Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN:1309-6575 *Kış* 2015 *Winter* 2015 *Cilt: 6- Sayı: 2* *Volume: 6- Issue: 2*

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 iki kez yayınlanan hakemli ulusal bir dergidir. Yayımlanan yazıların tüm sorumluğu ilgili yazarlara aittir.

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

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

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

Adnan KAN (Gazi Üni.)
Ahmet TURAN (Pearson)
Ali BAYKAL (Bahçeşehir Üni.)
Adnan ERKUŞ (Emekli Öğrtim Üyesi)
Arif ÖZER (Hacettepe Üni.)
Ayfer SAYIN (Gazi Üni.)
Aylin ALBAYRAK SARI (Hacettepe Üni.)
Ayşegül ALTUN (Ondokuz Mayıs Üni.)
Bayram BIÇAK (Akdeniz Üni.)
Bayram ÇETİN (Gazi Üni.)
Bilge BAŞUSTA UZUN (Mersin Üni.)
Bilge GÖK (Hacettepe Üni.)
Burak AYDIN (Recep Tayyip Erdoğan Üniversitesi)
Burcu ATAR (Hacettepe Üni.)
Burhanettin ÖZDEMİR (Siirt Üni.)
Cem Oktay GÜZELLER (Hacettepe Üni.)
Cindy M. WALKER (Duquesne University)
David KAPLAN (University of Wisconsin)
Deniz GÜLLEROĞLU (Ankara Üni.)
Derya ÇAKICI ESER (Kırıkkale Üni)
Derya ÇOBANOĞLU AKTAN (Hacettepe Üni.)
Dilara BAKAN KALAYCIOĞLU (ÖSYM)
Dilek GENÇTANRIM (Kırşehir Ahi Evran Üni.)
Durmuş ÖZBAŞI (Çanakkele Onsekiz Mart Üni)
Duygu GÜNGÖR (İzmir Üni.)
Elif Bengi ÜNSAL ÖZBERK (Adalet Bakanlığı)
Emine ÖNEN (Gazi Üni.)
Emrah GÜL (Hakkari Üni.)
Emre ÇETİN (Doğu Akdeniz Üni.)
Ergül DEMİR (Ankara Üni.)
Esin TEZBAŞARAN (İstanbul Üni)
Esin YILMAZ KOĞAR (Hacettepe Üni.)
Esra Eminoğlu ÖZMERCAN (MEB)
Evrim ÇETİNKAYA YILDIZ (Erciyes Üniversitesi)
Fatih KEZER (Kocaeli Üni.)
Fatih ORCAN (Karadeniz Teknik Üni)
Fatma BAYRAK (Hacettepe Üni.)
Fazilet TAŞDEMİR (Recep Tayyip Erdoğan Üni.)
Funda NALBANTOĞLU YILMAZ (Nevşehir Üni.)
Göksu GÖZEN (Mimar Sinan Güzel Sanatlar Üni.)
Gülden KAYA UYANIK (Sakarya Üni.)
Gülşen TAŞDELEN TEKER (Sakarya Üni.)
Hakan KOĞAR (Akdeniz Üni.)
Hakan Yavuz ATAR (Gazi Üni.)
Halil YURDUGÜL (Hacettepe Üni.)
Hatice KUMANDAŞ (Artvin Çoruh Üni.)
Hülya KELECİOĞLU (Hacettepe Üni.)
Hüseyin SELVİ (Mersin Üni.)
İbrahim Alper KÖSE (Abant İzzet Baysal Üni.)
İlker KALENDER (Bilkent Üni.)
İsmail KARAKAYA (Gazi Üni.)
Kaan Zülfikar DENİZ (Ankara Üni.)
Kübra ATALAY KABASAKAL (Hacettepe Üni.)
Mehmet KAPLAN (MEB)
Meltem ACAR GÜVENDİR (Trakya Üni.)
Mustafa ASİL (University of Otago)
Nagihan BOZTUNÇ ÖZTÜRK (Hacettepe Üni.)
Neşe GÜLER (Sakarya Üni.)
Neşe ÖZTÜRK GÜBEŞ (Mehmet Akif Ersoy Üni.)
Nuri DOĞAN (Hacettepe Üni.)
Nükhet DEMİRTAŞLI (Ankara Üni.)
Okan BULUT (University of Alberta)
Onur ÖZMEN (TED Üniversitesi)
Ömer KUTLU (Ankara Üni.)
Recep Serkan ARIK (Dumlupınar Üni.)
Sakine GÖÇER ŞAHİN (Hacettepe Üvi.)
Sedat ŞEN (Harran Üni.)
Seher YALÇIN (Ankara Üni.)
Selahattin GELBAL (Hacettepe Üni.)
Sema SULAK (Bartın Üni.)
Serdar ÇAĞLAK (Osmangazi Üniveristesi)
Seval KIZILDAĞ (Adıyaman Üni.)
Sevda ÇETİN (Hacettepe Üni.)
Sevilay KİLMEN (Abant İzzet Baysal Üni.)
Şeref TAN (Gazi Üni.)
Şeyma UYAR (Mehmet Akif Ersoy Üni.)
Tahsin Oğuz BAŞOKÇU (Ege Üni.)
Terry A. ACKERMAN (University of North Carolina)
Tülin ACAR (Parantez Eğitim)
Türkan DOĞAN (Hacettepe Üni.)
Yavuz AKPINAR (Boğaziçi Üni.)
Yeşim ÖZER ÖZKAN (Gaziantep Üni.)
Zekeriya NARTGÜN (Abant İzzet Baysal Üni.)

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

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2015; 6(2); 279-292

Türkiye'deki Araştırma Görevlilerinin Mesleki Sorunlarının İkili Karşılaştırma Yoluyla Ölçeklenmesi

Scaling of Research Assistants' Professional Problems in Turkey with Paired-Wise Comparison Method

Duygu ANIL * Levent ERTUNA ** İbrahim UYSAL***

Öz

Bu araştırma, araştırma görevlilerinin yaşadıkları mesleki sorunların önem düzeyini "karşılaştırmalı yargılar kanunu" kapsamında yer alan ikili karşılaştırma yoluyla ölçeklemeyi amaçlamaktadır. Bu kapsamda araştırma görevlilerinin mesleki sorunları cinsiyet, öğrenim durumu ve kadro türü değişkenlerine göre incelenmiştir. Araştırma genel tarama modeli ile desenlenmiştir. Araştırmanın örneklemini Türkiye'de Yüksek Öğretim Kurumuna bağlı devlet ve vakıf üniversitelerinde görev yapan 555 araştırma görevlisi oluşturmaktadır. Ölçme aracı iki kısımdan oluşmakta olup birinci kısımda demografik bilgiler (cinsiyet, öğrenim seviyesi ve kadro türü), ikinci kısımda ise araştırma görevlilerinin ikili olarak karşılaştıracakları sekiz mesleki sorun (akademik olmayan işlere yönlendirilme, fiziksel yetersizlikler, mobbing, yabancı dil sorunu, kadro güvencesi olmaması, ekonomik sorunlar, ödeneklerin yetersizliği ve idare tarafından verilen fakülte işlerinin yoğunluğu) bulunmaktadır. Ölçekleme Thurstone'un karşılaştırmalı yargılar kanununun III. Hal denklemi kullanılarak tam veri matrisinden gerçekleştirilmiştir. Araştırmanın sonucunda elde edilen bulgulara göre araştırma görevlilerinin en önemli sorunu ekonomiktir. Bunu sırasıyla mobbing, idare tarafından verilen fakülte işlerinin yoğunluğu, ödeneklerin yetersizliği ya da bulunmaması, kadro güvencesi olmaması, akademik olmayan işlere yönlendirme, fiziksel yetersizlikler ve yabancı dil sorunları takip etmiştir. Araştırma görevlilerinin karşılaştıkları mesleki sorunlar cinsiyet değişkenine göre incelendiğinde kadın araştırma görevlileri mobbingi en önemli sorun olarak görürken erkek araştırma görevlileri bu sorunu altıncı sırada görmektedir. Diğer bir değişken olan öğrenim seviyesinde, doktora eğitimini tamamlamış araştırma görevlilerinin en önemli sorununun kadro güvencesinin olmaması olduğu görülmüştür. Kadro türü değişkeni açısından ise 50. maddenin d bendince atanan araştırma görevlilerinin en önemli sorunu kadro güvencesinin olmamasıdır.

