı değildir. Fakat diğer günlerdeki kaymaların küçük bir kısmı istatistiksel olarak anlamlıdır. İlk güne göre üçüncü gün kaymaların % 38,10'u, ve dördüncü gün kaymalarının % 5'i istatistiksel olarak anlamlıdır.

Tablonun en alt sırasında, Zaman 1'e göre her zaman döneminde anlamlı olabilecek kadar büyük katılık düzeyinde bir değişiklik gösteren puanlayıcıların yüzdesi verilmektedir. Tabloda görüldüğü üzere iki gün arasındaki kaymanın 0,50 logitten büyük olan değişimlerinin yüzdesi oldukça

[3] Etkileşim raporu Ek-Şekil 1'de verilmiştir.

yüksektir. Bu durum, kaymaların büyüklüğünün istatistiksel olarak anlamlı olmasa da büyüklük olarak önemli olduğunu gösteriyor. Bu büyüklüklerin istatistiksel olarak anlamlı olamamasının nedeni, ölçmedeki hata payının yeteri kadar küçük olmamasıdır.

SONUÇLAR ve TARTIŞMA

Zaman içerisinde daha da popülerleşen akran değerlendirmesi yöntemi, hem biçimlendirmeye hem de düzey belirlemeye yönelik bir değerlendirme amacıyla kullanılabilen eğitsel bir araçtır. Buna rağmen diğer değerlendirme yöntemlerine göre daha az sayıda akademik çalışmaya konu olmuştur. Bir değerlendirme yöntemi olarak daha az dikkat çekmesinin sebebi, öğretmenlerin ya da öğretim elemanlarının öğrencilerin güvenilir ve nitelikli bir puanlayıcı/değerlendirici olduklarına inanmamaları olabilir. Alan yazınında akran değerlendirmesiyle elde edilen puanların geçerliliği ve güvenirliği üzerine sınırlı sayıda çalışma vardır ve var olan çalışmalarda genelde öğrenci-öğretmen puanlama ilişkisine bakılmıştır. Yüksek korelasyon değerleri, akran değerlendirmesiyle elde edilen puanların güvenirliği ve/veya geçerliği olarak kabul edilmiştir. Fakat tam olarak geçerliği belli olmayan öğretmen puanlamasını ölçüt olarak kabul edilen bir yöntemle geçerlik kanıtı sağlandığının iddia edilmesi soru işareti oluşturmaktadır.

Üniversite öğrencileriyle yapılan ve dört güne yayılan bu akran değerlendirme çalışmasında dereceli puanlama anahtarı kullanılarak puanlanan sunum performansları sürecinde genel olarak puanlayıcıların davranışları incelenmiş ve özel olarak puanlayıcı etkisi çeşitlerinden olan puanlayıcı katılığı kaymasının olup olmadığı ÇYRÖM analiziyle araştırılmıştır. Bu çalışmada ölçme sürecinin bir öğesi olarak akran puanlayıcı, puanlanan grup, puanlama için kullanılan dereceli puanlama anahtarı ayrı ayrı incelenmiştir. Analizler genel olarak akran puanlayıcıların arkadaşlarını oldukça cömert bir biçimde puanladıklarını göstermiştir. Puanlayıcılar kendi aralarında kıyaslandığında ise katılık/cömetlik seviyelerinin birbirlerinden farklı oldukları bulunmuştur. Puanlayıcılar katılık seviyelerine göre ± 1 logit ranjında dağılım göstermektedirler. Puanlayıcılar dereceli puanlama anahtarını kullanarak 9 sunumu niteliklerine göre birbirlerinden ayırt edebilmektedirler. Puanlayıcıların uyum indekslerine bakıldığında çok az sayıda puanlayıcının modele göre beklenenden farklı davranış sergilediği görülmektedir. Uyum istatistiklerine bakarak sunumların puanlayıcılar tarafından tutarlı bir şekilde niteliklerine göre sıralandıkları sonucuna varılabilir. Güvenirlik katsayısı 0,90 olması ve uyum indekslerinin istenen aralıkta olması, kullanılan dereceli puanlama anahtarının niteliğinin iyi olduğunu göstermektedir. Puanlama anahtarındaki madde kestirimlerine bakıldığında öğrenci performanslarında yerine getirilebilen üç ölçüt, zorluk sırasına göre şu şekilde sıralanmıştır; Madde 9 "sunumu yapan kişi göz temasını ve yüz ifadesini iyi kullandı," Madde 4 "Sunum ilgi çekiciydi" ve Madde 7 "Sunum akıcıydı". Bir başka deyişle, akran puanlayıcıların sunumlarda en zayıf buldukları ölçütler bunlardır. Sunumların en başarılı bulundukları ölçütler ise sırasıyla Madde 1 "Sunum detaylıydı," Madde 2 "Sunum sonunda konuyu anladım," Madde 10 "Slaytlar kalabalık değildi, takibi kolaydı".

Bu çalışmada puanlamanın ilk günü temel alınarak daha sonraki 3 puanlama gününde ilk güne göre puanlayıcı katılık/cömertlik seviyesinde bir değişik olup olmadığı incelenmiştir. Puanlayıcı katılığı kayması iki farklı yöntemle araştırılmıştır: standartlaştırılmış farklar (Z_{SAIrc}) ve etkileşim terimi (I_{rc}). İki yöntemle de oldukça benzer sonuçlara ulaşılmıştır. Gün-1 ve 2 arasında puanlayıcıların kestirimlerinde bir farklılık görülmemektedir. Her ne kadar ortalamada puanlayıcılar daha cömert puanlama yapsa da kaymalar istatistiksel olarak anlamlı değildir. Fakat kaymalardaki logit cinsindeki farklara bakıldığında etki büyüklüğü olarak kaymaların %60'tan fazlasında görülmesi beklenen minumum farktan daha büyük kayma gösterdiği tespit edilmiştir. Bu kadar kayma, daha fazla puanlamanın yapıldığı ve önemli kararların alındığı sınav sonuçlarında önemli sorunlara yol açabilir. Gün-1 ve 3 arasında puanlayıcıların kestirimlerinde önemli kaymaların olduğu puanlayıcıların oranı %38,10'dur. İki yönteme göre puanlayıcılar ortalamada yaklaşık 0,14 logit kayma gösterip daha katı puanlama davranışı sergilemişlerdir. Puanlayıcılara ayrı ayrı bakıldığında beş puanlayıcının daha katı olduğu, üç puanlayıcının ise daha cömert olduğu görülmektedir. Gün-1 ve 4 arasında puanlayıcıların kestirimlerinde önemli kaymaların olduğu puanlayıcıların sayısı standartlaştırılmış

farklar yöntemiyle üçgen, etkileşim terimi yöntemiyle birdir. Ortalamada iki yöntemle de puanlayıcılar daha katılaşmıştır. Ortalamada kaymanın en yüksek olduğu Gün-4'dür. Gün-3 ise kayma gösteren en yüksek sayıda puanlayıcının olduğu gündür.

Alan yazınında bu çalışmanın kapsamında yapılmış çalışma az olduğu için sonuçları karşılaştırabileceğimiz bulgular çok azdır. Giriş bölümünde bahsedilen meta analiz çalışmaları, çalışmaların büyük çoğunluğunda akran değerlendirmesinden elde edilen puanların geçerli ve güvenilir olduğunu raporlamaktadırlar. Bilgimiz dahilinde Casabianca, Lockwood ve McCaffrey'nin çalışması (2015) akran değerlendirmede puanlayıcı etkisini incelemiş tek çalışmadır ve bu çalışmanın sonuçlarıyla paralel olarak puanlayıcılar arasında katılık/cömertlik seviyeleri bakımından önemli farklılıklar görülmüştür. Bu çalışmada Casabianca ve diğerlerine ek olarak bazı puanlayıcıların birinci günden sonra katılık seviyelerini sınıf arkadaşlarına kıyasla sabit tutamadıkları görülmüştür. Bazıları daha katı davranırken bazıları da daha cömert davranmışlardır. Ortalamaya bakılarak sonuca varıldığında sonraki günlerde sunum yapan gruplar dezavantajlı olmuşlardır. Bu nedenle literatürde bahsedildiği gibi (ör. Congdon & McQueen, 2000 ve McKinley & Boulet, 2004) puanlama sürecinde puanlayıcı eğitimleri, puanlayıcı kaymasını tam olarak ortadan kaldırmasa da gereklidir.

KAYNAKÇA

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement. Analyzing and evaluating rater-mediated assessments*. Frankfurt am Main: Peter Lang.

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics, 13*(1), 1–18.

Braun, H. I., & Wainer, H. (1989). Making essay test scores fairer with statistics. In J. Tanur, F. Mosteller, W. H. Kruskal, E. L. Lehmann, R. F. Link, R. S. Pieters & G. S. Rising (Eds.), *Statistics: A guide to the unknown* (3rd ed., pp. 178–188). Pacific Grove, CA: Wadsworth.

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement, 75*(2), 311–337. doi: 10.1177/0013164414539163

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*(2), 163-178.

Demirbilek, M. (2015). Social media and peer feedback: What do students really think about using Wiki and Facebook as platforms for peer feedback? *Active Learning in Higher Education, 16*(3) 211–224. doi: 10.1177/1469787415589530

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and coassessment in higher education: A review. *Studies in Higher Education, 24*(3), 331-350.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement 31*(2), 93-112.

Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33*(1), 56–70.

Engelhard, G., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition Program with a many-faceted Rasch model* (Research Rep. 03-01). Princeton, NJ: Educational Testing Service

Falchikov, N. (1995) Peer feedback marking: Developing peer assessment. *Innovations in Education and Training International, 32*(2), 175-187.

Falchikov, N. (2001). *Learning together: Peer tutoring in higher education*. London: Routledge Falmer.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287–322.

Gabrielson, S., Gordon, B., & Engelhard, G. (1995). The effects of task choice on the quality of writing obtained in a statewide assessment. *Applied Measurement in Education 8*(4), 273-290.

Heyman J. E., & Sailors J. J. (2011). Peer assessment of class participation: Applying peer nomination to overcome rating inflation. *Assessment & Evaluation in Higher Education, 36*(5), 605-618. doi: 10.1080/02602931003632365

Hafner, J. C., & Hafner, P. M. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education, 25*(12), 1509–1528. doi: 10.1080/0950069022000038268

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement, 46*(1), 43-58. doi: 10.1111/j.1745-3984.2009.01068.x

Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed response items: An example from the Golden State Examination. *Journal of Educational Measurement, 38*(2), 121–146.

Kane, J. S., & Lawler, E. E (1978). Methods of peer assessment. *Psychological Bulletin, 85,* 555-586.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean?. *Rasch Measurement Transaction, 16*(2), 878.

Love, K. G. (1981). Comparison of peer assessment methods: Reliability, validity, friendship bias, and user reaction. *Journal of Applied Psychology, 66*(4),451-457.

Lumley, T., & McNamara, T. E (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71.

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions, 13,* 425-444.

McLaughlin, K., Ainslie, M., Coderre, S., Wright, B., & Violato, C. (2009). The effect of differential rater function over time (DRIFT) on objective structured clinical examination ratings. Medical Education, 43, 989–992. doi:10.1111/j.1365-2923.2009.03438.x

McNamara, T. F. (1996). *Measuring second language performance*. Harlow, UK: Addison Wesley Longman Limited.

McQueen, J., & Congdon, P. J. (April, 1997). *Rater severity in large-scale assessment: Is it invariant?* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Messick, S. (1994). Alternatıve modes of assessment, uniform standards of validity. ETS Research Report Series, 2,1-22.

Myford, C. M. (1991). *Judging acting ability: The transition from notice to expert*. Paper presented at the American Educational Research Association, Chicago IL.

Myford, C. M., & Wolfe, E. M. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Aplied Measurement, 4*(4), 386-422.

Myford, C. M., & Wolfe, E. M. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227.

Myford, C. M., & Wolfe, E. M. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale use. *Journal of Educational Measurement, 46*(4), 371-389. doi: 10.1111/j.1745-3984.2009.00088.x

Park, Y. S. (2011). *Rater drift in constructed response scoring via latent class signal detection theory and item response theory* (Unpublished doctoral dissertation). University of Columbia, NY.

Rowan, B., Harrison, D. M., & Hayes, A. (2004). Using instructional logs to study elementary school mathematics: A close look at curriculum and teaching in the early grades. *Elementary School Journal, 105,* 103-127.

Sadler, P. M., & Good, E. (2006). The impact of self and peer-grading on student learning. *Educational Assessment, 11,* 1–31. doi:10.4103/2229-516X.186961

Scruggs, T. E., & Mastropieri, M. A. (1998). Tutoring and students with special needs. In K. J. Topping & S. Ehly (Eds.), *Peer-assisted learning* (pp. 165–182). Mahwah, NJ: Lawrence Erlbaum Associates.

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*(6), 956-970.

Somervell, H. (1993). Issues in assessment, enterprise and higher education: The case for self-, peer and collaborative assessment. *Assessment & Evaluation in Higher Education, 18*(3), 221-233.

Topping, K. J. (1998). Peer assessment between students in college and university. *Review of Educational Research, 68*(3), 249–276.

Topping, K. (2003). *Self and peer assessment in school and university: Reliability, validity and utility. in optimising new modes of assessment: In search of qualities and standards*. M. S. Segers, Dochy,R., and E. C. Cascallar (Ed.). Netherlands.

Topping, K. J. (2005). Trends in peer learning. *Educational Psychology, 25*, 631–645. doi: 10.1080/01443410500345172

Topping, K. J. (2009). Peer assessment. *Theory Into Practice, 48*(1), 20-27. doi: 10.1080/00405840802577569

Topping, K. J., & Ehly, S. (Eds.). (1998). *Peer assisted learning*. Mahwah, NJ: Lawrence Erlbaum Associates

Tseng, S. C., & Tsai, C.-C. (2007). On-line peer assessment and the role of the peer feedback: A study of high school computer course. *Computers & Education, 49*(4), 1161-1174. https://doi.org/10.1016/j.compedu.2006.01.007

Weaver II, R., & Cotrell, H. W. (1986). Peer evaluation: A case study. *Innovative Higher Education, 11*(1), 25-39.

Wilson, M., & Case, H. (2000). An examination of variation in rater severity over time: A study of rater drift. *Objective measurement: Theory into practice, 5*, 113-134.

Wolfe, E. W., Moulder, B. C., & Myford, C. M. (2001). Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model. *Journal of Applied Measurement, 2*(3), 256–80.

Wolfe, E. W., Myford, C. M., Engelhard Jr. G., & Manalo, J. R. (2007). *Monitoring reader performance and DRIFT in the AP® English literature and composition examination using benchmark essays* (Research Report No. 2007-2). Retrieved from https://research.collegeboard.org/publications/content/2012/05/monitoring-reader-performance-and-drift-ap-english-literature-and

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press

Yang, R. (2010). *A many-facet Rasch analysis of rater effects on an oral English Proficiency Test* (Unpublished doctoral dissertation).Purdue University. IN.

Yang, Y. & Tsai, C. C. (2010). Conceptions of and approaches to learning through online peer assessment. *Learning and Instruction, 20*(1), 72-83. doi: 10.1016/j.learninstruc.2009.01.003

EXTENDED ABSTRACT

Introduction

As in each measurement process, the validity of the scores obtained from the peer assessment needs to be examined. Reliability and validity in the peer assessment literature are generally regarded as the agreement between students' scores and experts' score such as teachers score, not as agreement among peer raters or the consistency of the same peer raters across different time points. Such studies are based on the assumption that teacher scoring are highly reliable and valid. Since this is a suspicious assumption in some contexts, it should be debated whether these studies provide evidence for reliability or validity, or both (Topping, 2003; Topping, 2009). In these studies, the terms accuracy, validity and reliability are often used interchangeably.

Research on the reliability and validity of peer assessment is often included in higher education studies (Falchikov, 2001, Topping, 2003). The majority of the studies have reported high validity (Sadler & Good, 2006); some studies have reported different results (Falchikov & Goldfinch, 2000; Topping, 1998). The results of the studies conducted in the school settings have provided similar results to the studies conducted in the higher education (Toppings, 2003).

In this study, the behavior of peer raters in scoring presentation performance by using the rating scale during four different days will be examined and specifically the following two main research questions will be sought.

1. Are there any rater whose relative severity level drift during four days? In other words, are there any peer rater exhibiting differential severity?