Anahtar Kelimeler: Araştırma görevlisi mesleki sorunları, ikili karşılaştırma yöntemi, III. Hal denklemiyle ölçekleme.

Abstract

In this study, the significance level of research assistants' professional problems was measured with the paired comparison method within the scope of "law of comparative judgment". In this context, research assistants' professional problems are investigated for gender, educational background and type of position. This study was carried out in the form of general survey model. The study group consisted of 555 research assistants working in public or foundation universities which are accredited by the Council of Higher Education in Turkey. The data collection tool is composed of two parts. The first parts asks for demographic information about gender, educational background and type of position while the second part inquires about the 8 professional problems (being directed to non-academic duties, lack of physical equipment, mobbing, foreign language proficiency, lack of guaranteed positions, financial problems, insufficient funds and heavy faculty works) given by the administration. The data was scaled via the third conditional equation on the full data matrix of Thurstone's law of comparative judgment. The results of the study indicated that the most significant problem research assistants face was economic problems. They were followed by mobbing, faculty workload, insufficient

* Doç. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, e-posta: aduygu@hacettepe.edu.tr
** Arş. Gör., Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara-Türkiye, e-posta: leventertuna@gmail.com
*** Arş. Gör., Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara-Türkiye, e-posta: ibrahimuysal06@gmail.com

Geliş Tarihi: 24.07.2015
Kabul Tarihi: 30.11.2015

funding, job security, non-academic duties, physical deficiencies and poor foreign language skills respectively.When the research assistants' problems were analyzed in terms of gender variable, there was a meaningful difference in the scale values of mobbing. In fact, male research assistants puts mobbing in the sixth place while female research assistants regard it as the most important problem. Another variable is the educational background, it can be seen from the results that the most significant problem of those who have obtained their PhD degrees is lack of guaranteed positions. When research assistants' problems were examined in terms of type of position, it was found that lack of guaranteed positions is the most important problem for research assistants of 50/d position.

Keywords: Research asistants' problems, paired comparison method, the third conditional equation.

GİRİŞ

Araştırma görevlileri 2547 sayılı kanunun 33. maddesinde "yükseköğretim kurumlarında yapılan araştırma, inceleme ve deneylerde yardımcı olan ve yetkili organlarca verilen ilgili diğer görevleri yapan öğretim yardımcıları" olarak tanımlanmaktadır (YÖK, 1981). Yasada yer alan "yetkili organlarca verilen diğer görevleri yapar" ifadesi nedeniyle araştırma görevlilerinin yapacakları işler net olarak bilinmemektedir. Bu nedenle araştırma görevlilerine eğitim, öğretim ve araştırma faaliyetlerinde, bunun yanı sıra kurumsal birçok işte görev verilmektedir. Acar, Nemutlu, Gürhan ve Liman (2004) yöneticilerin araştırma görevlilerinden lisanüstü eğitimlerini başarıyla tamamlamalarını, sınavlarda gözetmenlik yapmalarını, derslerle ilgili uygulamalara, laboratuvar çalışmalarına, ödev ve proje değerlendirmelerine, araştırma ve deneylere, kayıt işlemlerine ve öğrenci danışmanlıklarına yardım etmelerini beklediğini ifade etmektedir. Korkut, Muştan ve Yalçınkaya (1999) araştırma görevlilerinin akademik olmayan işlerle de meşgul olduğundan bahsetmektedir.

Araştırma görevlilerinden bu denli çok işin beklenmesi onların akademik olarak yetiştirilmesine ket vurmaktadır. Ergün (2001)'ün belirttiği gibi araştırma görevlileri, bilimsel araştırma ve öğretim faaliyetlerini çok iyi şekilde gerçekleştirebilecek şekilde yetiştirilmelidir. Ayrıca yüksek öğretimin sürdürülebilirliği ve gücü için personelin gelişimi, kurumda tutulması ve yöneticiler çok önemlidir (Pienaar ve Bester, 2009). Bu noktada şu konudan bahsetmekte yarar vardır. Türkiye'de son yıllarda yüksek öğrenime yönelik ilginin artması ve yeni üniversiteler açılması nedeniyle öğretim elemanlarına olan ihtiyaç artmıştır (Karakütük ve Özdemir, 2011). Alt yapı hazırlığı yapılmadığı için de mevcut öğretim elemanlarının iş yükü artmaktadır. Bu durum öğretim elemanlarının zaman zaman akademik gelişimlerinden ödün vermelerine neden olmaktadır. Arı (2007)'nın araştırma sonucunun gösterdiği gibi üniversiteden ayrılmayı en çok isteyen ünvana sahip akademik personeller araştırma görevlileridir. Ayrıca öğretim elemanı olmaktan en az memnun olan ve iş doyumu en az olan yine araştırma görevlileridir.

Yüksek iş yükünün yanında araştırma görevlilerinin iş güvenceleri de bulunmamaktadır. 2547 sayılı kanunun 33. maddesinde araştırma görevlileriyle ilgili olarak "bunlar ilgili anabilim veya anasanat dalı başkanlarının önerisi, bölüm başkanı, dekan, enstitü, yüksekokul veya konservatuvar müdürünün olumlu görüşü üzerine rektörün onayı ile araştırma görevlisi kadrolarına en çok üç yıl süre ile atanırlar; atanma süresi sonunda görevleri kendiliğinden sona erer. Bunlar aynı usulle yeniden atanabilirler." ifadesi yer almaktadır. Oysaki YÖK (Yüksek Öğretim Kurumu) yasasına dek araştırma görevlileri devamlı statüden hiçbir zaman yoksun bırakılmamıştır (Bakioğlu ve Yaman, 2004). Bu madde araştırma görevlilerinin mobbinge uğramasına yol açmaktadır. Mobbing psikolojik bir saldırı olup işyerinde bireylere astları, üstleri ya da arkadaşları tarafından uygulanabilmektedir. Mobbingin sebepleri kuruluş değerlerinin çökmesi, negatif iklim, bireyler arası çatışmalar, ortama güvenmeme, saygının yitirilmesi ve yaratıcılığın engellenmesi şeklinde açıklanmaktadır (Yaman, 2010). Çok sayıda araştırma görevlisi kadro yenilememe ile tehdit edilmektedir. Korkut, Muştan ve Yalçınkaya (1999)'nın araştırmasında yer verdiği gibi Türkiye'de işten çıkarmada yeterlik dışı ölçütler kullanılabilmektedir. 33/a kadrosunun dışında ÖYP (Öğretim Üyesi Yetiştirme Programı) ile atanan araştırma görevlileri bulunmaktadır. ÖYP bahsedilen 33/a kadrosunun özelliklerini taşımakta bunun yanı sıra araştırma görevlilerine yurt dışında ya da bir üniversite adına başka bir üniversitede

eğitim alma hakkı tanımaktadır. 33/a maddesinden dolayı ÖYP araştırma görevlilerinin de kadro güvencesinin bulunmadığı bilinmektedir (Karakütük ve Özdemir, 2011). Belirtilen durumlardan daha da kötüsü bazı araştırma görevlilerinin 2547 sayılı kanunun 50. Maddesinin d bendine göre atanmasıdır. 50. maddenin d bendi "lisans üstü öğretim yapan öğrenciler, kendilerine tahsis edilebilecek burslardan yararlanabilecekleri gibi, her defasında bir yıl için olmak üzere öğretim yardımcılığı kadrolarından birine de atanabilirler" ifadesini içermektedir. Bu madde doğrultusunda her sene kadrosu yenilenen 50/d'li araştırma görevlilerinin yükseköğrenimleri bittiğinde üniversite ile ilişikleri kesilmektedir. Bu durum araştırma görevlilerinin gelecekleriyle ilgili umutsuzluğa düşmesine ve insan gücü kaybına neden olmaktadır (Bakioğlu ve Yaman, 2004). Her şeyden önemlisi kadro güvencesinin olmaması üniversitelere nitelikli zihinlerin çekilmesini zorlaştırmaktadır (Korkut, Muştan ve Yalçınkaya, 1999). Belirtilen sorunlar dışında 33/a maddesi kapsamında ya da ÖYP kapsamında atanan araştırma görevlileri bir üniversite adına başka bir üniversitede görevlendirildiklerinde yüklü miktarda senet imzalamakta ve bu durum araştırma görevlilerinin baskı altında hissetmesine ve mobbinge uğramasına neden olmaktadır.

Araştırma görevlileri kendilerini geliştirmekten ve öğretim üyesi olduklarında da nitelikli yayın yapmaktan sorumludur (Kahraman, 2010). Bunu sağlayabilmek için ise araştırma görevlilerinin yeterli yabancı dil becerisine sahip olması gerekmektedir. Yabancı dil bilme öğretim elemanı yetiştirmede en önemli konulardan biridir (Korkut, Muştan ve Yalçınkaya, 1999). İyi bir yabancı dil becerisinin bilim dünyasını takip etmede oldukça önemli olduğu bilinmektedir. Araştırma görevlilerinin yabancı dil becerilerinin yeterliğini arttırmaya yönelik olarak ÖYP kapsamında yurt dışında ya da yurt içinde yabancı dil eğitimi verilmektedir. Ancak yurt dışı yabancı dil eğitimi ile ilgili sorunlar yaşandığı görülmektedir. Bunun yanı sıra ÖYP kapsamında atanmayan araştırma görevlileri yabancı dil sorunuyla karşı karşıya kalabilmekte ve bu konuda bir önlem alınmamaktadır (Tuzgöl Dost ve Cenkseven, 2007).

Alanyazında araştırma görevlilerinin ekonomik yönde sorunları olduğu belirtilmektedir (Tuzgöl Dost ve Cenkseven, 2007). Acar, Nemutlu, Gürhan ve Liman (2005) gerçekleştirdikleri araştırmada araştırma görevlilerinin %91.6'sının maddi açıdan tatmin olmadığını belirlemiştir. Bu araştırmalardan araştırma görevlilerinin çok büyük bir kısmının maaşlarından memnuniyetsizlik duyduğu görülmektedir. Bunun yanında araştırma görevlerinin yurt içi ya da yurt dışı kongreler için görevlendirildiklerinde kaynak bulamamaları büyük bir sorun olarak görülmektedir. ÖYP kapsamında atanan araştırma görevlilerinin ödenekleri bulunmasına rağmen bu ödeneklerin aktarımıyla ilgili sorunlar yaşanmaktadır. Maddi olanakların yetersizliği de üniversitelerin çekiciliğini engellemektedir (Korkut, Muştan ve Yalçınkaya, 1999).

Araştırma görevlilerinin sorunlarından bir diğeri de fiziki yetersizliklerdir. Korkut, Muştan ve Yalçınkaya (2004) kitap, laboratuvar, araç ve gereçler yönünden üniversitelerin zengin olması gerektiğini belirtmektedir. Fakat özellikle bir üniversite adına başka bir üniversiteye eğitim almak üzere gönderilen araştırma görevlilerinin belirli bir süre araçsal donanımlardan yaralanamadıkları ve çalışma odalarının kalabalıklığı ya da olmaması nedeniyle zorluk çektikleri bilinmektedir (Karakütük ve Özdemir, 2011).

Alanyazında araştırma görevlilerinin mesleki sorunlarıyla ilgili belirli sayıda araştırma yer almaktadır. Bunlardan bazıları (Korkut, Muştan ve Yalçınkaya, 1999) araştırma görevlilerinin genel sorunlarını belirlemeyi amaçlarken; bazıları araştırma görevlilerinin iş memnuniyeti ve bunu etkileyen faktörler (Acar, Nemutlu, Gürhan ve Liman, 2004), kariyer gelişimlerine destek kültürü (Bakioğlu ve Pekince, 2011), kariyer gelişimindeki engeller (Bakioğlu ve Yaman, 2004), lisansüstü eğitim yapmak için başka bir üniversiteye gönderilen araştırma görevlilerinin yaşam tarzları ve problemleri (Kahraman, 2010), bilim insanı yetiştirme projesi ve öğretim üyesi yetiştirme programıyla (Karakütük ve Özdemir, 2011) ilgilidir.

Araştırmanın Amacı

Bu araştırmada, araştırma görevlilerinin yaşadıkları mesleki sorunların önem düzeyinin belirlenmesi için halihazırda araştırma görevlisi olarak görev yapan bireylerin yargılarına dayalı olarak mesleki

sorunlarının ikili karşılaştırma yoluyla ölçeklenmesi amaçlanmıştır. Bu amaçla aşağıdaki sorulara yanıt aranılmıştır.

1. Araştırma görevlilerinin mesleki sorunlarının ölçek değerlerine göre sıralaması nasıldır?
2. Cinsiyete, öğrenim durumuna ve kadro türüne göre araştırma görevlilerinin mesleki sorunlarının ölçek değerlerine göre sıralaması nasıldır?

Araştırmanın Önemi

Toplumun bilgi üreten en önemli kurumlarından biri olan üniversitelerde nitelikli öğretim üyelerinin bulunması gerekmektedir. Araştırma görevliliğinin de öğretim üyeliğinin temel basamağı olduğu düşünüldüğünde bu sürecin araştırma becerileri ve öğretim becerileriyle ilgili olarak mümkün olduğunca verimli geçmesi gerekmektedir. Fakat araştırma görevlilerinin yaşadıkları bazı mesleki sorunlar gelişimlerini olumsuz yönde etkilemektedir. Şüphesiz ki bu sorunların önem düzeyinin açığa çıkarılması ve bu yönde önlem alınması gerekmektedir. Bu çalışma araştırma görevlilerinin mesleki sorunlarının önem düzeyini psikometrik bir yapıda ele aldığı için önemli görülmektedir.

YÖNTEM

Araştırmanın Türü

Araştırma genel tarama modeli ile desenlenmiştir. Tarama modelleri çok sayıda elemanın yer aldığı bir evrende, evren ile ilgili karakteristik özellikleri ortaya koymayı amaçlamaktadır (Frankell, Wallen ve Hyun, 2011).

Evren ve Örneklem

Araştırmanın evrenini Türkiye'de YÖK'e bağlı devlet ve vakıf üniversitelerinde görev yapan araştırma görevlileri oluşturmaktadır. YÖK'ün istatistiklerine göre, 2014 yılında Türkiye çapında devlet ve vakıf üniversitelerinde gören yapan toplam 44.440 araştırma görevlisi bulunmaktadır (YÖK, 2014). Krejcie ve Morgan (1970)'a göre 50000 kişilik evreni temsil edecek örneklem sayısı en az 381 olmalıdır. Bu değer dikkate alınarak ulaşılan ve evreni temsil etmesi amaçlanan 555 araştırma görevlisi bu araştırmanın örneklemini oluşturmaktadır. Örneklemdeki katılımcılar evrenden "basit seçkisiz örnekleme" yöntemiyle seçilmiştir. Basit seçkisiz örnekleme, evrendeki tüm birimlerin örnekleme seçilme olasılığının eşit ve bağımsız olduğu durumlarda kullanılan bir yöntemdir (Cohen, Manion ve Morrison, 2007). Tablo 1.'de araştırmaya katılan araştırma görevlilerinin cinsiyet, kadro türü ve öğrenim seviyesine göre dağılımı görülmektedir.

Tablo 1. Örneklemin Cinsiyet, Kadro Türü ve Öğrenim Değişkenlerine göre Frekans Dağılımı

		Öğrenim Durumu				
Cinsiyet	Kadro Türü	Yüksek Lisans	Doktora	Bütünleşik Doktora	Doktora Mezunu	Toplam
Kadın	33/a	25	78	4	12	119
	50/d	23	48	1	1	73
	Öyp	56	55	14	2	127
	Toplam	104	181	19	15	319
Erkek	33/a	30	62	1	6	99
	50/d	14	39	0	1	54
	Öyp	34	40	8	1	83
	Toplam	78	141	9	8	236
Toplam		182	322	28	23	555

Tablo 1 incelendiğinde araştırmaya öğrenim seviyesine göre sırasıyla 182'si (%32,8) yüksek lisans, 322'si (%58) doktora, 28'i (%5) bütünleşik doktora ve 23'ü (%4,2) doktora mezunu olmak üzere

toplam 555 araştırma görevlisinin katıldığı ve bunlardan 319'unun (%57,5) kadın, 236'sının (%42,5) erkek olduğu görülmektedir.

Veri Toplama Araçları

Ölçme aracının hazırlanması aşamasında araştırma görevlilerin mesleki sorunları hakkında alanyazın taraması yapılarak sorunlar saptanmıştır. Daha sonra farklı üniversite, bölüm, kadro türü, cinsiyette seçilen araştırma görevlileri ile araştırma görevlilerin mesleki sorunları hakkında görüşmeler yapılmış, görüşme notlarındaki ve alanyazındaki ortak sorunlar tespit edilmiştir. Ölçme aracının son haline getirilmesi aşamasında birisi alanında doktorasını tamamlamış, diğeri doktoraya devam eden iki ölçme ve değerlendirme uzmanından yardım alınarak ölçme aracında yer alacak olan sekiz sorun belirlenmiştir.