2. Do the different approaches used to detect drift provide similar results?

Method

Participants

The data of this study were collected at the measurement and evaluation course offered in the college of education at the medium sized university in the spring semester of 2017. At the end of the presentation, the presentation was assessed by their peers. Twenty nine students attending the course participated in this study; however, on the day of each assessment, there was no full participation in the classroom. This has led to some data loss.

The total of nine presentations were held on four separate days, two presentations on each of the first three days and three presentations on the fourth day. The number of peer raters in each presentation scoring ranged from 19 to 25. Each presentation performance was rated by each peer, except for the

presenter group and the peers who did not attend the class on the particular day. However, since presentation performances are nested within days, there is no link between days. The first two presentations are nested within the first day, and the next two presentations are nested within the second day and so on. Each presentation was rated on the day the presentation had been made. For this reason, time (day) facet can be added the Many-Facet Rasch Measurement model as a dummy variable.

Data Collection Tool

The researcher of this study developed a scoring key for peer raters. This scoring key was developed to be used in the same course and had been finalized with the contribution of the students who were taking the same course one semester earlier. There were a total of 10 criteria (items) in the key, and each criterion was scored on a three-point rating scale. Sample items were: I understand the topic as a result; Elaborated upon ideas by giving examples/reasons, explanations; Presentation was engaging.

Measurement Model

A many-Facet Rasch Measurement (MFRM) (Linacre, 1989) model was used for analyses and they were performed by FACET v3.71.4 (1987-2014) software. In order to examine rater drift, two indexes were calculated using MFRM - *standardized differences* and *interaction term.*

Three Facet Rasch model (facets are presenter group, rater and the rating scale) was used to estimate rater severity for each day separately. Average of rater estimates was set to 0.00 logit. This arrangement allows us to follow the drift in the relative severity/leniency levels of peer rater at different times. Rater drift was calculated between baseline day (Day-1) and other days (Day-2, 3 and 4). The difference (SAI) between two days relative severity estimates were standardized (Z_{SAIrc}). Z values were used to test the null hypothesis that there is no difference between the level of severity at baseline day and the level of severity at time c.

The second index for rater drift, the interaction index (I_{rc}) obtained from the time dummy model was used (Linacre, M. personal communication, June 2017, Wolfe et al., 2007). In this Rasch model time was added as a dummy variable. Interaction between time and rater facet were examined

Results and Discussion

The estimates of the presentations vary between 1.56 and 3.75 logit (See Annex for more information). The separation index and reliability coefficient are 4.62 and .96 respectively [$\chi^2(8)=197.0$, $p <.000$]. All these statistics show that quality of the presentations are significantly different from each other. Peer raters vary according to their level of severity. The logit value of the most severe rater is 1.30, while the logit estimate of the most lenient rater is -1.27. The separation index and reliability coefficient are 1.88 and .78 respectively [χ^2 (28) = 160.9, $p <.0001$]. These results show that peers exhibit different severity behaviors when evaluating presentations. The rating scale used in scoring has nine items. It can be said that the reliability of the scale used in the scoring (0.90) is good. It is seen that the majority of the fit statistics of items are around the desired value '1'. The item separation index is 3.01, and the reliability coefficient is 0.90. As a result, statistics show that the quality of the scale is acceptable and its each item behaves as expected, and that the raters are a different in their severity/lenience.

Relative rater severity drift from Day-1 to 4 was examined. Standardized differences and interaction term provided very similar results. Between Day-1 and 2, there is no statistically significant difference in the estimates of the rater severity. Between Day-1 and 3, the percentage of scorers with significant drift in the estimates of the raters is 38.10%. According to the two methods, the raters show an average of about 0.14 logit shift and display a more severe scoring behavior. While five raters are more severe, three raters are more lenient. Between Day-1 and 4, the number of raters who had significant shifts in their estimates is three according to the standardized difference method,

while one according to interaction method. In the average, the raters became more severe. Among three comparison, Day-4 has the largest rater severity drift on the average; Day-3, however, has the highest number of raters with rater severity drift.

Ekler

Tablo 1: Sunum Performans İstatistikleri

Presentation	FairMAvge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ
1	1.43	1.56	0.14	1.05	0.62	1.04	0.54
2	1.76	2.92	0.16	1.13	1.36	1.08	0.64
3	1.83	3.38	0.17	0.94	-0.50	0.75	-1.71
4	1.50	1.82	0.13	0.95	-0.72	0.97	-0.32
5	1.82	3.26	0.17	1.08	0.80	1.19	1.37
6	1.72	2.74	0.14	1.09	1.04	1.10	0.94
7	1.87	3.67	0.20	0.96	-0.23	0.80	-1.01
8	1.65	2.40	0.16	0.93	-0.70	0.94	-0.49
9	1.88	3.75	0.20	0.89	-0.75	0.74	-1.30
		2.83	.16	1.00	.1	.96	-.1
		.78	.02	.09	.8	.16	1.1
RMSE:0.17	S.D.: 0.76	Ayırma İndeksi:4.62		Güvenirlik:0.96			
χ^2 (8): 197.0	*p*: 0.00						

Tablo 2: Akran Puanlayıcı İstatistikleri

Raters	Fair MAvge	Measure	S.E.	Infit MS	Infit Z	Outfit MS	Outfit Z
9	1.84	-0.56	0.48	1.43	1.35	***1.80***	1.63
6	1.76	-0.09	0.26	1.27	1.62	1.49	2.03
26	1.43	1.26	0.28	1.42	2.25	1.42	2.25
20	1.67	0.35	0.25	1.34	2.15	1.41	2.14
23	1.64	0.46	0.23	1.32	2.27	1.40	2.40
4	1.58	0.71	0.25	1.28	1.89	1.35	2.04
19	1.80	-0.31	0.28	1.04	0.31	1.19	0.81
21	1.61	0.58	0.24	1.04	0.34	1.12	0.80
17	1.46	1.14	0.22	1.08	0.64	1.06	0.49
22	1.80	-0.34	0.24	1.00	0.04	1.05	0.30
2	1.62	0.54	0.26	0.94	-0.37	1.02	0.19
18	1.86	-0.72	0.30	1.05	0.32	0.99	0.05
3	1.73	0.08	0.32	0.94	-0.28	0.94	-0.21
10	1.58	0.71	0.30	0.93	-0.37	0.93	-0.30
16	1.80	-0.30	0.44	0.84	-0.54	0.87	-0.21
25	1.66	0.38	0.23	0.88	-0.90	0.84	-1.00
29	1.61	0.58	0.33	0.88	-0.59	0.82	-0.83
1	1.42	1.30	0.24	0.83	-1.25	0.81	-1.41
5	1.66	0.41	0.35	0.79	-0.97	0.81	-0.71
8	1.84	-0.58	0.35	0.97	-0.05	0.80	-0.54

Tablo 2: Akran Puanlayıcı İstatistikleri – devam ediyor

Raters	Fair MAvge	Measure	S.E.	Infit MS	Infit Z	Outfit MS	Outfit Z
13	1.83	-0.48	0.28	0.88	-0.66	0.76	-0.96
28	1.81	-0.37	0.28	0.87	-0.74	0.73	-1.15
24	1.81	-0.37	0.28	0.81	-1.13	0.71	-1.25
27	1.79	-0.25	0.57	0.83	-0.38	0.70	-0.62
11	1.81	-0.37	0.28	0.78	-1.34	0.66	-1.47
12	1.87	-0.80	0.31	0.86	-0.64	0.61	-1.35
14	1.87	-0.80	0.31	0.81	-0.91	0.58	-1.49
7	1.91	-1.27	0.41	0.84	-0.44	0.54	-1.05
15	1.88	-0.89	0.32	0.77	-1.12	0.50	-1.77
Ortalama	1.70	0.00	0.31	0.99	0.00	0.95	0.00
SS	0.13	0.68	0.08	0.20	1.10	0.32	1.30
RMSE:0,32	S.D.: 0,61	Ayırma İndeksi:1,92		Güvenirlik:0,76			
χ^2 (28): 160,9	*p*: 0.00						

Tablo 3: Madde İstatistikleri

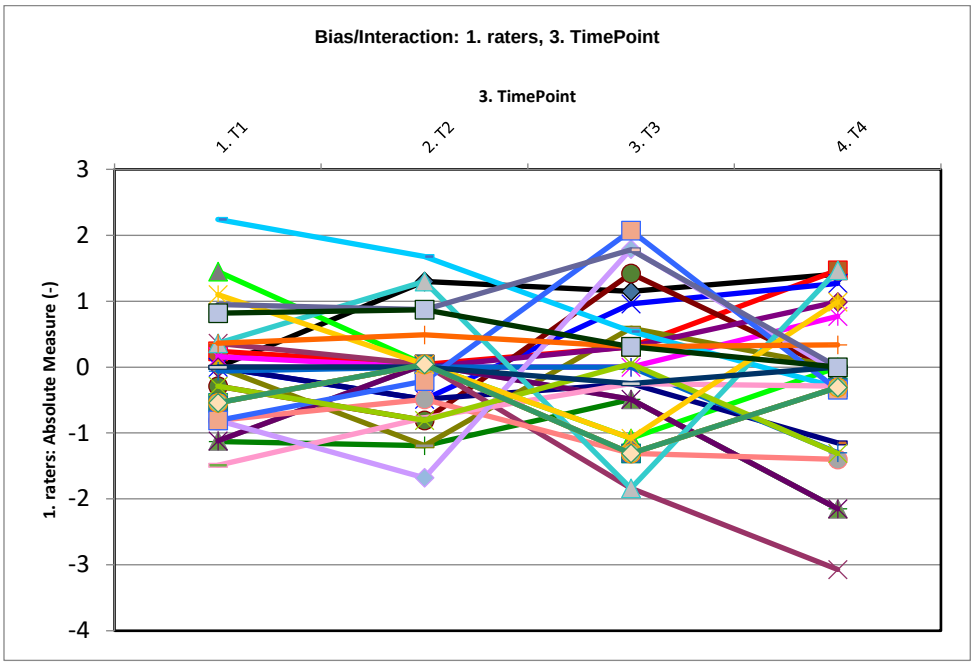
Maddeler	FairMAvge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ
I1	1.87	-0.84	0.2	1.2	1.39	1.05	0.31
I2	1.84	-0.59	0.19	1.1	0.82	1	0.05
I3	1.78	-0.19	0.17	1.05	0.48	1.06	0.45
I4	1.62	0.54	0.15	1.04	0.44	1.02	0.22
I5	1.78	-0.21	0.17	1.06	0.62	0.96	-0.19
I6	1.7	0.19	0.16	1.12	1.29	1.02	0.24
I7	1.67	0.33	0.16	0.91	-1	0.87	-1.13
I8	1.69	0.25	0.16	0.95	-0.56	0.87	-1.13
I9	1.53	0.92	0.15	0.8	-2.47	0.78	-2.57
I10	1.81	-0.4	0.18	1.04	0.4	0.92	-0.42
		0.00	0.17	1.03	0.1	0.96	-0.40
		0.54	0.02	0.11	1.2	0.09	0.90
RMSE:0.17	S.D.: 0.51	Ayırma İndeksi:3,01		Güvenirlik:0.90			
χ^2 (9): 91,6	*p*: 0.00						

Tablo 4: Ayrı Modele Dayalı Puanlayıcı Kestirimleri ve Standartlaştırılmış Farklar

Puanlayıcı	Gün 1		Gün 2		Gün 3		Gün 4		Z_SAIrc		
	Ölçüm	S.H.	Ölçüm	S.H.	Ölçüm	S.H.	Ölçüm	S.H.	Gün 2-1	Gün 3-1	Gün 4-1
1			-1.7	0.51	-1.3	0.45	-1.75	0.41			
2	-0.37	0.69	-0.07	0.54	-0.41	0.51	-1.86	0.49	0.34	-0.05	-1.76
3	-1.55	0.59	-0.07	0.54	1.05	0.76			1.85	***2.70***	
4			0.56	0.58	-1.1	0.46	-1.58	0.42			
5	-0.22	0.46					-1.09	0.55			-1.21
6	0.22	0.48	0.92	0.62	-1.81	0.63	0	0.51	0.89	***-2.56***	-0.31
7	0.99	1.06	1.35	0.69	0.61	0.81	2.25	1.04	0.28	-0.28	0.85
8			0.56	0.58	0.18	0.58	1.26	0.79			
9			1.35	0.69	-0.95	0.71					
10					-0.41	0.51	-1.23	0.42			
11	0.46	0.51	-0.07	0.54	1.45	1.07	0.28	0.54	-0.71	0.84	-0.24
12	1.05	0.58	-0.07	0.54	0.61	0.81	2.25	1.04	-1.41	-0.44	1.01
13	-0.42	0.45	-0.07	0.54	1.83	1.04	3.37	1.84	0.50	***1.99***	***2.00***
14	1.05	0.58	-0.07	0.54	0.61	0.81	2.25	1.04	-1.41	-0.44	1.01
15	0.74	0.54	0.56	0.58	1.45	1.07	1.46	0.76	-0.23	0.59	0.77
16	-0.01	0.47					1.5	1.88			0.78
17	-2.21	0.41	-2.24	0.53	-0.65	0.49	0.35	0.61	-0.04	***2.44***	***3.48***
18	1.42	0.64	0.87	0.63	0.18	0.58	0.3	0.62	-0.61	-1.44	-1.26
19	0.74	0.54	1.89	0.8	-2.19	0.61	0	0.51	1.19	***-3.60***	-1.00
20	0.74	0.54	0.23	0.56	-2.25	0.43	0.1	0.77	-0.66	***-4.33***	-0.68
21	-0.42	0.45	-1.7	0.51	1.83	1.04	-1.86	0.49	-1.88	***1.99***	***-2.16***
22	0.22	0.34	0.92	0.62	-0.07	0.43	1.33	1.88	0.99	-0.53	0.58
23	-1.1	0.44	-0.07	0.54	1.05	0.76	-1.23	0.42	1.48	***2.45***	-0.21
24	0.46	0.51	-0.07	0.54	1.45	1.07	0.28	0.54	-0.71	0.84	-0.24
25	-0.42	0.45	-0.63	0.52	-0.41	0.51	-0.46	0.46	-0.31	0.01	-0.06
26	-0.99	0.43	-1.14	0.76	-1.97	0.46			-0.17	-1.56	
27					0.18	0.58					
28	0.46	0.51	-0.07	0.54	1.45	1.07	0.28	0.54	-0.71	0.84	-0.24
29	-0.81	0.59	-1.14	0.76	-0.41	0.51			-0.34	0.51	

Target Num	Target ra	Target Measr	S.E.	Obs-Exp Average	N	Context TimeP	Target Measr	S.E.	Obs-Exp Average	N	Context TimeP	Target Contrast	Joint S.E.	Welch t	d.f.	Prob.
100120	17	2.24	.43	-.30	1	T1	-.31	.58	.30	4	T4	2.55	.72	3.54	37	.0011
100120	17	2.24	.43	-.30	1	T1	.54	.47	.15	3	T3	1.70	.64	2.67	37	.0113
12276	3	1.45	.60	-.33	1	T1	-1.08	.75	.16	3	T3	2.53	.97	2.62	26	.0146
100165	23	1.10	.44	-.17	1	T1	-1.08	.75	.23	3	T3	2.18	.87	2.49	35	.0176
100084	13	.36	.45	-.19	1	T1	-1.84	1.03	.12	3	T3	2.20	1.13	1.95	32	.0596
100156	21	.36	.45	.06	1	T1	-1.84	1.03	.31	3	T3	2.20	1.13	1.95	32	.0596
100018	6	-.29	.49	.04	1	T1	1.42	.62	-.32	3	T3	-1.71	.78	-2.18	22	.0401
100135	19	-.81	.54	.09	1	T1	1.79	.60	-.45	3	T3	-2.60	.81	-3.22	23	.0038
100150	20	-.81	.54	.25	1	T1	2.07	.43	-.44	3	T3	-2.88	.69	-4.20	37	.0002

Şekil 1. Etkileşim Raporu (Interaction Pairwise Report)



Şekil 2. Yanlılık/Etkileşim: Puanlayıcı vs. Zaman

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(4);490-510

Kayıp Veri ile Baş Etme Yöntemlerinin Madde Parametrelerine Etkisinin İncelenmesi*

Examination the Effect of Missing Data Techniques of Item Parameters

Ayfer SAYIN ** Alperen YANDI *** Esra OYAR ****

Öz

Bu araştırmada madde bazında kayıp veri oranlarının bulunduğu farklı örneklem büyüklüğündeki verilere ait madde ve test parametrelerinin kayıp veri ile baş etme yöntemlerinden nasıl etkilendiğini belirlemek amaçlanmıştır. PISA 2015 uygulamasına katılan ve çalışma içerisinde yer alan "hırs algısı" ölçeğine cevap veren 5073 öğrenci içerisinden rastgele seçilen 500, 1000 ve 2500 öğrenci, araştırmanın çalışma grubunu oluşturmaktadır. Öncelikle her bir veri setinde normallik, tek boyutluluk, yerel bağımsızlık ve model-veri uyumu varsayımları incelenmiştir. Ölçekte 5 madde yer almaktadır ve kayıp veriler madde bazında oluşturulmuştur. Bu doğrultuda tamamen rastsal olacak şekilde her bir maddeden sırasıyla %5, %10, %15 ve %20'lik kayıp veriler oluşturulmuş, ölçek maddelerinden birinde ise hiç kayıp veri olmayacak şekilde analizler gerçekleştirilmiştir. Kayıp verilerin tamamen rastsal dağılım gösterdiği belirlendikten sonra öncelikle tam ve eksik verilerle; daha sonra silme, ortalama atama, yakın noktalar ortalama ataması, yakın noktalar medyan ataması, doğrusal değerleme, noktada doğrusal eğilim, regresyon atama ve beklenti maksimizasyonu yöntemleri sonucunda elde edilen tam veri setleri ile hesaplamalar gerçekleştirilmiştir. Hesaplama sürecinde betimsel istatistikler ve Cronbach-alfa güvenirlik katsayısı; ardından Madde Tepki Kuramına dayalı Aşamalı Tepki Modeline göre ayırıcılık ve güçlük indeksleri ile marjinal güvenirlik katsayısı hesaplanmıştır. Araştırma sonucunda madde ve test parametrelerinin eksik veriden ve kayıp veri ile baş etme yöntemlerinden etkilendiği; tam veri setine en yakın kestirimi sunan sonuçların doğrusal değerleme yöntemi ile elde edildiği belirlenmiştir.