Ölçme aracının son hali iki kısımdan oluşmakta olup birinci kısımda demografik bilgiler (cinsiyet, öğrenim seviyesi ve kadro türü), ikinci kısımda ise araştırma görevlilerinin ikili olarak karşılaştıracakları sekiz mesleki sorun (akademik olmayan işlere yönlendirilme, fiziksel yetersizlikler, mobbing, yabancı dil sorunu, kadro güvencesi olmaması, ekonomik sorunlar, ödeneklerin yetersizliği ve idare tarafından verilen fakülte işlerinin yoğunluğu) bulunmaktadır. Ölçme aracı elektronik ortamda hazırlanmış ve katılımcılara e-posta yoluyla çevrimiçi olarak iletilmiştir.

Verilerin Analizi

Verilerin analizi aşamasında ölçme aracından elde edilen veriler üzerinde Thurstone'nun karşılaştırma yargılar kanunu V. Hal denklemi ile tam veri matrisinden ölçekleme işlemi yapılmıştır. Bunun için öncelikle araştırma görevlilerin mesleki sorunlarının ikili karşılaştırmasına ilişkin frekanslar belirlenerek frekans matrisi (*F*) oluşturulmuştur. Frekans matrisindeki her bir eleman araştırma görevlisi sayısına (*N*) bölünerek oranlar matrisi (*P*) bulunmuştur. Oranlar matrisindeki her birime karşılık gelen z değerleri belirlenerek birim normal sapmalar matrisi (*Z*) elde edilmiştir. Birim normal sapmalar matrisinin en alt satırına göre her bir sütunun ortalaması alınarak her bir uyarıcıya ilişkin ölçek değerleri (*Sj*) hesaplanmıştır. Eksen başlangıç değerini ölçek değerlerinin en küçüğüne ötelemek için en küçük ölçek değerinin mutlak değeri tüm ölçek değerlerine eklenmiştir. Böylece yeni ölçek değerleri (*Sc*) belirlenmiş ve sıralanmıştır. Son aşamada ölçek değerleri sayı doğrusunda gösterilmiştir (Güler ve Anıl, 1999; Turgut ve Baykul, 1992) .

Bu işlemlerden sonra Thurstone ikili karşılaştırma kanununun V. Hal denklemi için sahip olduğu varsayımları test etmek, gözlemcilerin uyarıcılara tepki verirken dikkat ve hassasiyet gösterip göstermediklerini kontrol etmek amacıyla ortalama hata miktarı ve bu hata miktarının anlamlı olup olmadığını test etmek amacıyla ki-kare testi ile iç tutarlılık hesaplanmıştır (Ki-Kare= 530,067; sd= 21). Ardından serberstlik derecesi için kritik değer (sd= 21 için Ki-Kare= 32,67) ile hesaplanan kritik değer karşılaştırılmış ve hesaplanan degerin ki-kare değerinden büyük (530,067>32,67) olmasından dolayı iç tutarlılığın düşük çıktığı görülmüştür. İç tutarlılığın düşük çıkması nedeniyle elde edilen verilerin, V. Hal denkleminin sayıltılarını karşılamadığı anlaşılmıştır (Turgut ve Baykul, 1992). Bu nedenle sorunların önem düzeyini belirlemek üzere III. Hal denklemi ile ölçekleme yapılmıştır. Bütün bu işlemler cinsiyet, öğrenim seviyesi ve kadro türü değişkenleri için tekrar edilmiştir ve bu değişkenler açısından da III. Hal denklemi ile ölçekleme yapılmıştır. Bu aşamada verilerin varyans değerleri kullanılarak ölçek değerleri elde edilmiş ve ölçek değerleri sıralanmıştır.

BULGULAR

Her araştırma görevlisinden 8 ayrı uyarıcıyı ikişer ikişer karşılaştırmaları istenmiştir. Bu karşılaştırma sonrası her bir uyarıcıya ilişkin frekans değerleri belirlenmiştir. Belirlenen frekans değerleri Tablo 2.'de gösterilmiştir.

Tablo 2. Frekans Matrisi

	A	B	C	D	E	F	G	H
A		376	280	436	289	219	236	254
B	179		232	408	255	158	177	189
C	275	323		413	292	271	295	286
D	119	147	142		158	101	105	123
E	266	300	263	397		274	282	276
F	336	397	284	454	281		304	304
G	319	378	260	450	273	251		264
H	301	101	269	432	279	251	291	

Frekans matrisi satırlardaki uyarıcıların sütunlardaki uyarıcılara tercih edilme sayısını göstermektedir. Örneğin A sorununun B sorununa tercih edilme frekansı 179, B sorununun A sorununa tercih edilme frekansı 376'dır. Aynı sorunun tercih edilme imkanı olmadığından köşegen üzerinde herhangi bir değer yer almamaktadır. Oranlar matrisinin bulunabilmesi için tüm değerler katılımcı sayısı olan 555'e bölünmüştür. Oranlar matrisine ilişkin değerler Tablo 3.'de gösterilmiştir.

Tablo 3. Oranlar Matrisi

	A	B	C	D	E	F	G	H
A		0,68	0,50	0,79	0,52	0,39	0,43	0,46
B	0,32		0,42	0,74	0,46	0,28	0,32	0,34
C	0,50	0,58		0,74	0,53	0,49	0,53	0,52
D	0,21	0,26	0,26		0,28	0,18	0,19	0,22
E	0,48	0,54	0,47	0,72		0,49	0,51	0,50
F	0,61	0,72	0,51	0,82	0,51		0,55	0,55
G	0,57	0,68	0,47	0,81	0,49	0,45		0,48
H	0,54	0,18	0,48	0,78	0,50	0,45	0,52	

Oranlar matrisinde esas köşegene göre simetrik olan değerlerin toplamının 1'e eşit olduğu görülmektedir. Oranlar matrisinden ortalaması 0, standart sapması 1 olan birim normal sapmalar matrisine geçilerek z değerleri elde edilmiştir. III. Hal denkleminin hesaplanması için kullanılacak normal sapmalar matrisi değerleri Tablo 4'de gösterilmiştir.

Tablo 4. Birim Normal Sapmalar Matrisi

	A	B	C	D	E	F	G	H
A		0,461	0,011	0,791	0,052	-0,267	-0,189	-0,106
B	-0,461		-0,207	0,628	-0,102	-0,569	-0,471	-0,411
C	-0,011	0,207		0,656	0,066	-0,029	0,079	0,038
D	-0,791	-0,628	-0,656		-0,569	-0,908	-0,881	-0,767
E	-0,052	0,102	-0,066	0,569		-0,016	0,020	-0,007
F	0,267	0,569	0,029	0,908	0,016		0,120	0,120
G	0,189	0,471	-0,079	0,881	-0,020	-0,120		-0,061
H	0,106	-0,908	-0,038	0,767	0,007	-0,120	0,061	

Birim normal standart sapmalar matrisi incelendiğinde esas köşegene göre simetrik değerlerin birbirinin işaret olarak tersi olduğu görülmektedir. Birim normal standart sapmalar matrisi üzerinde gerekli hesaplamalar yapılarak (σ_i değerleri bulunarak karesi alınmıştır) varyans değerleri bulunmuştur. Varyans değerleri ikişer ikişer toplanarak esas köşegen üzerindeki hücrelere yazılmıştır. Bu şekilde varyans toplamları matrisine ulaşılmıştır. Varyans toplamları matrisi Tablo 5'de gösterilmiştir.

Tablo 5. Varyans Toplamları Matrisi

		A	B	C	D	E	F	G	H
		0,468	0,014	2,711	1,163	3,528	0,725	0,507	1,066
A	0,468		0,481	3,179	1,631	3,996	1,193	0,975	1,534
B	0,014			2,725	1,177	3,541	0,739	0,520	1,080
C	2,711				3,875	6,239	3,437	3,218	3,777
D	1,163					4,691	1,889	1,670	2,230
E	3,528						4,253	4,034	4,594
F	0,725							1,232	1,791
G	0,507								1,573
H	1,066								

Varyans toplamları matrisindeki değerlerin tümünün karekökü alınarak varyans toplamlarının karekökü matrisi elde edilmiştir. Varyans toplamlarının karekökü matrisi Tablo 6'da gösterilmiştir.