Anahtar Kelimeler: Kayıp veri, tamamen rastsal kayıp veri, madde tepki kuramı, aşamalı tepki modeli

Abstract

In this study, the aim is to determine how the item and test parameters affect the missing data techniques for different sample sizes and different items with different missing data rates. 500, 100 and 2500 students randomly selected from the 5073 students who participated in the PISA 2015 study and responded to the "ambition perception" scale included in the study constitute the study group of the research. First of all, the assumptions of normality, unidimensionality, local independence and model-data fit were examined for each data set. Afterwards, 5%, 10%, 15%, and 20% missing data were formed for four out of five items and there was no missing data in one item, then analyses were carried out. Once it is determined that the missing data are missing completely random, first with complete and incomplete data, then with serial mean, median of nearby points, mean of nearby points, linear interpolation, linear trend at point, regression, expectation maximization algorithm data item and test parameters were estimated. In the estimated process, descriptive statistics and

* Bu çalışma 20-23 Nisan 2017 tarihlerinde gerçekleştirilen 26. Uluslararası Eğitim Bilimleri Kongresi'nde sözlü bildiri olarak sunulmuştur.

** Dr., Gazi Üniversitesi, Gazi Eğitim Fakültesi, Ankara-Türkiye, ayfersayin@yahoo.com, ORCID ID: orcid.org/0000-0003-1357-5674.
*** Arş. Gör, Dr. Abant İzzet Baysal Üniversitesi, Eğitim Fakültesi, Bolu-Türkiye, e-posta:alperenyandi@gmail.com, ORCID ID: orcid.org/0000-0002-1612-4249
**** Arş. Gör. Gazi Üniversitesi, Gazi Eğitim Fakültesi, Ankara-Türkiye, esra.tas18@gmail.com, ORCID ID: orcid.org/0000-0002-4337-7815

DOI: 10.21031/epod.355506
Geliş Tarihi: 17.11.2017
Kabul Tarihi: 15.12.2017

cronbach alpha reliability coefficient and marginal reliability coefficient; the threshold parameters and the difficulty indices were estimated according to the graded response theory, which is one of the IRT models. The results of the study showed that the item and test parameters were influenced by incomplete and missing data techniques; it was determined that the best estimation results were obtained by linear interpolation method with different data.

Keywords: Missing data, missing data techniques, item response theory, graded response theory

GİRİŞ

Eğitim ve psikoloji alanlarında araştırmacılar genellikle, belirledikleri psikolojik yapılara bireylerin ne düzeyde sahip olduğunu ortaya koymayı hedeflemektedir. Bu hedef doğrultusunda psikolojik yapılarla ilgili olarak elde edilen sonuçlara dayalı olarak birey veya gruplarla ilgili önemli kararlar alınmaktadır. Bireylerin psikolojik yapılara sahip olma düzeylerinin doğrudan ölçülmesi çoğu zaman mümkün değildir. Gizil değişken özelliğine sahip olan bu yapıların ölçülmesinde bireylerin ölçme araçlarındaki gözlenen değişkenlere (maddeler) vermiş olduğu tepkilerden yararlanılmaktadır (Hambleton, Swaminathan, Cook, Eignor ve Gifford, 1977).

Bireylerin psikolojik yapılarını belirlemeyi amaçlayan bu çalışmalarda, araştırmacılar bireylerin gözlenen değişkenlere eksiksiz şekilde yanıt vermelerini sağlamaya çalışmaktadır. Bu durumun nedeni olarak ölçme araçlarının psikometrik özelliklerinin incelenmesi ve belirlenen amaç doğrultusunda belirlenmeye çalışılan sonuçlar için yapılan analizlerde eksiksiz veri ile çalışılmasının gerekli olması durumu gösterilebilir. Ancak psikolojik yapılarla ilgili yapılan ölçme uygulamalarında bireylerin kendilerini ifade etmesi yoluyla veri toplama süreci yürütüldüğünden, çeşitli nedenlerden dolayı eksik verilerin ortaya çıkma durumu ile karşı karşıya kalınabilmektedir. Uygulama sonucunda elde edilen veri setlerinde ortaya çıkan eksiklikler kayıp veri olarak nitelendirilmektedir. Çok sayıda madde içeren ölçme araçlarının kullanılması, veri kayıt sürecinde ortaya çıkan teknik hatalar, bireylerin ölçülen konuya karşı hassas olmaları durumundan dolayı maddelere cevap vermek istememesi, fiziksel nedenler, konu ile ilgili duyarsızlık ve bilgi eksiklikleri, zamanı yetiştirememe problemi gibi nedenler kayıp verinin ortaya çıkmasına sebep olmaktadırlar (Field, 2005; Goregebeur, De Boeck ve Molenberghs, 2010).

Araştırmalar kapsamında elde edilen kayıp veriler, ölçülmek istenen psikolojik yapının tam olarak ölçülememesine yol açmaktadır. Kayıp verilerin varlığı nedeniyle, bireyler veya gruplar ilgili psikolojik özellik açısından yanlı kestirimlerde bulunulabilir. Bu durum elde edilen sonuçların doğruluğunu olumsuz yönde etkileyebilir. Leeuw, Hox ve Huisman (2003), eksik verinin en temelde bilgi eksikliğine yol açtığını, aynı zamanda kestirimlerin etkisini azalttığını ve istatistiksel test gücünü azalttığını belirtmektedir. Geniş ölçekli testlerde de öğrencilerin hem testlerde hem de ölçek ya da anket maddelerindeki eksik verilerinin olması durumunda bireyler hakkında eksik bilgiye ulaşıldığı için yanlılık olabileceği ifade edilmektedir (Rose, Davier ve Xu, 2010). Benzer şekilde Doğanay Erdoğan (2012) da kayıp verinin bilgi kaybına neden olduğuna ve bu durumda ölçülmek istenen özelliğin doğru ve güvenilir olarak ölçülüp ölçülmediği sorununun ortaya çıktığını ifade etmektedir. Araştırma sürecinde gerçekleştirilen istatistiksel analizlerin işleyişinde kayıp verilerin yer aldığı veri setleri için problemlerle karşılaşılabilir (Bal, 2003) çünkü faktör analizi gibi bazı istatistikler tam veri seti matrisi gerektirmektedir (Peng, Harwell, Liou ve Ehman, 2007). Örneğin özellikle küçük örneklem büyüklüklerinde kayıp veri II. tip hatanın artmasına neden olabilir, pozitif olmayan kovaryans matrisi oluşturulmasına neden olabilir, varyansı azaltabilir (Acock, 2005). Kayıp verilerin olduğu durumlarda, özellikle kayıp veri göz ardı edilerek hesaplamalar yapılırsa sonuçlar yanlı hesaplanır ve bu da alınan kararların yanlış olmasına neden olmaktadır (Ambler ve Omar, 2007). Örneğin Hedeker, Mermelstein ve Demirtas (2007) tarafından sigara içmeye yönelik yürütülen bir çalışmada eksik verisi bulunan kişilerin gözlemlenen kişilere göre sigara içme olasılıklarının daha yüksek olduğu, başarısız bir sigara içmeyi bırakma süreçlerinin olduğu belirlemiş, eksik verilerin göz ardı edilmesi durumunda bu bilgilerin kaybolacağına dikkat çekilmiştir. Kayıp değerler hem iki kategorili hem de çok kategorili puanlanan verilerde testin

parametre kestirimlerinde önemli bir yanlılık sebebi olabilir (Demir ve Parlak, 2012; Demir, 2013). Finch (2008) de gerçekleştirdiği çalışmasında kayıp verinin olduğu durumlarda maddelerin güçlük ve ayırıcılığının yanlı kestirildiğine işaret etmektedir. Parametrelerde olduğu gibi kayıp verinin bireylerin yeteneklerinin de beklenenin üstünde ya da altında kestirilmesine neden olduğu belirlenmiştir (Ayala, Plake ve Impara, 2001; Hohensinn ve Kubinger). Tüm bu süreçler dikkate alındığında kayıp veride oluşan sorunların dört başlıkta toplanabileceği (Peng ve diğerleri, 2007); (i.) kayıp verinin olması durumunda eksik veriler bilinemediği için yanlılık oluşacağını, bunun da sonuçların temsil edilebilirlik yani ait olduğu grubu ya da evreni tanımlamada hatalara neden olabileceğini belirtmektedir. Ayrıca (ii.) istatistiksel analizlerin gücünün azalmasına, standart hatanın artmasına sebebiyet verebileceğini; (iii.) faktör analizi gibi bazı analizlerin tam veri ile gerçekleştirilmesi gerektiğine, aksi durumda hesaplama yapılamayacağına; (iv.) verilerin yeniden toplama sürecinin çaba, zaman ve maliyet açılarından da sıkıntı oluşturabileceğine dikkat çekmektedir (Peng ve diğerleri, 2007). Bütün bu nedenlerden dolayı kayıp veriler araştırma süreçlerinde üzerinde durulması gereken öncelikli konular arasında yer almaktadır.

Kayıp değerlere ilişkin olarak yapılan incelemelerin ilk adımı verinin örüntü varlığını incelenmek ile başlamaktadır. Veri setinde yer alan kayıp değerlerin oluşturduğu problemin ne derecede önemli olduğu, örüntüye sahip olup olmadığına bağlıdır (Çokluk ve Kayri, 2011). Veri setleri içerisinde belirlenen kayıp verilerin rastsal olarak dağılım göstermesi, bu verilerle baş etme açısından daha düşük düzeyde bir problem teşkil etmektedir. Ancak kayıp değerlerin bir örüntüye bir başka ifadeyle rastsal olmayan bir dağılıma sahip olması; araştırmacıları belirlenen örüntü bağlamında bir yol izlemeye götürdüğünden daha önemli düzeyde bir problem oluşturmaktadır.

Rubin (1976) ve Little ve Rubin (1987) araştırmasında kayıp verilerin oluşum mekanizmasını üç başlıkta ele alınmıştır: tamamıyla rastsal kayıp-TRK (missing completely at random-MCAR), rastsal kayıp-RK (missing at random-MAR) ve rastsal olmayan kayıp-ROK (missing not at random- M-NAR). Allison (2002) ise kayıp veri mekanizmalarını iki varsayım altında incelemektedir. Bu varsayımlar tamamıyla rastsal kayıp-TRK (missing completely at random-MCAR) ve rastsal kayıp-RK (missing at random-MAR) şeklinde sıralanabilir. Acock (2005) da benzer doğrultuda üç tür kayıp veri olduğundan söz eder: tamamıyla rastsal kayıp-TRK, rastsal kayıp-RK ve ihmal edilemez kayıp (nonignorable-NI).

TRK varsayımı, içerisindeki kayıp veri barındıran bir Y değişkeninde bulunan kayıp verilerin olasılığının, bu Y değişkeninin kendi değerine ve veri setindeki diğer değişkenlerle ilişkisiz olmasını ifade etmektedir. TRK varsayımı, Y değişkenindeki kayıp verilerin, veri setindeki bir başka değişkendeki kayıp verilerle ilişkili olmasına izin vermemektedir. Ancak bu durumda dahi verilerin tamamen rastsal olarak kayıp veri olması mümkündür. Veriler büyük bir matris olarak düşünüldüğünde TRK varsayımında kayıp veriler, matrise rastsal bir şekilde dağılmaktadır (Acock, 2005). RK varsayımı ise Y değişkenindeki kayıp verilerin, Y değişkeninin kendi değeri ile ilişkisiz olmasını ifade etmektedir. Bir başka ifadeyle bu varsayım hem Y hem de veri setindeki bir başka değişken olan X değişkeni birlikte ele alınırken, Y değişkeninde kayıp veri görülme olasılığı ile sadece X değişkeni ele alındığında Y değişkeninde kayıp veri görülme olasılığının eşit olması durumudur. Araştırmacılar Little'ın MCAR testi ile TRK varsayımını test etme imkanına sahipken RK varsayımının test edilmesi daha güçtür ve henüz önerilen kesin bir test bulunmamaktadır. Tam veri seti Y'de kayıp verilerin kendisi, gözlenmeyen cevapların yani kayıp verilerin oluşmasına neden oluyorsa bu durumda kayıp veri mekanizması rastsal olmayan kayıp-ROK (M-NAR) olacaktır. ROK veri yapısı kayıp verilerin kendisinin yanı sıra gözlenen diğer verilere de bağlı olabilmektedir (Rubin, 1976). Üniversite öğrencilerinin akademik performanslarının incelendiği bir panel çalışmasında üniversiteden ayrılan öğrencilerin akademik performanslarının daha düşük olma ihtimali bulunmaktadır ve bu durumda oluşan kayıp veri de "ihmal edilemez kayıp" olarak adlandırılmaktadır (Acock, 2005).

Veri setinde yer alan kayıp verilerin belli bir örüntü oluşturmadığı durumlarda kayıp verilerin veri setinden çıkarılması ve tamamlanması gibi farklı çözüm yolları önerilmektedir (Allison, 2002; Carpita ve Manisera, 2011; Demir ve Parlak, 2012; Şahin Kürşad ve Nartgün, 2015). Bu yöntemler

silme, yaklaşık değer atama ve yeni yaklaşımlar (Demir ve Parlak, 2012) ya da silme-basit atama ve model tabanlı atama yöntemleri (Schafer ve Graham, 2009) olarak sınıflandırılabilir.

a) Silme Yöntemi: Bu yöntemde kayıp veri bulunan satırların veri seti dışında bırakılması söz konusudur. Ancak örneklem büyüklüğündeki azalma, çalışmanın geçerlik ve güvenirliğinde yanlı kestirimler elde edilmesine sebep olmaktadır (Akbaş ve Tavşancıl, 2015; Baygül, 2007; Çüm ve Gelbal, 2015; Şahin Kürşad ve Nartgün, 2015; Yılmaz, 2014).

b) Yaklaşık değer atama yöntemleri (Çokluk ve Kayri, 2011): Bu yöntemde geçmiş bilgileri kullanmak, ortalama değer atamak ve regresyon işlemleri gerçekleştirilebilir.

   i) Seri ortalaması (SO): Bu yöntemde kayıp veri içeren değişkende mevcut olan diğer değerlerin tüm katılımcılar için ortalaması alınarak atama işlemi gerçekleştirilir.

   ii) Yakın noktaların ortalaması (YNO): Kayıp verinin bulunduğu hücrenin yakınındaki değerlerin aritmetik ortalaması üzerinden atama işlemleri gerçekleştirilmektedir. Atama işlemi esnasında kayıp verinin bulunduğu hücrenin altındaki ve üstündeki tam değerlerden yararlanılmaktadır.

   iii) Yakın noktaların medyanı (YNM): Bir önceki yöntemle benzer şekilde atama işlemi gerçekleştirilen bu yöntemde kayıp veri hücresinin altında ve üstünde yer alan tam veriler kullanılarak hesaplanan medyan değeri atanarak işlem yapılmaktadır.

   iv) Doğrusal değerleme (DD): Bu yöntemde ise kayıp veri hücresinden önceki ve sonraki ilk tam verinin ortalaması, kayıp değer yerine atanarak işlem yapılmaktadır.

   v) Noktanın doğrusal eğimi (NDE): Bu yöntemde kayıp veri dışında kalan yöntemlerden yararlanılması söz konusudur. Mevcut tam verilerin sahip olduğu yükseliş veya düşüş eğilimi doğrultusunda atama işlemi gerçekleştirilir.

   vi) Regresyon ataması (RA): Kayıp veriler dışında kalan tam veriler kullanılarak elde edilen regresyon modeli aracılığı ile kayıp veriler yerine atama yapılır.

c) Yeni yaklaşımlar (Yılmaz, 2014):

   i) Beklenti maksimizasyonu (BM): Bu yöntem iteratif şekilde tekrar eden iki aşama üzerinden atama işlemi yapılmaktadır. İlk aşama olan beklenti aşamasında, kayıp verilere başlangıç değerleri atanır. İkinci aşama olan maksimizasyon aşamasında ise, bu başlangıç değerleriyle oluşan beklentiler maksimize edilir. Bu beklenti-maksimize etme döngüsü, bundan sonra, atanan değerler, önceden belirlenmiş bir yakınsama kriterine dayalı olarak benzer hale gelene kadar tekrarlanmaktadır.