Tablo 6. Varyans Toplamlarının Karekökü Matrisi

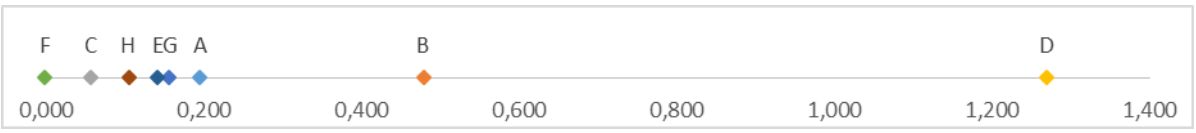
	A	B	C	D	E	F	G	H
A		0,694	1,783	1,277	1,999	1,092	0,987	1,239
B			1,651	1,085	1,882	0,860	0,721	1,039
C				1,968	2,498	1,854	1,794	1,944
D					2,166	1,374	1,292	1,493
E						2,062	2,009	2,143
F							1,110	1,338
G								1,254
H								

Varyans toplamlarının karekökü matrisiyle birim normal sapmalar matrisinin esas köşegeni üzerindeki tüm değerler çarpılarak S matrisi elde edilmiştir. S matrisinin esas köşegeninin altındaki değerler üstündeki değerlerin simetriği ve ters işaretlisidir. Bu yüzden sütun ortalamaları toplamı ve matris toplamı sıfırdır. S matrisi Tablo 7'de gösterilmiştir.

Tablo 7. Araştırma Görevlilerinin Mesleki Sorunlarına İlişkin S Matrisi

	A	B	C	D	E	F	G	H
A		0,320	0,020	1,010	0,104	-0,292	-0,187	-0,131
B	-0,320		-0,342	0,681	-0,192	-0,489	-0,340	-0,427
C	-0,020	0,342		1,291	0,165	-0,054	0,142	0,074
D	-1,010	-0,681	-1,291		-1,232	-1,248	-1,139	-1,145
E	-0,104	0,192	-0,165	1,232		-0,033	0,040	-0,015
F	0,292	0,489	0,054	1,248	0,033		0,133	0,161
G	0,187	0,340	-0,142	1,139	-0,040	-0,133		-0,077
H	0,131	0,427	-0,074	1,145	0,015	-0,161	0,077	
Toplam	-0,844	1,428	-1,940	7,747	-1,148	-2,409	-1,273	-1,561
Sj	-0,106	0,179	-0,243	0,968	-0,143	-0,301	-0,159	-0,195
Sc	0,196	0,480	0,059	1,270	0,158	**0,000**	0,142	0,106

Tablo 7'de görüldüğü gibi öncelikle sütun toplamları alınmıştır. Ardından sütun toplamları uyarıcı sayısına bölünmüştür. Elde edilen en küçük ölçek değeri başlangıç noktasına kaydırılarak bu değerin mutlak değeri kadar tüm ölçek değerlerine ekleme yapılmıştır. Elde edilen yeni ölçek değerleri Şekil 1'deki sayı doğrusunda gösterilmiştir.



Şekil 1. Uyarıcıların Sayı Doğrusu Üzerinde Gösterimi

Sonuçta 0 değerini taşıması nedeniyle F uyarıcısının araştırma görevlilerinin en önemli sorunu olduğu bulunmuştur. Ayrıca Tablo 7'deki ölçek değerlerine göre sorunların önem sıraları belirlenmiştir. Tablo 8'de araştırma görevlilerinin mesleki sorunları önem sırasına göre gösterilmektedir.

Tablo 8. Araştırma Görevlilerinin Mesleki Sorunlarının Ölçek Değerleri ve Uyarıcı Sıraları

Araştırma Görevlilerinin Mesleki Sorunları	Ölçek Değerleri	Uyarıcı Sıraları
Akademik Olmayan İşlere Yönlendirilme	0,196	6
Fiziki Yetersizlikler	0,480	7
Mobbing	0,059	2
Yabancı Dil Sorunu	1,270	8
Kadro Güvencesi Olmaması	0,158	5
Ekonomik Sorunlar	**0,000**	**1**
Ödeneklerin Yetersizliği	0,142	4
İdare tarafından verilen fakülte işlerinin yoğunluğu	0,106	3

Tablo 8 incelendiğinde araştırma görevlilerinin mesleki açıdan en önemli sorununun ekonomik olduğu görülmektedir. Bunu sırasıyla mobbing, idare tarafından verilen fakülte işlerinin yoğunluğu, ödeneklerin yetersizliği, kadro güvencesinin olmaması, akademik olmayan işlere yönlendirilme, fiziki yetersizlikler ve yabancı dil sorunu izlemektedir.

Araştırma görevlilerinin mesleki sorunları cinsiyetlerine göre ayrı ayrı ölçeklendiğinde elde edilen ölçek değerleri ve uyarıcı sıraları Tablo 9'da gösterilmiştir.

Tablo 9. Cinsiyete Göre Araştırma Görevlilerinin Mesleki Sorunlarının Ölçek Değerleri ve Uyarıcı Sıraları

	Kadın		Erkek	
Araştırma Görevlilerinin Mesleki Sorunları	Ölçek Değerleri	Uyarıcı Sıraları	Ölçek Değerleri	Uyarıcı Sıraları
Akademik olmayan işlere yönlendirilme	0,365	6	0,252	2
Fiziki yetersizlikler	0,608	7	0,579	7
Mobbing	**0,000**	**1**	0,377	6
Yabancı dil sorunu	1,397	8	1,402	8
Kadro güvencesi olmaması	0,227	5	0,344	5
Ekonomik sorunlar	0,184	3	**0,000**	**1**
Ödeneklerin yetersizliği	0,206	4	0,338	4
İdare tarafından verilen fakülte işlerinin yoğunluğu	0,180	2	0,295	3

Tablo 9 incelendiğinde kadın ve erkek araştırma görevlilerinin mesleki sorunlarının önem derecelerinin değiştiği görülmektedir. Kadın araştırma görevlilerinin en önemli sorunu mobbing iken erkek araştırma görevlilerinin en önemli sorunu ekonomiktir. Kadın araştırma görevlilerinin sorunlarında mobbing'i sırasıyla idare tarafından verilen fakülte işlerinin yoğunluğu, ekonomik sorunlar, ödeneklerin yetersizliği, kadro güvencesinin olmaması, akademik olmayan işlere yönlendirilme, fiziki yetersizlikler ve yabancı dil sorunu izlemektedir. Erkek araştırma görevlilerinin sorunlarında ise ekonomik sorunları sırasıyla akademik olmayan işlere yönlendirilme, idare tarafından verilen fakülte işlerinin yoğunluğu, ödeneklerin yetersizliği, kadro güvencesinin olmaması, mobbing, fiziki yetersizlikler ve yabancı dil sorunu izlemektedir.

Araştırma görevlilerinin mesleki sorunları öğrenim durumlarına göre ayrı ayrı ölçeklendiğinde elde edilen ölçek değerleri ve uyarıcı sıraları Tablo 10'da gösterilmiştir.

Tablo 10. Öğrenim Durumuna Göre Araştırma Görevlilerinin Mesleki Sorunlarının Ölçek Değerleri ve Uyarıcı Sıraları

	Yüksek Lisans		Doktora		Bütünleşik Doktora		Doktora Sonrası	
Araştırma Görevlilerinin Mesleki Sorunları	Ölçek Değ.	Uyarıcı Sır.	Ölçek Değ.	Uyarıcı Sır.	Ölçek Değ.	Uyarıcı Sır.	Ölçek Değ.	Uyarıcı Sır.
Akademik olmayan işlere yönlendirilme	0,152	5	0,251	6	0,185	2	0,749	2
Fiziki yetersizlikler	0,406	7	0,500	7	0,760	7	1,275	6
Mobbing	0,050	3	0,047	2	**0,000**	**1**	1,121	5
Yabancı dil sorunu	1,232	8	1,254	8	1,435	8	2,290	8
Kadro güvencesi olmaması	0,334	6	0,080	3	0,519	5	**0,000**	**1**
Ekonomik sorunlar	**0,000**	**1**	**0,000**	**1**	0,188	3	0,918	3
Ödeneklerin yetersizliği	0,054	4	0,160	5	0,450	4	1,170	6
İdare tarafından verilen fakülte işlerinin yoğunluğu	0,038	2	0,126	4	0,521	6	1,024	4

Tablo 10 incelendiğinde yüksek lisans, doktora ya da bütünleşik doktora yapan ve doktora mezunu olan araştırma görevlilerinin mesleki sorunlarının önem derecelerinin değiştiği görülmektedir. Yüksek lisans ve doktora yapan araştırma görevlilerinin en önemli sorunu ekonomik iken bütünleşik doktora yapan araştırma görevlilerinin mobbing, doktora mezunlarının ise kadro güvencesinin olmamasıdır. Yüksek lisans, doktora ya da bütünleşik doktora yapan ve doktora mezunu olan araştırma görevlilerinin önem düzeyi en düşük sorunları yabancı dil ile ilgilidir.