Kayıp veri ile baş etme yöntemlerinin araştırma sonuçları üzerine etkileri ile ilgili olarak alanyazında farklı birçok çalışma mevcuttur. Çokluk ve Kayri (2011), Ankara Üniversitesi Eğitim Bilimleri Fakültesi Sınıf Öğretmenliği bölümünde öğrenimine devam eden 200 öğrencinin Fatalizm Ölçeği maddelerine verdikleri yanıtlardan oluşan verileri %15 - %20 ve %0 - %50 oranında kayıp veri içerecek şekilde düzenlemiştir. Çalışma kapsamında elde edilen bu veri setlerindeki kayıp verilerin tamamlama işlemi sonrasında faktör yapıları, düzeltilmiş madde-toplam korelâsyonları ve Cronbach-alfa iç tutarlık katsayıları karşılaştırılmıştır. Sonuç olarak çalışma kapsamında tam veri için elde edilen faktör yapılarının benzer olduğu ancak tamamlanmış veriler üzerinden elde edilen açıklanan varyans, öz değer ve iç tutarlılık katsayısında bir düşüş olduğu gözlenmiştir. Köse ve Öztemur (2014) ise örneklem büyüklüğü ve kayıp veri oranının t testi ve ANOVA olmak üzere test istatistikleri üzerindeki etkisini incelemiş, kayıp veri yöntemlerinin fark testlerine etki ettiğini ortaya koymuşlardır. Akbaş ve Tavşancıl (2015) da araştırmalarında liste bazında silme tekniğinin test istatistiğinde yanlı sonuçlara neden olduğunu ve beklenti maksimizasyonu ve çoklu değer regresyon atama tekniklerinin ise genel olarak daha yüksek kestirimler gerçekleşmesine neden olduğunu belirtmişlerdir. Çüm ve Gelbal (2015), yapmış oldukları çalışmada PISA 2012 Türkiye örneklemi, tam veri seti üzerinden tamamen rastsal ve tamamen rastsal olmayacak şekilde %20 ve %30 oranında kayıp veri olacak şekilde veri setlerini düzenlemiştir. Düzenlenen bu veri setlerinde kayıp veriler

yerine 10 farklı yöntemle eksik verilerle baş etme yöntemlerine dayalı tam hâle getirilmiş ve bu durumun model veri uyumu değerlerine etkisi ele alınmıştır. Araştırma sonucunda kayıp verilerin tamamıyla rastsal olarak dağıldığı durumlarda regresyonla atama yöntemi sonrası elde edilen veri üzerinden kestirilen model veri uyum değerlerinin tam veri setinin model uyum değerlerine en yakın sonuçları verdiği tespit edilmiştir. Çalışmada ayrıca, yaklaşık değer atamalarının veri seti için belirlenen dağılımları önemli düzeyde etkilediği, bu nedenle araştırmacıların uygun atama yöntemlerini kullanarak süreci gerçekleştirmeleri önerilmiştir. Şahin Kürşad ve Nartgün (2015) ise yapmış oldukları çalışmada PISA 2012 "Matematik Çalışma Etiği" ölçeği Türkiye örneklemi verileri içerisinden 200 kişilik bir alt örneklem seçmiştir. Seçilen bu veri seti üzerinde %5, %10 ve %20 oranında, tamamen rastsal olacak şekilde veri silme işlemi yapılmıştır. Oluşturulan kayıp verili setleri farklı yöntemlerle tamamlandıktan sonra geçerlik ve güvenirliğe ilişkin kestirimler gerçekleştirilmiştir. Araştırma sonucunda değer atama yöntemleri ile oluşturulan veri setlerinden elde edilen parametre değerlerinin kayıp veri oranının düşük olduğu durumlarda genel olarak tam veri setinden elde edilen değerlere yakın veya aynı değerler verdiği rapor edilmiştir. Bunun yanı sıra ele alınan tüm durumlar için çoklu atama, beklenti maksimizasyonu ve regresyon ataması yöntemlerinin tam veri setinden elde edilen değerlere en yakın değer veren yöntemler olduğu belirtilmiştir. Soysal ve Akın Arıkan (2017) da kayıp veri yöntemleri ile örneklem büyüklüğünün faktörleşmeye olan etkisini araştırdıkları çalışmalarında, faktörleştirme tekniklerinin hemen hemen her koşulda benzer performans gösterdiği ve atama yöntemleri açısından farklılaşmadığı sonucuna ulaşmışlardır. Uluslararası alanyazında da benzer şekilde kayıp veri yöntemlerinin açımlayıcı faktör analizi gibi genel test istatistikleri üzerindeki etkilerinin incelendiği çalışmalar mevcuttur (Josse ve Husson, 2012; McNeish, 2016; Weaver ve Maxwell, 2014).

Ele alınan bu çalışmalar kapsamında kayıp veri ile baş etme yöntemlerinin test parametrelerini etkilediği görülmektedir çünkü alanyazındaki çalışmalarda veri setlerindeki kayıp, testin geneli için oluşturulmuş, veriler madde bazına indirgenmemiştir. Bir başka ifadeyle maddelerde bulunan kayıp veri oranının değişiklik göstermesinin diğer maddeler üzerinde farklı etkileme durumuna yönelik inceleme yapılmamıştır. Ayrıca bu yöntemlerin Madde Tepki Kuramı (MTK) temelli parametreler üzerinde etkisi sınırlı çalışmada ele alınmış (Koçak ve Çokluk Bökeoğlu, 2017), çalışmalarda genel olarak Klasik Test Kuramı temelinde incelemeler yapılmıştır. Madde bazında ortaya çıkan kayıp veri miktarının diğer maddelerin parametreleri üzerindeki etkisinin araştırılması, araştırmacılara kullandıkları kayıp veri ile baş etme yönteminin seçiminde bilgi sağlayabilir. Ayrıca MTK'nın temel varsayımlarından biri olan madde parametresi kestirimlerinin maddelerin birbirlerinden bağımsız gerçekleştirilip gerçekleştirilmediği hakkında da bilgi verebilir. Araştırmacılar çalışma kapsamındaki veri setinde bulunan maddelerdeki kayıp veri oranlarına göre uygun olan yöntemi seçmeleri ile birlikte elde edilen sonuçların doğruluğunu olumlu yönde etkileyebilir. Ayrıca MTK kapsamında ele alınan parametrelerin kayıp veri baş etme yöntemlerinden ne şekilde etkilendiğinin belirlenmesi, bu kuram kapsamında temellendirilen çalışmaların sürecinde araştırmacılara katkı sağlayabilir.

Araştırmanın Amacı

Bu araştırmada diğer araştırmalardan farklı olarak likert tipi derecelendirilmiş bir ölçekte yer alan kayıp veriyle baş etme yöntemlerinin Aşamalı Madde Tepki Kuramı (AMTK) ile kestirilen parametrelere etkisi incelenmiştir. Ayrıca yine diğer araştırmalardan farklı olarak kayıp veri oranı testin geneli için değil, madde bazında değiştirilmiş ve ölçekteki bir madde hiç eksik veri olmayacak şekilde düzenlenmiştir. Bu doğrultuda da bir ölçme aracındaki herhangi bir maddede bulunan kayıp veri miktarının eksik verisi olmayan diğer madde parametrelerine etkisi incelenmiştir. Alanyazında gerçekleştirilen çalışmalarda kayıp veri, örneklem setinin tamamı üzerinden yapılmışken bu çalışmada madde bazlı yapılmış olması ve kestirimlerin aşamalı tepki modeli ile gerçekleştirilmiş olmasının araştırmayı önemli kıldığı düşünülmektedir.

YÖNTEM

Bu araştırmada, farklı örneklem büyüklüklerinde ve kayıp veri oranlarına sahip maddeleri bulunan veri setlerinde madde parametrelerinin kayıp veri baş etme yöntemlerinden ne düzeyde etkilendiğinin belirlenmesi amaçlanmaktadır. Bu doğrultuda araştırmanın temel bir araştırma niteliğinde olduğu söylenebilir.

Örneklem

PISA 2015 çalışmasında yer alan "hırs algısı" ölçeğine eksiksiz cevap veren 5073 öğrenci içerisinden rastgele seçilen 500, 1000 ve 2500 öğrenci, araştırmanın çalışma grubunu oluşturmaktadır. Hesaplamalar MTK'ya göre gerçekleştirildiği için örneklem büyüklüklerinin yeterli sayıda olmasına dikkat edilmiştir. MTK modelleri için alanyazında farklı örneklem büyüklükleri önerilmektedir. Tsutakawa ve Johnson (1990) parametrelerin doğru kestirilmesi için en az 500 örneklem büyüklüğüne ihtiyaç olduğunu belirtmişlerdir.

Verilerin Elde Edilmesi

Araştırmada PISA 2015 uygulamasında yer alan ve öğrencilerin hırs algısını belirlemeyi amaçlayan 5 maddelik "hırs algısı" ölçeği kullanılmıştır. Bu ölçeğin seçilmesinde maddelerin tek boyutta toplanması göz önünde bulundurulmuştur. Ölçekte yer alan maddeler şu şekildedir:

- Derslerimin çoğunda veya tamamında en yüksek notu almak istiyorum.
- Mezun olduğumda bana uygun en iyi fırsatlardan birini seçmek istiyorum.
- Ne yaparsam yapayım en iyisi olmak istiyorum.
- Kendimi hırslı bir insan olarak görüyorum.
- Sınıftaki en iyi öğrencilerden biri olmak istiyorum.

4'lü likert tipinde derecelendirilen ölçek maddeleri olumsuz ifade içermemektedir.

İşlem

Araştırma verilerine OECD'nin internet adresinden (www.oecd.org/pisa/) ulaşılmıştır. Veri seti içerisinden Türkiye örneklemi alınmıştır. Örneklem, OECD sekreteryasında Westat (ABD) liderliğinde seçilmiş ve örneklemin evreni temsil ettiği belirtilmiştir.

Verilerin Analizi

Verilerin analizi sürecinde öncelikle MTK varsayımlarının karşılanıp karşılanmadığının incelenmesi, ardından veri seti içerisinde kayıp verilerin oluşturulması ve kayıp verilerin tamamen rastsal olup olmama durumunun test edilmesi, son olarak da MTK'ya dayalı olarak madde parametrelerinin hesaplanması işlemleri gerçekleştirilmiştir.

Varsayımların İncelenmesi

Araştırma kapsamında hesaplamalar MTK'ya dayalı olarak gerçekleştirildiği için MTK'nın temel varsayımları olan normallik, tek boyutluluk, yerel bağımsızlık ve değişmezlik varsayımları seçilen 500, 1000 ve 2500 kişilik örneklem grupları için incelenmiştir.

Örneklemden elde edilen verilerin evrene ait uzayda normal dağılım gösterip göstermediğinin incelenmesi MTK'nın temel varsayımlarından biridir. Tek değişkenli normalliğin test edilmesinde, çarpıklık ve basıklık katsayıları hesaplanmış ve sonuçlar Tablo 1'de gösterilmiştir.

Tablo 1. Farklı Örneklem Büyüklüğündeki Setlerin Çarpıklık ve Basıklık Katsayıları

Değerler	Örneklem büyüklükleri		
	n=500	n=1000	n=2500
Çarpıklık	-0,975	-0,951	0,929
Basıklık	0,677	0,720	0,691

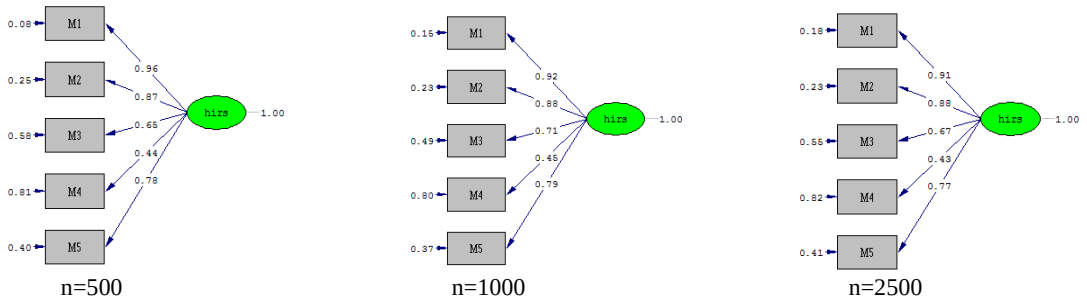
Çarpıklık ve basıklık katsayılarının ±1 arasında hesaplanması, verilerin normal dağılımdan aşırı bir sapma göstermediğini belirletmektedir (Büyüköztürk, 2017). Tablo 1'de yer alan bilgiler incelendiğinde veri setlerinin genel olarak normal bir dağılım gösterdiği görülmektedir. Çarpıklık ve basıklık katsayılarına ek olarak verilere ilişkin histogram grafikleri de incelenmiş, veri setlerinin normal bir dağılım gösterdiği tespit edilmiştir.

MTK'nın diğer önemli varsayımlarından biri tek boyutluluktur. Tek boyutluğun test edilmesi için açımlayıcı ve doğrulayıcı faktör analizi teknikleri kullanılmıştır. Açımlayıcı faktör analizi sonuçlarına Tablo 2'de yer verilmiştir.

Tablo 2. Farklı Örneklem Büyüklüğündeki Setlerin Açımlayıcı Faktör Analizi Sonuçları

Değerler	Örneklem büyüklükleri		
	n=500	n=1000	n=2500
Özdeğer	2,760	2,813	2,704
Açıklanan varyans (%)	55,204	56,251	54,084

Tablo 2'de yer alan değerler incelendiğinde tüm örneklem setlerinde 5 maddeden oluşan ölçek maddelerinin tek bir boyutta toplandığı ve boyutun özdeğerinin 2'den, açıklanan varyans oranının ise %30'dan büyük (Çokluk, Şekercioğlu ve Büyüköztürk, 2014) olduğu belirlenmiştir. Ardından doğrulayıcı faktör analizine başvurulmuş ve analizin sonucunda tüm maddelerin anlamlı t değerlerine sahip oldukları, başka bir ifadeyle anlamlı bir açıklayıcılıklarının bulunduğu belirlenmiştir. Bu analiz sonucunda oluşturulan yol (path) diyagramları Şekil 1'de gösterilmiştir.



n=500 n=1000 n=2500

Şekil 1. Farklı Örneklem Büyüklüğündeki Setlerin Doğrulayıcı Faktör Analizi Yol Diyagramları

Açımlayıcı ve doğrulayıcı faktör analizi sonuçlarına göre ölçek maddelerinin 500, 1000 ve 2500 kişilik örneklem setlerinde tek boyutta toplandığı tespit edilmiştir. Yerel bağımsızlık tek boyutlulukla ilişkili bir özelliktir. Eğer bir ölçek tek boyutluluk özelliğini gösteriyorsa, bu ölçekte yer alan maddelerin yerel bağımsızlık özelliğine sahip olduğu söylenebilir (Hambleton ve Swaminathan, 1997).