Araştırma görevlilerinin mesleki sorunları kadrolarına göre ayrı ayrı ölçeklendiğinde elde edilen ölçek değerleri ve uyarıcı sıraları Tablo 11'de gösterilmiştir.

Tablo 11. Kadro Türlerine Göre Araştırma Görevlilerinin Mesleki Sorunlarının Ölçek Değerleri ve Uyarıcı Sıraları

	33/a		Öyp		50/d	
Araştırma Görevlilerinin Mesleki Sorunları	Ölçek Değ.	Uyarıcı Sır.	Ölçek Değ.	Uyarıcı Sır.	Ölçek Değ.	Uyarıcı Sır.
Akademik olmayan işlere yönlendirilme	0,270	4	0,394	5	1,232	4
Fiziki yetersizlikler	0,546	7	0,535	6	1,801	7
Mobbing	0,179	3	0,225	3	1,054	2
Yabancı dil sorunu	1,286	8	1,275	8	2,953	8
Kadro güvencesi olmaması	0,512	6	0,793	7	**0,000**	**1**
Ekonomik sorunlar	0,103	2	**0,000**	**1**	1,172	3
Ödeneklerin yetersizliği	0,282	5	0,068	2	1,407	6
İdare tarafından verilen fakülte işlerinin yoğunluğu	**0,000**	**1**	0,365	4	1,277	5

Tablo 11 incelendiğinde kadro türüne göre araştırma görevlilerinin mesleki sorunlarının önem derecelerinin değiştiği görülmektedir. 33. maddenin a bendi kapsamında atanan araştırma görevlilerinin en önemli sorunu idare tarafından verilen fakülte işlerinin yoğunluğu iken ÖYP ile atanan araştırma görevlilerinin en önemli sorunu ekonomik, 50. maddenin d bendince atanan araştırma görevlilerinin en önemli sorunu ise kadro güvencesinin olmamasıdır. Tüm kadro türlerinin önem düzeyi en düşük sorunları ise yabancı dil ile ilgilidir.

SONUÇLAR ve TARTIŞMA

Bu çalışmada araştırma görevlilerin meslek hayatlarında karşılaştıkları sorunların önem düzeyi ikili karşılaştırma yöntemi kullanılarak belirlenmiştir. Araştırma sonucunda araştırma görevlilerinin en önemli sorununun ekonomik olduğu bunu sırasıyla mobbing, idare tarafından verilen fakülte işlerinin yoğunluğu, ödeneklerin yetersizliği ya da bulunmaması, kadro güvencesi olmaması, akademik olmayan işlere yönlendirme, fiziksel yetersizlikler ve yabancı dil sorununun izlediği bulunmuştur.

Araştırma sonucuna göre araştırma görevlileri tarafından en önemli mesleki sorun olarak *ekonomik sorunların* seçilmesi araştırma görevlilerinin maddi açıdan yetersizlik yaşadıklarının göstergesidir. Bu araştırma kamuoyunda akademik zam olarak ifade edilen akademik personelin maaşlarında yapılan iyileştirmeden sonra gerçekleştirilmiş olup yapılan iyileştirmeye rağmen araştırma görevlilerinin ekonomik sıkıntılarının devam ettiği görülmüştür. Bakioğlu ve Bülbül (2006) öğretim görevlileri üzerinde yaptığı çalışmada öğretim görevlilerinin ücret yetersizliğini meslekleri açısından en hoşnutsuz durum olarak gördüklerini belirtmiştir. Murat (2003) ise maddi imkansızlıkları öğretim elemanlarının en çok rahatsız eden ikinci problemi olarak raporlaştırmıştır. Korkut, Yalçınkaya ve Muştan (1999), araştırma görevlilerin yaşadığı ekonomik sorunların üniversiteden ayrılma kararında en önemli etken olduğunu belirtmiştir. Bu açıdan bakıldığında geleceğin öğretim üyeleri olan araştırma görevlilerinin maddi açıdan yaşadıkları sıkıntıların iş doyumlarını, örgüt bağlılıklarını etkileyeceği bir gerçektir (Bülbül, 2006). Buna ek olarak araştırma görevlileri maddi yetersizlikten dolayı asıl odaklanması gereken akademik üretkenlik ve mesleki gelişim konusunda sıkıntı yaşamaktadır (Kahraman, 2010).

Araştırma görevlileri açısından ikinci mesleki sorun ise mobbing olarak bulunmuştur. Mobbingin hem Türkiye'de hem de yurtdışında akademik camiada yaygın olarak karşılaşılan bir sorun olduğunu gösteren birçok çalışma bulunmaktadır (Celep ve Konaklı, 2013; Tüzel, 2008; Yaman, 2007; Westhues, 2007). Yaman (2007) yaptığı çalışmada akademik mobbing yüzünden öğretim görevlilerinin örgütsel aidiyetlerinin azaldığını ve bu durumun onları kötü yönde etkilediğini belirtmiştir. Benzer şekilde araştırma görevlilerinin de mobbingden yakındıkları görülmektedir.

İdare tarafından verilen fakülte işlerinin yoğunluğu, bu araştırmada araştırma görevlilerin karşılaştığı mesleki açıdan en önemli üçüncü sorun olarak bulunmuştur. Aynı sorun Korkut, Yalçınkaya ve Muştan (1999) tarafından belirtilmiş olsa da gözlenme oranı açısından bu çalışmanın sonucunu desteklememektedir. Yine bu çalışmada araştırma görevlilerine yöneltilen açık uçlu sorularda araştırma görevlilerinin diğer sorunlarının ödeneklerin yetersizliği ya da bulunmaması, kadro güvencesi olmaması, akademik olmayan işlere yönlendirme, fiziksel yetersizlikler ve yabancı dil sorunu olduğu bulunmuştur. Alanyazında akademik olmayan işlere yönlendirme sorununun temelde YÖK'ün araştırma görevlisi tanımı yapmamasının kaynaklandığı belirtilmektedir (Bakioğlu ve Pekince, 2011).

Araştırma görevlilerinin karşılaştıkları mesleki sorunlar cinsiyet değişkenine göre incelendiğinde mobbing konusunda ölçek değerleri açısından dikkat çeken bir fark olduğu görülmüştür. Kadın araştırma görevlileri mobbingi en önemli sorun olarak görürken erkek araştırma görevlileri bu sorunu altıncı sırada görmektedir. Kadınların akademik dünya gibi birçok iş kolunda mobbinge maruz kaldıkları bilinmektedir (Ceylan, 2005; Kök, 2006; Leymann, 1996). Celep ve Konaklı (2013) yaptıkları nitel çalışmada kadın öğretim görevlilerinin dış görünüşleri ve özel hayatlarından dolayı yıldırma yaşantılarına maruz kaldıklarını belirtmiştir. Şahin (2013) de benzer şekilde yaptığı vaka analizlerinde bu durumdan bahsetmektedir. Bu durum toplumdaki erkek egemenliğinin bir sonucu olarak kadınlara mobbingin daha kolay uygulanması ile ilişkili olabilir (İhan vd, 2009). Araştırma görevlilerinin sorunlarının cinsiyet değişkeni açısından incelenmesi sonucunda mobbing ve akademik olmayan işlere yönlendirilme sorunu dışında kadın ve erkek araştırma görevlilerinin sorunlarının benzerlik gösterdiği görülmüştür.