MTK'nın bir diğer varsayımı madde parametrelerinin kestirimin yapıldığı gruptan; yetenek parametrelerinin ise kestirimin yapıldığı maddelerden bağımsız olmasıdır (Wright, 1997). Bu varsayımın test edilmesi amacıyla öncelikle madde parametrelerini kestirmek için grup rastgele ikiye bölünmüş, sonrasında her iki grup için de madde parametreleri kestirilmiştir. Kestirilen parametreler arasında Pearson korelasyon katsayısı hesaplanmış ve korelasyon değerlerinin kabul edilebilir aralıkta olduğu (>.70) belirlenmiştir. Parametre değişmezliğinin ikinci kısmında ölçekte yer alan maddeler tek ve çift maddeler olarak ayrılmış sonrasında ölçeği alan bireyler için iki ayrı yetenek parametresi kestirilmiştir. Kestirilen yetenek parametreleri arasında Pearson korelasyon katsayısı hesaplanmış ve korelasyon değerlerinin anlamlı olduğu bulunmuştur.

Hırs algısı ölçeğinde yer alan beş maddeye ilişkin verilerin model veri uyumu için "-2likelihood (negative twice the loglikelihood)" değeri hesaplanmıştır. Bu değer 500 örneklem büyüklüğünde 2215,3 hesaplanmış, 1000 örneklem büyüklüğünde -5972.8 olarak; 2500 örneklem büyüklüğünde -19003.9 olarak hesaplanmıştır; -2likelihood değerinin yüksek olması modelin uyumlu olduğunun bir göstergesidir.

Kayıp Verilerin Oluşturulması ve Test Edilmesi

Varsayımların incelenmesinin ardından tamamen rastsal kayıp veri-TRK (MCAR) setleri oluşturulabilmesi için tüm gözlemlerin kayıp olma ihtimallerinin birbirine eşit olması gerekmektedir. Bu nedenle öncelikle gözlemler ile değişkenler birbiri ile çarpılarak toplam hücre sayısı bulunmuştur. Her bir hücre içinde eksiltilecek veri sayısı belirlenmiştir. Ardından MS Excell dosyasında oluşturulan bir betikle (random seçim; aynı casede dört veri silme vb.) silme işlemi gerçekleştirilmiştir. Bu kısımda yapılan veri silme işlemi tamamen rastsaldır ve TRK mekaznizmasının mantığını uygun yapılmaktadır. Öncelikli olarak kayıp veri oluşturulmuş ve sonrasında kayıp verinin rastsal olup olmadığı Little'ın MCAR testi uygulanarak test edilmiştir. Daha önce de belirtildiği gibi eksiltme işlemleri maddeler bazında gerçekleştirilmiştir. Her bir veri setinde birinci maddenin %20'si; ikinci maddenin %15'i; üçüncü maddenin %10'u; dördüncü maddenin %5'i kayıp veri durumuna getirilmiştir. Beşinci maddede ise herhangi bir kayıp veri oluşturulmamıştır. Bu doğrultuda her bir veri setinde %12 kayıp veri bulunmuştur.

Kayıp veriler oluşturulduktan sonra kayıp verilerin, veri setlerinde tamamen rastsal bir şekilde dağılıp dağılmadığının belirlenmesi amacıyla Little'ın MCAR testi hesaplanmıştır. Hesaplama sonuçları Tablo 3'te gösterilmiştir.

Tablo 3. Farklı Örneklem Büyüklüğündeki Setlerdeki Eksik Verilerin Tamamen Rastsallığına İlişkin Hesaplanan Little'ın MCAR Testi Sonuçları

Değerler	Örneklem büyüklükleri		
	n=500	n=1000	n=2500
X2	38,001	27,381	48,309
sd	36	40	40
p	0,378	0,935	0,172

Little'ın MCAR testi eksik verilerin TRK olarak dağılıp dağılmadığının belirlenmesinde kullanılan en yaygın testtir. Little'ın MCAR testi için hesaplanan p değerinin anlamlı olmaması yani 0,05'ten büyük olması, eksik verilerde herhangi bir örüntü bulunmadığını içeren yokluk hipotezinin kabul edildiğini, kayıp verinin TRK olduğunu göstermektedir (Garson, 2015). Bu doğrultuda Tablo 3 incelendiğinde her bir veri setinde yer alan eksik verilerin tamamen rastsal bir dağılım gösterdiği belirlenmiştir (p>0,05).

Verilerin Çözümlenmesi

Verilerin çözümleme sürecinde öncelikle eksik veriler için silme, ortalama atama, yakın noktalar ortalama ataması, yakın noktalar medyan ataması, doğrusal değerleme, noktada doğrusal değerleme, regresyon atama ve beklenti maksimizasyonu atama olmak üzere 8 kayıp veri ile baş etme yöntemine dayalı olarak eksik veri setleri tamamlanmıştır. Bu doğrultuda her bir örneklem büyüklüğü için tam veri seti, eksik (eksiltilmiş, kayıplar oluşturulmuş) veri seti, 8 farklı kayıp veri başa çıkma yöntemine göre tam hâle getirilmiş veri setleri olmak üzere 10'ar veri seti oluşturulmuştur. Bu veri setleri üzerinde test istatistikleri için ortalama, standart sapma, Cronbach alfa güvenirlik ve marjinal güvenirlik katsayısı hesaplanmıştır. Ardından da Aşamalı Tepki Modeli'ne göre ayırıcılık ve güçlük indeksleri hesaplanarak sonuçlar karşılaştırılmıştır. Araştırmada kullanılan ölçek maddeleri çok kategorili puanlandığı için Samejima'nın Aşamalı Tepki Modeli olarak adlandırılan model kullanılmıştır. Bu kuram cevaplayıcının yetenek düzeyi ile belli bir kategoriye tepki vermesi arasında doğrusal olmayan ilişkilere dayanmaktadır ve bireylerin yetenekleri hakkında bilgi elde edilmektedir. Eşik parametresi, bir maddenin her bir kategoriye kadar olan sınıflarının 0,50 olasılıkla yanıtlanması için gerekli olan düzeyi tanımlar ve kategori sayısının bir eksiği kadar eşik parametresi hesaplanır. Modelde bir madde için her bir eşik değeri için bir işlem karakteristik eğrisi, her eşik için güçlük indeksi ve maddenin tümü için bir ayırıcılık parametresi kestirilmektedir (Emretson ve Reise, 2000). Madde parametrelerinin hesaplanmasında MULTILOG programı kullanılmıştır.

BULGULAR

1. 500 örneklem büyüklüğünde kayıp veri ile baş etme yöntemlerine dayalı olarak test ve madde istatistikleri nasıl bir değişim göstermektedir?

Araştırma kapsamında 500 örneklem büyüklüğü için oluşturulan 10 veri için betimsel istatistikler ve güvenirlik katsayıları hesaplanmış, sonuçlar Tablo 4'te gösterilmiştir.

Tablo 4. Betimsel İstatistikler ve Güvenirlik Katsayıları (n=500)

Veri setleri	Ortalama	SS	Cronbach alfa güvenirlik	Marginal güvenirlik
Tam veri	17,080	2,580	0,775	0,738
Eksik veri	15,370	3,250	0,772*	0,731
Silme	17,160	2,540*	0,764	0,735
Ortalama atama	17,120	2,350	0,731	0,711
Yakın noktalar ortalama ataması	17,100*	2,510	0,770	0,735
Yakın noktalar medyan ataması	17,110	2,530	0,771	0,734
Doğrusal değerleme	17,100*	2,530	0,770	0,736*
Noktada doğrusal eğilim	17,005	2,400	0,735	0,727
Regresyon atama	17,110	2,540	0,770	0,732
Beklenti maksimizasyonu	17,110	2,520	0,782	0,734

* tam veri parametrelerine en yakın olan değerler

Tablo 4'te görüldüğü gibi tam veri setinden elde edilen değerler, yani veri setinin hiçbir eksik verisi bulunmayan hâlinde hesaplanan değerler, (ilk satır) hem sonradan oluşturulan eksik veri seti için hem de eksik verilerle baş etme yöntemlerine dayalı tam hâle getirilmiş veri setleri için bir referans oluşturmaktadır. Bu doğrultuda eksik veriler ile baş etme yöntemlerine dayalı tam hâle getirilmiş veri setlerinde hesaplanan değerler; veri setlerinin hesaplama yanlılığından ne düzeyde etkilendiğini belirlemek amacıyla tam veri setinden elde edilen değerler ile karşılaştırılmıştır.

Tablo 4'te yer alan bilgiler incelendiğinde, 500 örneklem büyüklüğüne tam veri seti ile eksik veri setinin ortalama ve standart sapma betimsel istatistikleri arasında farklılıklar olduğu görülmektedir. Bununla birlikte kayıp veri ile baş etme yöntemlerine dayalı olarak oluşturulan veri setlerinin ortalama, standart sapma ve güvenirlik katsayılarının tam veri ile büyük ölçüde benzerlik gösterdiği belirlenmiştir. Noktada doğrusal atama yöntemi (-0,075 fark) dışındaki eksik veri ile baş etme

yöntemleri sonucunda ortalama değerin tam veriye yakın olmakla birlikte biraz daha yüksek kestirildiği belirlenmiştir. Güvenirlik katsayılarının yüksek kestirilmiş olmakla birlikte gerçek değere oldukça yakın olduğu görülmektedir. Başka bir ifadeyle 500 örneklem büyüklüğünde eksik veri için kullanılan kayıp veri ile baş etme yöntemlerinin test istatistiklerine olan etkisi benzerlik göstermektedir.

500 örneklem büyüklüğünde kayıp veri ile baş etme yöntemlerinin madde parametreleri üzerindeki etkisini belirlemek amacıyla aşamalı tepki modeli ile madde parametreleri hesaplanmış ve sonuçlar Tablo 5'te gösterilmiştir.

Tablo 5. Aşamalı Madde Tepki Modeline Dayalı Kestirilen Ayırıcılık ve Güçlük İndeksleri (n=500)

Madde no	Madde parametreleri	Veri setleri									
		1	2	3	4	5	6	7	8	9	10
M1 (%20 eksik veri)	a	5,86	7,51	7,58	3,63	7,79	8,69	6,81*	4,44	7,05	11,5
	b_1	-2,26	-2,21	-2,23	-2,51	-2,24	-2,24	-2,25*	-2,45	-2,3	-2,15
	b_2	-1,81	-1,84	-1,82*	-2,1	-1,78	-1,77	-1,8*	-2,06	-1,82*	-1,72
	b_3	-0,52	-0,57	-0,61	-0,81	-0,54	-0,54	-0,53	-0,55	-0,55	-0,52*
M2 (%15 eksik veri)	a	3,2	2,25	3,39	2,01	3,02	3,41	3,17*	2,27	2,99	4,64
	b_1	-2,44	-2,41*	-2,32	-2,83	-2,51	-2,49	-2,55	-2,86	-2,55	-2,35
	b_2	-1,95	-2	-1,98*	-2,36	-2,1	-1,98*	-2,08	-2,39	-1,99	-1,89
	b_3	-0,61	-0,64	-0,59*	-0,85	-0,67	-0,66	-0,63*	-0,75	-0,63*	-0,58
M3 (%10 eksik veri)	a	1,69	1,75	1,81	1,96	1,65	1,65	1,69*	2,0	1,96	1,75
	b_1	-3,08	-2,97	-2,79	-3,01	-3,08*	-3,09	-3,07	-3,08*	-2,96	-2,98
	b_2	-1,64	-1,62	-1,55	-1,68	-1,7	-1,71	-1,68	-1,71	-1,64*	-1,66
	b_3	-0,03	-0,07	-0,13	0,11	-0,05	-0,05	-0,04*	0,01	-0,08	-0,02*
M4 (%5 eksik veri)	a	0,95	1,01	0,88	0,98	0,99	0,99	0,95*	1,11	1,01	0,95*
	b_1	-3,47	-3,41	-3,54	-3,43	-3,39	-3,39	-3,52	-3,12	-3,3	-3,47*
	b_2	-1,43	-1,4	-1,57	-1,45	-1,41	-1,44*	-1,44*	-1,05	-1,39	-1,42*
	b_3	0,97	0,94	0,9	1,03	0,91	0,9	0,96	0,95	0,97*	1,09
M5 (%0 eksik veri)	a	2,3	2,1	2,12	2,05	2,24	2,23	2,28*	1,94	2,08	2,03
	b_1	-2,5	-2,6	-2,65	-2,62	-2,51	-2,52	-2,5*	-2,72	-2,6	-2,6
	b_2	-1,75	-1,82	-1,92	-1,84	-1,76	-1,77	-1,75*	-1,89	-1,82	-1,81
	b_3	-0,1	-0,12	-0,14	-0,12	-0,11*	-0,11*	-0,11*	-0,12	-0,11*	-0,09*

1: tam veri, 2: eksik veri, 3: silme, 4: ortalama atama, 5: yakın noktalar ortalama ataması, 6: yakın noktalar medyan ataması, 7: doğrusal değerleme, 8: noktada doğrusal eğilim, 9: regresyon atama, 10: beklenti maksimizasyonu
* tam veri parametrelerine en yakın olan değerler

Tablo 5'te görüldüğü gibi birinci maddede %20 eksik veri bulunmaktadır. Ortalama atama (-2,23 fark), doğrusal değerleme (-0,95 fark) ve noktada doğrusal eğilim (-1,42 fark) yöntemleri ile tamamlanan veri setlerindeki ayırıcılık indekslerinin tam veridekinden daha düşük; diğer yöntemlerle kestirilen verilerin de daha yüksek olduğu belirlenmiştir. Ayırıcılığın en yakın hesaplandığı yöntemin 0,95 farkla doğrusal değerleme olduğu tespit edilmiştir. Benzer şekilde %15 eksik verisi bulunan ikinci maddede ortalama atama (-1,19 fark), yakın noktalar ortalama ataması (-0,18 fark), noktada doğrusal eğilim (-0,93 fark), regresyon atama (-0,21 fark) başa çıkma yöntemleriyle oluşan setlerindeki ve eksik veri setindeki (-0,95 fark) ayırıcılık indekslerinin daha düşük; diğer yöntemlerle olanlarda daha yüksek olduğu belirlenmiştir. Tam veri ile hesaplanan ayırıcılık indeksinin en yakın olduğu yöntemin doğrusal değerleme yöntemiyle (0,03 fark) yapılan kestirim olduğu belirlenmiştir. Üçüncü maddede %10 eksik veri bulunmaktadır ve doğrusal değerleme yöntemi kullanılarak kayıp verilerin atandığı veri seti ile tam veri setinin ayırıcılık indeksleri bire bir aynı (0,00 fark) hesaplanmıştır. Diğer eksik verilerle baş etme yöntemlerine dayalı tam hâle getirilmiş veri setleri ile gerçekleştirilen hesaplamaların da genel olarak benzerlik gösterdiği belirlenmiştir. %5 eksik verisi bulunan dördüncü maddenin ayırıcılık indekslerinin noktada doğrusal eğilim yöntemi (0,16 fark) ile oluşturulan veri setinde farklılaştığı, diğer veri setlerinde tam veri seti ile büyük ölçüde benzerlik gösterdiği belirlenmiştir. Benzer şekilde doğrusal değerleme (0,00 fark) ve beklenti maksimizasyonu (0,00 fark) yöntemleriyle oluşturulan veri setlerinin madde ayırıcılık indekslerinin tam veri seti ile aynı çıktığı görülmektedir. Ölçeğin beşinci ve son maddesinde eksik veri bulunmamaktadır. Ancak

diğer maddelerdeki eksik verilerle baş etme yöntemlerinden madde ayırıcılık indeksinin etkilendiği belirlenmiştir. Doğrusal değerleme yöntemi (0,02 fark) ile gerçekleştirilen atama sonucunda beşinci maddenin ayırıcılık indeksinin diğer veri setlerinden daha yakın olduğu tespit edilmiştir. Madde güçlük indekslerinin veri setleri bazında büyük farklılıklar göstermediği; tam veri seti ile en fazla benzerliğin yine doğrusal değerleme yöntemiyle atanan veri setinde olduğu belirlenmiştir.

500 örneklem büyüklüğünde maddedeki kayıp veri oranı azalma gösterdikçe; kayıp veri ile baş etme yöntemleri sonucunda oluşturulan veri setlerinin ayırıcılıklarının tam veride hesaplanan madde ayırıcılık indeksine yaklaştığı belirlenmiştir. Aynı zamanda maddenin ayırıcılık indeksi artış gösterdikçe de farkın arttığı tespit edilmiştir. Araştırma kapsamında beşinci maddede hiç kayıp veri bulunmamasına ve MTK`nın yerel bağımsızlık varsayımı karşılanmasına karşın beşinci maddenin parametre kestiriminin diğer maddelerdeki kayıp verilerden etkilendiği saptanmıştır. Madde güçlük indekslerinin ise madde ayırıcılık indeksine göre kayıp veri ile baş etme yöntemlerinden daha az etkilendiği belirlenmiştir. Tam veri seti üzerinden hesaplanan madde ayırıcılık ve madde güçlük indeksinin en çok doğrusal değerleme yönteminin kullanıldığı veri seti sonuçları ile benzerlik gösterdiği belirlenmiştir.

2. 1000 örneklem büyüklüğünde kayıp veri ile baş etme yöntemlerine dayalı olarak test ve madde istatistikleri nasıl bir değişim göstermektedir?

1000 öğrenciden oluşan veri setinin de benzer şekilde birinci maddesinin %20'si, ikinci maddesinin %15'i, üçüncü maddesinin %10'u, dördüncü maddesinin %5'i eksiktir, beşinci maddede ise eksik veri bulunmamaktadır. 1000 örneklem büyüklüğünde de benzer şekilde tam veri, eksik veri ve 8 farklı kayıp veri baş etme yöntemi ile tam hale getirilmiş veri setleri üzerinde betimsel istatistikler ve güvenirlik katsayıları hesaplanmış, sonuçlar Tablo 6'da gösterilmiştir.

Tablo 6. Betimsel İstatistikler ve Güvenirlik Katsayıları (n=1000)

Veri setleri	Ortalama	SS	Cronbach alfa güvenirlik	Marginal güvenirlik
Tam veri	17,180	2,570	0,779	0,741
Eksik veri	15,400	3,300	0,784	0,738
Silme	17,220	2,560	0,784	0,743
Ortalama atama	17,180*	2,370	0,743	0,723
Yakın noktalar ortalama ataması	17,180*	2,550	0,779*	0,740
Yakın noktalar medyan ataması	17,190	2,550	0,778	0,738*
Doğrusal değerleme	17,180*	2,550	0,777	0,738
Noktada doğrusal eğilim	17,180*	2,380	0,738	0,722
Regresyon atama	17,160	2,570*	0,778	0,736
Beklenti maksimizasyonu	17,170	2,540	0,792	0,743

* tam veri parametrelerine en yakın olan değerler

Tablo 6 incelendiğinde, 1000 örneklem büyüklüğünde tam veri ile hesaplanan ortalama değer ile kayıp veri ile baş etme yöntemleri sonucunda oluşturulan veri seti sonuçlarının büyük oranda benzerlik gösterdiği görülmektedir. Ancak eksik veri ile hesaplanan betimsel istatistiklerin tam veri setinden uzaklaştığı belirlenmiştir. Güvenirlik katsayılarının da benzer şekilde 9 veri setinde tam veri seti ile hesaplanan sonuçlara oldukça yakın olduğu tespit edilmiştir.

1000 örneklem büyüklüğünde hırs algısı ölçeğinde yer alan maddelerin ayırıcılık ve güçlük indeksleri AMTK'ya dayalı olarak hesaplanmış; sonuçlar Tablo 7'de gösterilmiştir.

Tablo 7. Aşamalı Madde Tepki Modeline Dayalı Kestirilen Ayırıcılık ve Güçlük İndeksleri (n=1000)

Madde no	Madde parametreleri	Veri setleri									
		1	2	3	4	5	6	7	8	9	10
M1 (%20 eksik veri)	a	4,45	4,39	5,79	3,26	4,61	4,58	4,47*	3,18	4,39	6,23
	b_1	-2,43	-2,45	-2,29	-2,67	-2,4	-2,4	-2,42*	-2,78	-2,45	-2,32
	b_2	-1,8	-1,83	-1,77	-2,02	-1,79*	-1,79*	-1,79*	-2,1	-1,82	-1,76
	b_3	-0,66	-0,69	-0,67	-0,91	-0,66*	-0,66*	-0,67	-0,61	-0,64	-0,61
M2 (%15 eksik veri)	a	3,41	3,19	3,52	2,3	3,22	3,18	3,4*	3,13	3,47	4,69
	b_1	-2,52	-2,61	-2,52*	-3,01	-2,68	-2,69	-2,65	-2,71	-2,62*	-2,41
	b_2	-2,01	-2,08	-2,05*	-2,4	-2,1	-2,13	-2,11	-2,2	-1,98	-1,88
	b_3	-0,57	-0,58	-0,59	-0,8	-0,62	-0,63	-0,62	-0,57*	-0,56	-0,55
M3 (%10 eksik veri)	a	2,12	2,16	2,24	1,93	2,15	2,1*	2,1*	1,68	1,99	2,07
	b_1	-2,91	-2,88	-3,04	-3,11	-2,93*	-2,96	-2,96	-3,42	-2,96	-2,95
	b_2	-1,49	-1,51*	-1,53	-1,66	-1,52	-1,53	-1,53	-1,79	-1,56	-1,55
	b_3	-0,11	-0,14	-0,16	0,04	-0,11*	-0,12	-0,12	0,03	-0,11*	-0,09
M4 (%5 eksik veri)	a	1,08	1,07	1,09	1,07	1,07	1,07	1,08*	0,87	0,99	1,03
	b_1	-3	-2,99*	-3,01*	-3,07	-3,02	-3,03	-3,01*	-3,59	-3,13	-3,11
	b_2	-1,17	-1,15	-1,09	-1,23	-1,18*	-1,19	-1,18*	-1,41	-1,21	-1,2
	b_3	0,81	0,84	0,74	0,94	0,81*	0,81*	0,81*	1,08	0,9	0,98
M5 (%0 eksik veri)	a	2,59	2,61*	2,73	2,73	2,61*	2,67	2,61*	2,38	2,5	2,49
	b_1	-2,42	-2,4	-2,27	-2,37	-2,39	-2,39	-2,39	-2,55	-2,45	-2,41*
	b_2	-1,6	-1,6*	-1,53	-1,59	-1,59	-1,59	-1,59	-1,68	-1,63	-1,6*
	b_3	-0,22	-0,23	-0,19	-0,23	-0,22*	-0,22*	-0,22*	-0,22*	-0,22*	-0,21

1: tam veri, 2: eksik veri, 3: silme, 4: ortalama atama, 5: yakın noktalar ortalama ataması, 6: yakın noktalar medyan ataması, 7: doğrusal değerleme, 8: noktada doğrusal eğilim, 9: regresyon atama, 10: beklenti maksimizasyonu
* tam veri parametrelerine en yakın olan değerler

Tablo 7'de yer alan bilgiler doğrultusunda maddelerin ayırıcılık parametrelerinin güçlük parametrelerine göre kayıp veriden ve kayıp veri yöntemlerinden daha fazla etkilenmiş olduğu görülmektedir. Ölçekte yer alan tüm maddelerin tam veri ile hesaplanan ayırıcılık indekslerinin en yakın kayıp verileri doğrusal değerleme yöntemiyle atanan veri setinde hesaplandığı belirlenmiştir. %20 kayıp veriye sahip birinci maddenin ayırıcılık indeksinin doğrusal değerleme yöntemi ile verilerin atandığı veri seti ile farkının 0,02; ikinci maddenin (%15 kayıp verili) farkının 0,01; üçüncü maddenin (%10 kayıp verili) ayırıcılık indeksleri arasındaki farkın 0,02; dördüncü maddenin (%5 eksik verili) ayırıcılık indeksi arasındaki farkın 0,00; beşinci maddenin (%0 kayıp verili) ayırıcılık indeksleri arasındaki farkın da 0,01 olduğu belirlenmiştir.

1000 örneklem büyüklüğünde de maddelerdeki kayıp veri oranının ve madde ayırıcılık gücünün kayıp veri ile baş etme yöntemlerine etkisi olduğu tespit edilmiştir. Aynı zamanda maddede kayıp veri olmasa da diğer maddelerdeki kayıp verinin madde parametreleri üzerinde etkisi olduğu belirlenmiştir.

3. 2500 örneklem büyüklüğünde kayıp veri ile baş etme yöntemlerine dayalı olarak test ve madde istatistikleri nasıl bir değişim göstermektedir?

Araştırma kapsamında son olarak 2500 kişilik örneklem büyüklüğünde tam veri seti ile birlikte 10 veri seti oluşturulmuştur. Veri setlerine yönelik önce betimsel istatistikler ve güvenirlik katsayıları hesaplanmış, sonuçlar Tablo 8'de gösterilmiştir.

Tablo 8. Betimsel İstatistikler ve Güvenirlik Katsayıları (n=2500)

Veri setleri	Ortalama	SS	Cronbach alfa güvenirlik	Marginal güvenirlik
Tam veri	17,140	2,550	0,766	0,735
Eksik veri	15,380	3,300	0,769	0,734
Silme	17,200	2,560	0,769	0,734
Ortalama atama	17,150	2,350	0,729	0,712
Yakın noktalar ortalama ataması	17,130	2,550*	0,769	0,735
Yakın noktalar medyan ataması	17,140*	2,550*	0,768	0,734
Doğrusal değerleme	17,140*	2,550*	0,767	0,734
Noktada doğrusal eğilim	17,150	2,370	0,730	0,725
Regresyon atama	17,130	2,540	0,766*	0,736*
Beklenti maksimizasyonu	17,140*	2,530	0,782	0,7412

* tam veri parametrelerine en yakın olan değerler

Tablo 8'de yer alan bilgiler doğrultusunda 2500 örneklem büyüklüğünde tam veri setinde ölçek maddelerine verilen cevapların ortalaması 17,14; standart sapması 2,55 olarak hesaplanmıştır. Örneklemin Cronbach alfa güvenirlik katsayısı 0,766; marginal güvenirlik katsayısı da 0,7351 olarak hesaplanmıştır. Kayıp verilere herhangi bir müdahalede bulunulmadan veri setinin %12'si kayıp veri durumundayken yapılan hesaplama sonucunda ortalama değerin gerçek değerden daha düşük kestirildiği (-1,76 fark) belirlenmiştir. Bununla birlikte eksik verilerle farklı baş etme yöntemlerine dayalı tam hâle getirilmiş veri setleri gerçekleştirilen hesaplamalar sonucunda ortalama değer, standart sapma ve güvenirlik katsayılarının tam veri seti ile büyük ölçüde benzerlik gösterdiği tespit edilmiştir.

Araştırmada 2500 kişilik örneklem setindeki maddelerin ayırıcılık ve güçlük parametreleri aşamalı tepki modeli ile kestirilmiş, sonuçlar Tablo 9'da gösterilmiştir.

Tablo 9. Aşamalı Madde Tepki Modeline Dayalı Kestirilen Ayırıcılık ve Güçlük İndeksleri (n=2500)

Madde no	Madde parametreleri	Veri setleri									
		1	2	3	4	5	6	7	8	9	10
M1 (%20 eksik veri)	a	4,2	4,09	4,34	2,82	4,18	4,17	4,19*	3,59	4,18	5,58
	b_1	-2,22	-2,23*	-2,24	-2,54	-2,23*	-2,23*	-2,23*	-2,39	-2,27	-2,16
	b_2	-1,74	-1,78	-1,71	-2,07	-1,74	-1,74*	-1,75	-1,96	-1,8	-1,7
	b_3	-0,62	-0,63	-0,66	-0,88	-0,62*	-0,62*	-0,62*	-0,57	-0,58	-0,57
M2 (%15 eksik veri)	a	3,55	3,23	3,39	2,42	3,58	3,49	3,57*	3,37	3,4	4,57
	b_1	-2,4	-2,47*	-2,42	-2,81	-2,43	-2,42*	-2,42*	-2,53	-2,49	-2,33
	b_2	-1,91	-1,97	-1,9*	-2,25	-1,92*	-1,92*	-1,93	-2,05	-1,89	-1,81
	b_3	-0,57	-0,59	-0,58	-0,8	-0,58	-0,59	-0,58	-0,57*	-0,54	-0,53
M3 (%10 eksik veri)	a	1,8	1,8*	1,8*	1,66	1,81	1,8*	1,8*	1,51	1,7	1,79
	b_1	-2,93	-2,94*	-2,95	-3,15	-2,98	-3	-3,01	-3,35	-3,04	-2,98
	b_2	-1,62	-1,62*	-1,6	-1,78	-1,63	-1,64	-1,64	-1,88	-1,67	-1,66
	b_3	-0,09	-0,09*	-0,14	0,1	-0,09*	-0,09*	-0,09*	0,11	-0,07	-0,04
M4 (%5 eksik veri)	a	0,99	1,03	1,07	1,03	0,99*	0,99*	0,99*	0,91	1,01	0,98
	b_1	-3,44	-3,34	-3,34	-3,39	-3,45*	-3,45*	-3,47	-3,73	-3,41	-3,49
	b_2	-1,41	-1,39	-1,36	-1,46	-1,41*	-1,41*	-1,41*	-1,59	-1,39	-1,46
	b_3	0,84	0,79	0,71	0,89	0,84*	0,84*	0,84*	0,97	0,81	0,95
M5 (%0 eksik veri)	a	2,32	2,35	2,4	2,39	2,31	2,32*	2,32*	2,13	2,23	2,21
	b_1	-2,54	-2,55	-2,48	-2,52	-2,54*	-2,54*	-2,54*	-2,66	-2,59	-2,57
	b_2	-1,63	-1,64	-1,61	-1,63*	-1,64	-1,64	-1,63*	-1,71	-1,66	-1,64
	b_3	-0,15	-0,16	-0,18	-0,16	-0,15*	-0,15*	-0,15*	-0,15*	-0,15*	-0,13

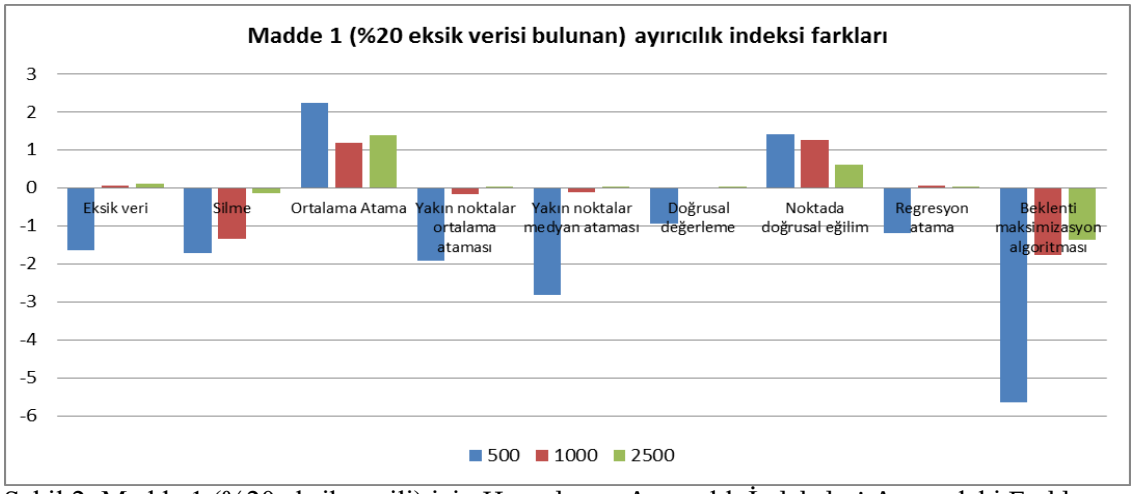
1: tam veri, 2: eksik veri, 3: silme, 4: ortalama atama, 5: yakın noktalar ortalama ataması, 6: yakın noktalar medyan ataması, 7: doğrusal değerleme, 8: noktada doğrusal eğilim, 9: regresyon atama, 10: beklenti maksimizasyonu

* tam veri parametrelerine en yakın olan değerler

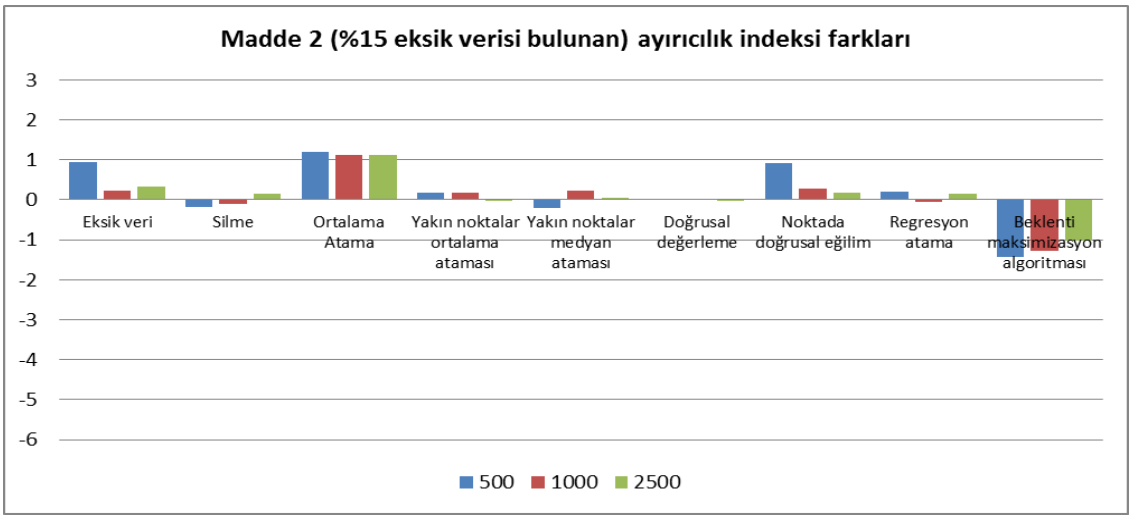
Tablo 9'da yer alan bilgiler doğrultusunda 2500 örneklem büyüklüğünde tam veri setinde hesaplanan madde ayırıcılık ve madde güçlük indekslerinin eksik veri setinde ve 8 kayıp veri ile baş etme yöntemleri ile elde edilen veri setlerinde büyük ölçüde benzerlik gösterdiği; farkların oldukça düşük olduğu tespit edilmiştir. Diğer örneklem büyüklüklerine benzer şekilde, tam veri setinden hesaplanan ayırıcılık indekslerine en yakın olan değerlerin doğrusal değerleme kayıp veri atama yöntemi ile elde edilen veri setlerine ait olduğu belirlenmiştir. Birinci maddede (%20 eksik verili) farkın 0,01; ikinci maddede (%15 eksik verili) farkın 0,02; üçüncü maddede (%10 eksik verili), dördüncü maddede (%5 eksik verili) ve beşinci maddede (%0 eksik verili) ise farkın olmadığı (0,00), madde ayırıcılık indeksinin bire bir aynı hesaplandığı belirlenmiştir.