Araştırma görevlilerinin mesleki açıdan sorunlarının incelendiği diğer bir değişken ise öğrenim seviyesidir. Bu değişken açısından ölçek değerlerine bakıldığında doktora eğitimini tamamlamış araştırma görevlilerinin en önemli sorununun kadro güvencesinin olmaması olduğu görülmüştür. Bakioğlu ve Pekince (2011) yaptıkları çalışmada bu durumun sebebini doktora sonrasında kadroların başvuru ve atanma ölçütlerinin değişken olmasıyla açıklamıştır. Özkal (2010) araştırma görevlilerinin bu duruma ilişkin algılarının onların gelecek ile ilgili kaygılarının artmasına sebep olduğunu belirtmiştir. Bazı araştırmacılar bu durumun çözümünü yardımcı doçent kadrolarının arttırılarak doktora eğitimini başarıyla tamamlayan ve uygun şartları yerine getiren adayların bu kadrolara atanmasına olanak sağlanması olarak belirtmektedir (Bakioğlu ve Yaman, 2004; Korkut, Muştan ve Yalçınkaya, 1999).

Kadro türü değişkeni açısından araştırma görevlilerinin mesleki sorunları incelendiğinde ise 33/a kadro türü için temel sorunun idare tarafından verilen fakülte işlerin yoğunluğu olduğu görülmüştür. Bu durum 2547 sayılı YÖK kanununun 33. maddesindeki araştırma görevlisi tanımında yer alan "yetkili organlarca verilen ilgili diğer görevleri yapan öğretim yardımcılarıdır" ifadesinin açık olmamasından kaynaklandığı düşünülebilir. Kadro türü olarak "50/d" kadrosunda görev yapan araştırma görevlileri için en önemli sorun ise kadro güvencesinin olmamasıdır. Bulunan bu sonuç Bakioğlu ve Pekince (2011)'nin yaptığı çalışmadaki sonuçlar ile örtüşmektedir. Özkal (2010), 50/d kadrosunda çalışan araştırma görevlilerinin kadro konusunda endişe ettiklerini ve bunun da araştırma görevlilerinin motivasyonu bozan bir etmen olduğunu belirtmiştir. Ayrıca aynı araştırmada, 50/d'li araştırma görevlilerinin kadrosuz kalma korkusundan dolayı kendilerini daha az hür olarak algıladıkları ifade edilmiştir. Diğer bir kadro türü olan ÖYP kadrosundaki araştırma görevlilerinin en çok sıkıntı duydukları konular ise sırasıyla ekonomik sorunlar ve ödeneklerin yetersizliğidir. Bu iki sorun birbiriyle ilişkili olup öğretim üyesi yetiştirme programında vaad edilen ödeneklerin aktarılmamasından kaynaklı olabilir.

Sonuç olarak öğretim üyeliğinin temel basamağı olarak görülen araştırma görevliliğinde mesleki sorunlarının temelinde ekonomik sıkıntılar yatmaktadır. Diğer mesleki sorunların ise kadro tanımındaki sorunlardan kaynaklandığı düşünülmektedir. Bu nedenlerle öncelikle araştırma görevlierinin mali konularda (akademik teşvik, ödenek, vb.) desteklenmesi ve ardından yetkili organların yasal düzenlemelere giderek mesleki sorunlara çözüm bulması beklenmektedir. Bu çalışmada ele alınmayan özel ve devlet üniversitelerindeki araştırma görevlisi mesleki sorunlarının değişimi ileriki araştırmalarda incelenebilir. Ayrıca bu çalışmada belirlenen mesleki sorunların sebepleri nitel araştırmalarla derinlemesine incelenebilir.

KAYNAKÇA

Acar, A., Nemutlu, E., Gürhan, G. ve Liman, V. (2004). HÜ Eczacılık Fakültesi araştırma görevlilerinin iş memnuniyeti ve bunu etkileyen faktörler. *HÜ Eczacılık Fakültesi Dergisi, 24*(2), 95-106.

Arı, A. (2007). Üniversite öğretim elemanlarının sorunları. *Kırgızistan Türkiye Manas Üniversitesi Sosyal Bilimler Dergisi, 17*, 66-74.

Bakioğlu, A. ve Pekince, D. (2011). Araştırma görevlilerinin kariyer gelişimlerine bölümlerindeki destek kültürünün etkisi. *Uluslararası Yükseköğretim Kongresi: Yeni Yönelişler ve Sorunlar (UYK-2011)*, 27-29 Mayıs 2011, İstanbul, 2(XI), pp. 1272-1280.

Bakioğlu, A. ve Yaman, E. (2004). Araştırma görevlilerinin kariyer gelişimleri: Engeller ve çözümler. *Marmara Üniversitesi Atatürk Eğitim Fakültesi Dergisi, 20*, 1-20.

Bülbül, T. (2006). *Üniversite öğretim elemanlarının ücretlerinin akademik yaşama yansımalarının değerlendirilmesi.* Yayımlanmamış Doktora Tezi, Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Celep, C. ve Konaklı, T. (2013). Öğretim elemanlarının yıldırma yaşantıları: Nedenleri, sonuçları ve çözüm önerileri. *Kuram ve Uygulamada Eğitim Bilimleri, 13*(1), 175-199.

Ceylan, L. (2005). *Psikolojik baskı ve sınıf öğretmenleri.* Yayımlanmamış Yüksek Lisans Tezi, Niğde Üniversitesi, Sosyal Bilimler Enstitüsü, Niğde.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th Edition). London: Routledge Falmer.

Ergün, M. (2001). *Üniversitelerde öğretim etkinliğinin geliştirilmesi*. 2000 Yılında Türk Milli Eğitim Örgütü ve Yönetimi Ulusal Sempozyumu'nda sunulan bildiri, 11-13 Ocak 2001, Ankara.

Frankel, R. J., Wallen, E. N. and Hyun, H. H. (2011). *How to design and evaluate research in education* (8th Edition). New York: McGraw-Hill.

Güler, N. ve Anıl, D. (2009). Scaling through pair-wise comparison method in required characteristics of students applying for post graduate programs. *International Journal of Human Sciences, 6*(1), 627-639.

İlhan, M., Özkan, S., Kurtcebe, Ö. ve Aksakal, N. (2009). Gazi Üniversitesi Tıp Fakültesi Hastanesi'nde çalışan araştırma görevlileri ve intörn doktorlarda şiddete maruziyet ve şiddetle ilişkili etmenler. *Toplum Hekimliği Bülteni, 28*(3), 15-23.

Kahraman, A.B. (2010). Lisansüstü eğitim yapmak amacıyla başka bir üniversitede görevlendirilen araştırma görevlilerinin yaşam tarzı profilleri ve problemleri (Hacettepe Üniversitesi örneği). *Journal of World of Turks, 2*(2), 243-257.

Karakütük, K. ve Özdemir, Y. (2011). Bilim İnsanı Yetiştirme Projesi (BİYEP) ve Öğretim Üyesi Yetiştirme Programı'nın (ÖYP) değerlendirilmesi. *Eğitim ve Bilim, 36*(161), 26-38.

Korkut, H., Yalçınkaya, M. ve Muştan T. (1999). Araştırma görevlilerin sorunları. *Eğitim Yönetimi, 17,* 19-36.

Kök, S. B. (2006). İş yaşamında psiko-şiddet sarmalı olarak yıldırma olgusu ve nedenleri. *14. Ulusal Yönetim ve Organizasyon Kongresi Bildiri Kitabı*, 25-27 Mayıs 2006, Atatürk Üniversitesi İ.İ.B.F., Erzurum. 161-170.

Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and Psychological Measurement, 30,* 607-610.

Leymann, H. (1996). The Content and Development of Mobbing at Work. *European Journal of Work and Organizational Psychology*, *5*(2), 165–84.

Murat, M. (2003). Üniversite öğretim elemanlarında tükenmişlik. *Türk Psikolojik Danışma ve Rehberlik Dergisi, 2*(19), 25-34.

Özkal, F. M. (2010). *Akademik personel yetiştirme sürecini baltalayan sabıkalı istihdam maddesi: 50/d*. Cumhuriyetimizin Yüzüncü Yılına Doğru Üniversite Vizyonumuz Sempozyumu, 16-18 Nisan, Ankara.

Pienaar, C. & Bester, C. (2009). Addressing career obstacles within a changing higher education work environment: Perspectives of academics. *Psychological Society of South Africa, 39*(3), 376-385.

Şahin, M. (2013). Mobbing olgusuna anatomik bir bakış: Üniversite özelinde vaka analizi. *Akademik Bakış Dergisi Uluslararası Hakemli Sosyal Bilimler E-Dergisi, 38,* 1-20.