4. Farklı örneklem büyüklüklerinde kayıp veri ile baş etme yöntemlerine dayalı olarak hesaplanan madde ayırıcılık indeksi ve tam veri setinden hesaplanan madde ayırıcılık indeksi farkları nedir?

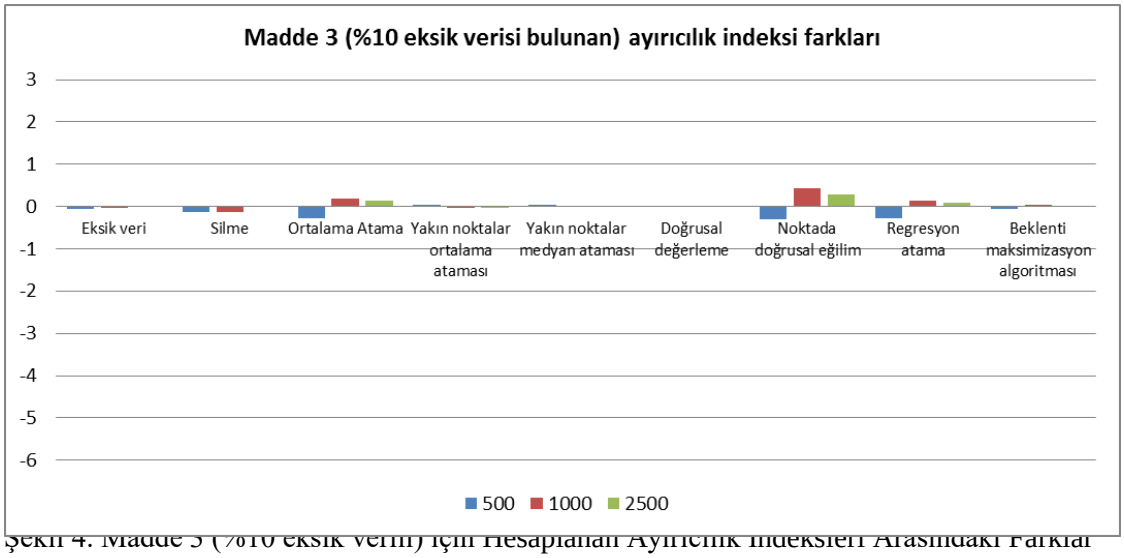
Araştırmada kayıp veriden ve kayıp veri ile baş etme yöntemlerinden en çok etkilenen değerin madde ayırıcılık parametresi olduğu belirlenmiştir. Bu doğrultuda her bir madde için eksik veri seti ve kayıp veri baş etme yöntemlerine dayalı oluşturulmuş veri setlerinin ayırıcılık indekslerinin tam veri setinde hesaplanan değerle arasındaki fark hesaplanmış; sonuçlar Şekil 2-Şekil 6'da gösterilmiştir.



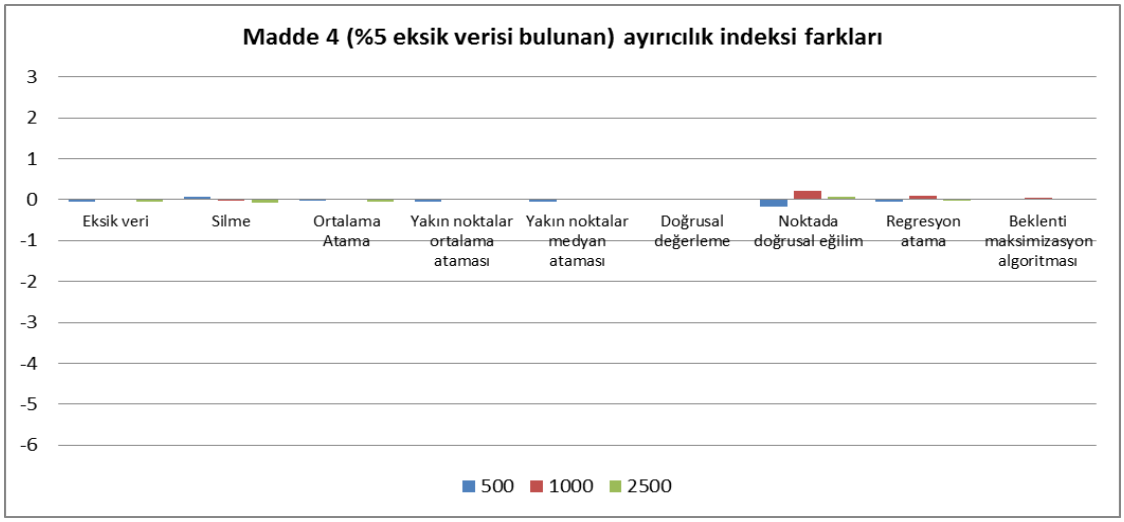
Şekil 2. Madde 1 (%20 eksik verili) için Hesaplanan Ayırıcılık İndeksleri Arasındaki Farklar



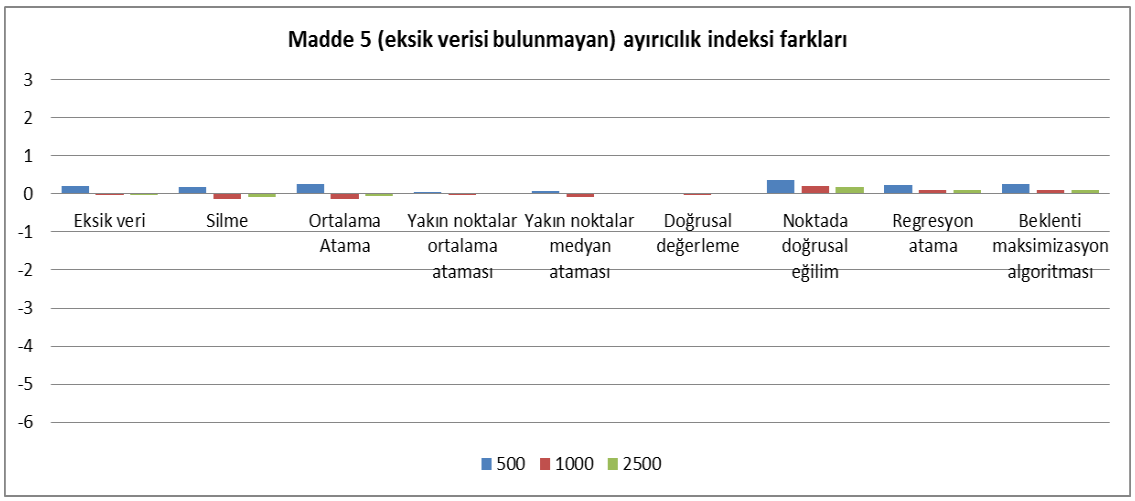
Şekil 3. Madde 2 (%15 eksik verili) için Hesaplanan Ayırıcılık İndeksleri Arasındaki Farklar



Şekil 4. Madde 3 (%10 eksik verili) için Hesaplanan Ayırıcılık İndeksleri Arasındaki Farklar



Şekil 5. Madde 4 (%5 eksik verili) için Hesaplanan Ayırıcılık İndeksleri Arasındaki Farklar



Şekil 6. Madde 5 (eksik verisi bulunmayan) için Hesaplanan Ayırıcılık İndeksleri Arasındaki Farklar

Şekil 2-6 incelendiğinde tüm maddelerde örneklem büyüklüğüne bağlı olarak tam veri ile eksik veri ve kayıp veri ile baş etme yöntemlerine dayalı olarak oluşturulan veri setlerinde hesaplanan ayırıcılık indeksleri arasındaki farkların azalma gösterdiği belirlenmiştir. Maddelerdeki kayıp veri oranları azalma gösterdikçe ayırıcılık indeksleri arasındaki farkın da azalma gösterdiği tespit edilmiştir. Bununla birlikte ölçek maddeleri içerisinde en yüksek ayırıcılığa birinci madde; en düşük ayırıcılığa ise dördüncü madde sahiptir. Düşük ayırıcılıktaki maddeler arasındaki fark da ranjı daha küçük olduğu için daha az hesaplanmıştır. Beşinci maddede hiç kayıp veri olmamasına rağmen madde ayırıcılık indekslerinde ufak da olsa farklılaşmalar saptanmıştır.

SONUÇLAR ve TARTIŞMA

Bu araştırma kapsamında PISA 2015 uygulamasına katılan ve çalışma içerisinde yer alan "hırs algısı ölçek maddelerine" cevap veren 5073 öğrenci arasından rastgele 500, 1000 ve 2500 büyüklüğünde örneklemler seçilmiştir. Hırs algısı ölçeğinde tek boyutta toplandığı belirlenen 4'lü likert tipinde derecelendirilmiş 5 madde yer almaktadır. Her bir örneklem büyüklüğünde birinci maddenin %20'si, ikinci maddenin %15'i, üçüncü maddenin %10'u ve dördüncü maddenin %5'ine denk gelen veriler silinmiştir. Beşinci madde ise eksik veri içermeyecek şekilde aynen bırakılmıştır. Allison (2002) ile Tabachnick ve Fidell (2014) kayıp verinin TRK olduğu durumda, veri setindeki kayıp veri oranının %5 ve altında olması hâlinde test ve madde istatistiklerinde önemli sorunlar yaşanmayacağını, ancak kayıp veri oranının %5'in üzerine çıkması durumunda sonuçların yanlı olarak kestirilebileceğini belirtmişlerdir. Bu doğrultuda bu araştırmada ele alınan veri setlerinde %12'lik kayıp veri oluşturulmuştur. Veri setlerindeki eksik verilerin mekanizmasının tamamen rastsal olduğu Little´ın MCAR testi ile test edilmiştir. Her bir örneklem büyüklüğünde tam veri, eksik veri ve silme, ortalama atama, yakın noktalar ortalama ataması, yakın noktalar medyan ataması, doğrusal değerleme, noktada doğrusal değerleme, regresyon atama ve beklenti maksimasyonu algoritması atama yöntemleriyle eksik verilerin tamamlanması sonucunda 10'ar olmak üzere toplam 30 veri seti oluşturulmuştur. Veri setlerinde önce test istatistikleri (ortalama, standart sapma, Cronbach alfa ve marginal güvenirlik katsayıları) ardından madde ayırıcılık ve madde güçlük parametreleri Aşamalı Tepki Modeline dayalı olarak hesaplanmıştır.

Hesaplamalar sonucunda eksik veri ve eksik verilerle baş etme yöntemlerine dayalı tam hâle getirilmiş veri setlerine ait madde istatistiklerinin test istatistiklerine göre daha fazla etkilendiği belirlenmiştir. Tam veri setleri ölçüt (referans) alınarak karşılaştırmalar yapılmış; eksik verilerle baş etme yöntemlerine dayalı tam hâle getirilmiş veri setlerinde test istatistiklerinin tam veri seti ile büyük ölçüde benzerlik gösterdiği saptanmıştır. Ancak eksik veri seti ile hesaplanan ortalama değerlerin tam veri setinden elde edilen ortalama değerinden düşük olduğu belirlenmiştir. Araştırma sonuçları ve bu bilgiler doğrultusunda kayıp veri ile baş etme yöntemlerinin testin betimsel istatistiklere olan etkisinin yordayıcı istatistiklere olan etkisinden daha az olduğu söylenebilir. Eksik veri ve kayıp veri ile baş etme yöntemlerinin kullanılmasıyla elde edilen veri setlerine ait madde parametreleri incelendiğinde, madde ayırıcılık indeksinin madde güçlük indekslerinden daha fazla etkilendiği tespit edilmiştir. Maddedeki kayıp veri oranının ve maddenin ayırıcılık indeksinin ayrıca kayıp veri ile baş etme yöntemleri üzerinde etkisi olduğu belirlenmiştir. Koçak ve Çokluk Bökeoğlu (2017) çalışmalarında MTK´ya dayalı test istatistiklerini karşılaştırmış, tüm setlerde kayıp veri oranı artış gösterdikçe kayıp veri ile baş etme yöntemleri sonuçları arasındaki farkın artış gösterdiğini tespit etmiştir.

Bu araştırmada madde parametreleri Aşamalı Tepki Modeline dayalı olarak kestirilmiştir. Ölçeğin beşinci maddesinde hiç kayıp veri bulunmamasına karşın madde ayırıcılık ve güçlük indeksinin veri setleri bazında değişiklik gösterdiği saptanmıştır. Bu durum, eksik verisi bulunan maddelerin diğer maddelerin parametrelerine de etki ettiğini göstermektedir. Bu nedenle araştırmacıların kayıp veri ile karşılaştığı durumda veri setini inceleyerek kayıp verisi için en uygun baş etme yöntemini kullanması gerekmektedir. Araştırma sonucu dikkate alınarak eksik verisi olmayan maddelerin de süreçten etkilendiği göz önünde bulundurulmalıdır. Garrett (2009) yapmış olduğu çalışmasında

değişen madde fonksiyonun kayıp veri olduğu durumda referans ve odak grup kestirimlerinin benzerlik gösterdiğini belirlemiştir. Ergün (2013) 0-1 veri setiyle çalıştığı araştırmasında kayıp veri yöntemlerinin madde ayırıcılık ve güvenirlik katsayılarına etki ettiğini; en çok olabilirlik ve çoklu veri atama yöntemlerinde madde parametre kestirimlerinin benzer sonuçlar ürettiğini tespit etmiştir.

Örneklem büyüklüğü artış gösterdikçe kayıp veri ile baş etme yöntemlerinin madde ve test parametreleri üzerindeki etkisinin azaldığı, gerçek değere oldukça yaklaştığı belirlenmiştir. Soysal ve Akın Arıkan (2017) tamamen rastsal kayıp verilerle gerçekleştirdikleri açımlayıcı faktör analizi sonucunda örneklem büyüklüğünün artış gösterdikçe hata oranının da azalma gösterdiğini tespit etmişlerdir.

Araştırma sonucunda her bir örneklem büyüklüğünde tam veri setine en yakın madde ayırt edicilik indeksi sunan sonuçların doğrusal değerleme yöntemi ile elde edildiği belirlenmiştir. Kayıp verinin kendisinden bir önceki ve bir sonraki veri ortalaması alınarak hesaplandığı bu yöntemin genellikle boylamsal veriler ya da ardışık kategorilerin olduğu veri setlerinde kullanıldığı belirlenmiştir. Bu araştırmalarda kayıp verilerle baş etme amacıyla doğrusal değerleme yöntemi kullanılmış ve tam veri setine yakın sonuçlar elde edilmiştir. (Junninen, Niska, Tuppurainen, Ruuskanen ve Kolehmainen, 2004; Kohn ve Ansley, 1986; Twisk ve de Vente, 2002). Yapılan bu çalışmada da likert tipinde derecelendirilmiş bir ölçek kullanılmıştır ve ölçek dereceleri ardışıklık özelliği göstermektedir. Aşamalı Tepki Modeli ile kestirilen eşik parametresi (bij), bireylerin örtük özellik boyunca 0,50 olasılıkla bir kategori eşiğinin üzerinde tepki verme olasılığını tanımlar. bij, aynı zamanda bireylerin örtük özellik boyutunda o kategoride ya da daha üst kategorilerde 0,50 tepki verme olasılığının da bir tanımıdır ve bu nedenle eşik parametreleri sıralı özellik göstermektedir. Eğim parametresi olan a_i de sıralı kategorilerin tamamı için kestirilmektedir (Emretson ve Reise, 2000; Ostini ve Nering, 2006). Bu bağlamda araştırmada kayıp gözlemlere doğrusal değerleme yöntemi ile atanan veriler sonucunda hesaplanan ayırıcılık ve güçlük değerlerinin tam veri setine yakın hesaplanması, cevap kategorilerinin kendi içinde ardışıklık göstermesinden kaynaklanmış olabilir. Bu doğrultuda başka çalışmalarda bu yöntemin yinelenmesi, iki kategorili veri setlerinde de uygulanarak sonuçlarının karşılaştırılması önerilmektedir.

Araştırma kapsamında doğrusal değerleme dışında da kayıp veri ile baş etme yöntemleri kullanılmış ve tam veri seti ile, özellikle küçük örneklem büyüklüklerinde, benzer olmayan ayırıcılık indeksleri hesaplanmıştır. Çokluk ve Kayri (2011) de yaklaşık değer atama yöntemlerinin test maddelerinin açıklanan varyans oranında ve güvenirlikte düşüş meydana getirdiğini belirlemiştir. Bernaards ve Sijtsma (2000) kayıp veri yöntemleri içerisinde beklenti maksimizasyonu yönteminin diğer yöntemlerden daha iyi sonuç verdiği bilgisine ulaşmıştır. Sezgin ve Çelik (2013) araştırmalarında kayıp veri ile baş etme sürecinde durum düzeyinde silme yönteminin kullanılmasının basit olmasına rağmen veri kaybı yaşattığı için varyansı değiştirdiğini; regresyon atamasının da özellikle ilişkisizlik durumunda hata oluşumuna neden olduğunu belirtmektedir. Garlett (2009) de silme yöntemi ile değişen madde fonksiyonunun farklılık gösterdiğini ortaya koymuştur. Köse ve Öztemur (2014) da araştırmalarında silme yönteminin özellikle t testi sonuçlarına etki ettiğini, sonuçların yanlı kestirilmesine neden olduğunu ortaya koymuşlardır. Acock (2005), silme yönteminin TRK dışında özellik gösteren kayıp veriler için hiç uygun olmadığını; TRK özellik gösteren kayıp veri setlerinde de özellikle küçük örneklem büyüklüklerinde testin istatistiksel gücünde bir azalma meydana getirdiğini ifade etmektedir. Nartgün (2015) araştırmasında silme yönteminin, tüm koşullarda, tam veri setinden elde edilen değere en uzak değerleri verdiğini, en yakın değerlerin ise çoklu veri atama ve regresyon ataması yöntemleriyle sağlandığını belirlemiştir. Bu çalışmada örneklem büyüklüğü artış gösterdikçe kayıp veri ile baş etme yöntemleri arasındaki farklar benzerlik gösterse de tüm örneklem büyüklüğünde ve kayıp veri oranlarında tam veri setine en yakın sonuçların doğrusal değerleme yönteminin kullanıldığı koşullarda ulaşıldığı tespit edilmiştir. Diğer araştırma bulgularına benzer şekilde diğer eksik verilerle baş etme yöntemlerinin tam veri seti ile benzerlik gösterdiği; ancak söz konusu çalışmaların çoğunda doğrusal değerleme yöntemi kullanılmadığı için başka çalışmalarda da doğrusal değerleme yönteminin incelenmesi ve özellikle bu yöntemin küçük örneklem gruplarından elde edilen veri setleri üzerinde kullanılması önerilmektedir.

Bu araştırmada sekiz farklı kayıp veri ile baş etme yönteminin madde ve test parametreleri kestirimindeki performansları karşılaştırılmıştır. Başka çalışmalarda eksik veriler yerine atama ya da silme yöntemleri yerine kayıp verilerin de analizin içine dahil edildiği tam bilgi en çok olabilirlik (full information maximum likelihood) gibi yöntemler ile de hesaplamalar yapılıp elde edilen sonuçlar karşılaştırılabilir.

KAYNAKÇA

Acar, T. ve Kelecioğlu, H. (2010). Maddenin farklı fonksiyonlaşmasını belirleme tekniklerinin karşılaştırılması: GADM, LR ve MTK-OO. *Kuram ve Uygulamada Eğitim Bilimleri, 10*(2), 639-649.

Acock, A. (2005). Working with missing values. *Journal of Marriage and Family, 67*(4), 1012-1028.

Akbaş, U. ve Tavşancıl, E. (2015). Farklı örneklem büyüklüklerinde ve kayıp veri örüntülerinde ölçeklerin psikometrik özelliklerinin kayıp veri baş etme teknikleri ile incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 6*(1), 38-57.

Allison, P. D. (2002). *Missing data*. California: Sage Publication, Inc.

Ambler, G., Omar, R. Z., & Royston, P. (2007). A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Statistical Methods in Medical Research, 16*(3), 277-298.

Bal, C. (2003). *Çok gruplu veri setlerinde eksik gözlem sorununun çözümlenmesi ve sağlık alanında bir uygulama* (Yayımlanmamış Yüksek Lisans Tezi, İstanbul Üniversitesi, Sağlık Bilimleri Enstitüsü, İstanbul). Erişim adresi: http://tez2.yok.gov.tr

Baygül, A. (2007). *Kayıp veri analizinde sıklıkla kullanılan etkin yöntemlerin değerlendirilmesi.* (Yayımlanmamış Doktora Tezi, Osmangazi Üniversitesi, Sağlık Bilimleri Enstitüsü, Eskişehir). Erişim adresi: http://tez2.yok.gov.tr

Bernaards, C. A., & Sijtsma, K. (2000). Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research, 35*(3), 321 – 364.

Carpita, M., & Manisera, M. (2011). On the imputation of missing data in surveys with Likert-type scales. *Journal of Classification, 28*(1), 93-112.

Çokluk, Ö., Şekercioğlu, G. ve Büyüköztürk, Ş. (2014). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları*. Ankara: Pegem Akademi.

Çokluk, Ö. ve Kayri, M. (2011). Kayıp değerlere yaklaşık değer atama yöntemlerinin ölçme araçlarının geçerlik ve güvenirliği üzerindeki etkisi. *Kuram ve Uygulamada Eğitim Bilimleri, 11*(1), 289-309.

Çüm, S. ve Gelbal, S. (2015). Kayıp veriler yerine yaklaşık değer atamada kullanılan farklı yöntemlerin model veri uyumu üzerindeki etkisi. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi,* 35, 87-111.

De Ayala, R. J., Plake, B. S., & Impara, J. C. (2011). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement, 38*(3), 213-234.

De Leeuw, E. D., Hox, J., & Huisman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics, 19*(2), 153-176.

Demir, E. (2013). Kayıp verilerin varlığında çoktan seçmeli testlerde madde ve test parametrelerinin kestirilmesi: SBS örneği. *Eğitim Bilimleri Araştırmaları Dergisi, 3*(2), 48-68.

Demir, E. (2013). *Kayıp verilerin varlığında iki kategorili puanlanan maddelerden oluşan testlerin psikometrik özelliklerinin incelenmesi*. (Yayımlanmamış Doktora Tezi. Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara). Erişim adresi: http://tez2.yok.gov.tr

Demir, E. ve Parlak, B. (2012). Türkiye'de eğitim araştırmalarında kayıp veri sorunu. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 3*(1), 230-241.

Doğanay-Erdoğan, B. (2012). *Çoklu atama yöntemlerinin Rasch modelleri için performansının benzetim çalışması ile incelenmesi*. (Yayımlanmamış Doktora Tezi, Ankara Üniversitesi, Sağlık Bilimleri Enstitüsü, Ankara). Erişim adresi: http://tez2.yok.gov.tr

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

Field, A. (2005). *Discovering statistics with SPSS*. California: Sage Publication, Inc.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement, 45*(3), 225-245.

Garrett, P. L. (2009). *A Monte Carlo study investing ating missing data, differential item functioning and effect size* (Yayımlanmamış doktora tezi, Georgia State University, Atlanta). Erişim adresi: https://scholarworks.gsu.edu.

Garson, D. (2015). *Missing data analysis & data imputation*. Asheboro: Statistical Publishing Associates.

Goegebeur, Y., De Boeck, P., & Molenberghs, G. (2010). Person fit for test speededness: Normal curvatures, likelihood ratio tests and empirical Bayes estimates. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 6*(1), 3-16.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. (1977). Developments in latent trait theory: models, technical issues, and applications. *Review of Educational Research, 48*(4), 467–510.

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

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2007). Analysis of binary outcomes with missing data: Missing = smoking, last observation carried forward, and a little multiple imputation. *Addiction, 102*(10), 1564-1573.

Hohensinn, C., & 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.

Josse, J., & Husson, F. (2012). Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique, 153*(2), 79-99.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment, 38*, 2895–2907.

Koçak, D. ve Çokluk Bökeoğlu, Ö. (2017). Kayıp veriyle baş etme yöntemlerinin model veri uyumu ve madde model uyumuna etkisi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 8*(2), 200-223.

Kohn, R., & Ansley, C. F. (1986). Estimation, prediction, and interpolation for ARIMA models with missing data. *Journal of the American Statistical Association, 81*, 751-761.

Köse, A. ve Öztemur, B. (2014). Kayıp veri ele alma yöntemlerinin t-testi ve ANOVA parametreleri üzerine etkisinin incelenmesi. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 14*(1), 400-412.

McNeish, D. (2016). Exploratory factor analysis with small samples and missing data. *Journal of Personality Assessment, 99*(6), 637-652.

Nartgün, Z. (2015). Kayıp veri sorununun çözümünde kullanılan farklı yöntemlerin farklı kayıp veri koşulları altında ölçeklerin psikometrik nitelikleri ve ölçme sonuçları bağlamında karşılaştırılması. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 7*(4), 252-265.

Ostini, R., & Nering, M. L. (Eds.). (2006). *Polytomous item response theory models*. U.S.A: Sage Publication.

Peng, C. Y., Harwell, M. R., Liou, S. M., & Ehman, L. H. (2007). Advances in missing data methods and implications for educational research. In S. S. Sawilowsky (Ed.), *Real data analysis* (pp. 31-78). Charlotte, NC: New Information Age.

Rose N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)* (Report No: RR-10-11). Princeton: ETS Research Report Series.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147 – 177

Sezgin E. ve Çelik Y. (2013). *Veri madenciliğinde kayıp veriler için kullanılan yöntemlerin karşılaştırılması,* Akademik Bilişim Konferansı, Akdeniz Üniversitesi, Antalya.

Soysal, S. ve Akın Arıkan, Ç. (2017). Kayıp veri atama yöntemlerinin faktörleşme teknikleri üzerindeki etkisi. PegemA Yayıncılık (Ed.), *Küreselleşen Dünyada Eğitim*, (syf. 316-332) Ankara: PegemA Yayıncılık.

Şahin Kürşad, M. ve Nartgün, Z. (2015). Kayıp veri sorununun çözümünde kullanılan farklı yöntemlerin ölçeklerin geçerlik ve güvenirliği bağlamında karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 6*(2), 254-267.

Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics*. USA: Pearson Education Limited.

Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55*, 371–390.

Twisk, J., & de Vente, W. (2002). Attrition in longitudinal studies: How to deal with missing data. *Journal of Clinical Epidemiology, 55*, 329–337.

Weaver, B., & Maxwell, H. (2014). Exploratory factor analysis and reliability analysis with missing data: A simple method for SPSS users. *The Quantitative Methods for Psychology, 10*(2), 143-152.

Witta, E. L. (2000). C*omparison of missing data treatments in producing factor scores* (Report No:1). Honolulu: Annual Meeting of the American Educational Resarch Association.

Wright, B. D. (1997). A History of social science measurement. *Educational measurement: Issues and practice, 16*, 33–45.

Yılmaz, H. (2014). *Random forests yönteminde kayıp veri probleminin incelenmesi ve sağlık alanında bir uygulama*. (Yayımlanmamış Yüksek Lisans Tezi, Eskişehir Osmangazi Üniversitesi, Sağlık Bilimleri Enstitüsü, Eskişehir). Erişim adresi: http://tez2.yok.gov.tr

EXTENDED ABSTRACT

Introduction

One of the major problems in the literature in social and behavioural sciences is missing data. Since large samples are needed especially in items response theories, missing data can be seen in the data sets. In this study, item and test parameters were estimated according to the graded response theory, which is one of the item response theory (IRT) models. In the cases of missing data, researchers generally prefer to listwise or pairwise deletion techniques for the missing data cases from the analysis. However, this situation effects the validity and reliability of the results (Akbaş ve Tavşancıl, 2015; Baygül, 2007; Çüm ve Gelbal, 2015; Kaspar, 2011; Kürşad, 2014). And also different methods have been developed in order to solve the missing data problem in the studies so there are a lot of techniques of handling missing data.

Unlike other studies, in this study the missing data rate is not based on the test, but it is item-based. In this direction, the effect of the amount of missing data on any item to other non-missing item parameters is examined. The other importance of the study is that this study is made on the basis of items and the estimations are done by using the graded response model.

So in this study, the aim is to determine how the item and test parameters are affected by the missing data techniques for different sample sizes and different items with different the missing data rates.

Method

PISA data was used to achieve the purpose of the study. In the PISA study, there is an ambition perception scale with 5 items. All items are rated in 4-point likert type and there isn't any reverse item. In the study in 2015, there are a total of 5073 Turkish students responding to all of the scale items. 500, 100 and 2500 students randomly selected from 5073 students. First of all, the assumptions of normality, uni-dimensionality, local independence and model-data fit are examined for each data set. Afterwards, 5%, 10%, 15% and 20% missing data were formed from for the items of the scale and there is no missing data in one (fifth) item. Tabachnick and Fidell (2014) indicate that the missing data above 5% of the test will lead to biased estimates. In the scope of the research, 12% of data sets have missing data.

First, whether the consistency of data had missing completely at random (MCAR) mechanism was also controlled with Little's MCAR test. Then 10 data sets were obtained for each sample. Complete data, incomplete data and 8 data sets which were obtained by missing data techniques including series mean, median of nearby points, mean of nearby points, linear interpolation, linear trend at point, regression.

For the analysis, descriptive statistics and Cronbach alpha reliability coefficient and marginal reliability coefficient were estimated for each sample. And then the threshold parameters and the difficulty indices were estimated according to the graded response theory from the item response theory models. Estimates were done primarily for each sample size, the parameters of the items with different missing data rates were compared. The results were then compared in terms of sample sizes.

Results and Discussion

As a result of the study, it was determined that the item statistics were affected from missing data and missing data techniques, which was more than the test statistics. When only the listwise deletion method was used, the mean value was estimated lower than the actual value.

It was found that the item threshold parameters were affected from missing data and missing data techniques for each sample size, which was more than the item difficulty indices. Sample size, missing data techniques and value of the threshold index had an effect on the threshold index.

In this study, item parameters were estimated based on the item response theory. Item response theory assumes that the item parameters are independent of each other. But this study showed that last items which belong to complete data (no missing data) were affected by other items and missing data techniques.

In the study, it was determined that as the sample size increased, the difference between item parameters in terms of the missing data techniques decreased. For this reason, the missing data technique is not important in large samples (ex. n=2500).

The results of the study showed that the item and test parameters were influenced by incomplete and missing data and missing data techniques; it was determined that the best estimation results were obtained by linear interpolation method with different data.

In this study, the results of 8 different missing data techniques were compared. For the future research, the full information maximum likelihood method can also be used and the results can be compared. By using full information maximum likelihood method, missing values are not replaced or imputed, but the missing data is handled within the model.