Turgut, M. F. ve Baykul, Y. (1992). *Ölçekleme teknikleri.* Ankara:ÖSYM Yayınları.

Tuzgöl Dost, M. ve Cenkseven, F. (2007). Devlet ve vakıf üniversitelerinde çalışan öğretim elemanlarının mesleki sorunları. *Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 16*(2), 203-218.

Tüzel, E. (2008). *Araştırma görevlilerinin maruz kaldıkları yıldırma (mobbing) davranışlarının araştırma görevlilerinin sahip oldukları çeşitli değişkenlere göre incelenmesi (Gazi Eğitim Fakültesi Örneği)*. Yayınlanmamış Yüksek Lisans Tezi Gazi Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Westhues, K. (2007). *The unkindly art of mobbing*. http://arts.uwaterloo.ca/~ kwesthue/ mobbing.htm Erişim tarihi: 12.01.2015.

Yaman, E. (2007). *Üniversitelerde bir yönetim sorunu olarak öğretim elemanlarının maruz kaldığı informal cezalar: Nitel bir araştırma*. Yayınlanmamış Doktora Tezi, Marmara Üniversitesi, Eğitim Bilimleri Enstitüsü, İstanbul.

YÖK (1981). 2547 Sayılı Yükseköğretim Kanunu.

YÖK (2014). Yükseköğretimde Kalite İçin. Ankara Üniversitesi Basımevi Müdürlüğü: Ankara.

EXTENDED ABSTRACT

Introduction

In the 33. article of the 2547 numbered law, research assistants are described as the assistant teaching staff who help the academic members in their research, analysis and experiments and carry out the duties given by the authorities. The duties and responsibilities of research assistants are not clear due to the following statement in the law: "they carry out the duties given by the authorities." Therefore, research assistants carry out a great many organizational duties as well as educational, teaching and research activities. Although they have heavy workload, they do not have safety at work. Another statement about research assistants in the article 33 of the 2547 numbered law is as follows: " They are appointed for maximum three years by the suggestion of the related head of the departments and positive opinion of the head of the department, dean or head of the related institute, college or conservatory; and their commissions expire automatically at the end of three years. They can be appointed to the same position by the same way." This article results in the exposure of research assistants to mobbing in the workplace. In fact, many research assistants are threatened with the renewal/extension of their contracts. There are research assistants who are assigned to this position within the scope of Academic Member Training Program (ÖYP). This program holds the features of 33/a vacancy and gives research assistants the opportunity to study abroad and at another university in the country in the name of their home universities. It is known that ÖYP research assistants do not have the guarantee of their positions because of the 33/a article. The worse is the fact that some research assistants are appointed according to 50/d article of the 2547 numbered law. Clause d of the 50th article includes the following statement: "graduate students can benefit from the scholarships

allocated for them and can be appointed to one of the assistant teaching staff positions for one year at each time." In this context, research assistants appointed for 50/d vacancies are from the university at the end of their graduate education, which drives them to dispair about their future and causes the loss of qualified labour force. As a result, it becomes difficult to employ qualified students as research assistants as the vacancies are not ensured. Moreover, research assistants of 33/a or ÖYP have to sign a substantial amount of deed if they are assisgned to study at another university and this puts them under pressure and makes them mobiing victims. Another issue is foreign language proficiency as it is one of the most siginificant issues in training academic members. Research assistants should have sufficient proficiency in at least one foreign language. In fact, ÖYP allocates some budget for improving the language skills of its research assistants by sending them to language courses either at home or around the world. However, there are some problems with overseas language courses. Besides, research assistants who are not appointed within the scope of ÖYP may have foreign language problems but not any precautions are taken to avoid such problems. It is also known that research assistants are not content with their salaries. Moreover, it is a big problem that they cannot find resources to meet their expenses when they are assigned to participate in national or international congresses. Although ÖYP research assistants have their own budget, they encounter with some problems in allocating funds. In this regard, insufficient financial resources influences the attraction of universities negatively. Another problem of research assistants is the insufficient physical opportunities. It is known that research assistants, especially the ones who are sent to other universities for education, cannot benefit from equipment for a while and have difficulties as they do not have rooms or share small rooms with a lot of people. This study aims to scale professional problems of research assistants with comparative measurement of their judgements, in order to define significance level of problems faced. In this context, this paper tried to answer the following questions:

- How do the problems of research assistants rank according to their scale values?
- How do ranking of the problems of research assistants change according to the paramaters: gender, educational background and type of position?

Method

This study was carried out in the form of general survey model. The population comprises the research assistants employed in the state or private universities affiliated to Higher Education Council (YÖK) in Turkey. According to the statistics of Higher Education Council, the total number of research assistants working at state or private universities is 44.400. The sample of the research was selected by simple random sampling method and the sample comprised 555 research assistants. The data collection tool is composed of two parts. The first parts asks for demographic information about gender, educational background and type of position while the second part inquires about the 8 problems (being directed non-academic duties, lack of physical equipment, mobbing, foreign language proficiency, lack of guaranteed positions, financial problems, insufficient funds and heavy faculty works given by the administration. While the data collection tool was prepared, the problems of research assistants were depicted by literature review.

Following this, research assistants from different universities, departments, positions and genders were interviewed about their problems. These problems were compared with ones available in the literature and mutual ones were depicted. Two experts in the field of assessment and evaluation were consulted to organize the assesment tool before it was used. During the data analysis process, data gathered through the assessment tool were evaluated in the form of full data matrix by fifth condition of state of Thurstone's Law of Comparative Judgment.

After this process, chi-square test was applied for internal consistencey in order to test the assumptions of the fifth condition of Thurstone's Law of Comparative Judgment. Since the internal consistency was low, it was understood that the collected data did not correspond to the assumptions of the V. State Equation. Thus, scaling was made with third condition in order to depict the

significance level of the problems. The same process was repeated for the gender, educational background and vacancy type variables.

Results and Discussion

The results of the study showed that the most significant professional problem of research assistants is financial ones which was followed by mobbing, heavy faculty work given by the administrators, insufficient or lack of funds, lack of guaranteed positions, being directed to non-academic duties, lack of physical equipment and foreign language proficiency.

When the research assistants' problems were analyzed in terms of gender variable, there was a meaningful difference in the scale values of mobbing. In fact, male research assistants puts mobbing in the sixth place while female research assistants regard it as the most significant problem. Educational background was the other variable according to which the problems of research assistants were studied. The results indicate that the most significant problem of those who have obtained their PhD degrees is lack of guaranteed positions. When research assistants' problems were examined in terms of type of vacancies, it was found that research assistants of 33/a position considered the heavy faculty work given by the administrators as their basic problem. As a result, it is clear that financial issues are the primary concern of research assistantship which is the first step for academia. The other problems are thought to be resulting from the problems in defining jobs and positions. Therefore, it is necessary to financially support the research assistants using academic encouragement, funds and etc. and solve the problems in job definitions through legal arrangements.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2015; 6(2); 177-193

Ankara Üniversitesi Uzaktan Eğitim Programına Katılan Öğrencilerin Akademik Başarılarını Yordayan Faktörler*

The Factors Predicting Academic Achievement of Ankara University Distance Education Students

Selma ŞENEL ** Ömer KUTLU ***

Öz
Bu çalışmanın amacı, Ankara Üniversitesi Uzaktan Eğitim Merkezi (ANKUZEM)'nde eğitim gören öğrencilerin akademik başarılarını yordayan faktörlerin neler olduğunu belirlemektir. Çalışma grubu, 2010-2011 eğitim-öğretim yılında ANKUZEM önlisans programlarının birinci sınıfında okuyan 302 öğrenciden oluşmaktadır. Veriler; bireye ilişkin, aileye ve çalışma ortamına ilişkin, bilgi ve iletişim teknolojilerinin kullanımına ilişkin ve eğitimsel veriler olarak dört grupta toplanmış, analiz edilmiş ve yorumlanmıştır. Verilerin çözümlenmesinde, aşamalı çoklu regresyon analizi kullanılmıştır. Analiz sonucunda kişisel özelliklerden; yaş, tam zamanlı bir işte çalışma ve boşanmış olmanın akademik başarıyı yordadığı gözlenmiştir. Aile ve çalışma ortamına ilişkin özelliklerden; birlikte yaşanılan birey sayısı, annenin tam günlük bir işte çalışması ve annenin lise mezunu olmasının akademik başarıyı yordadığı belirlenmiştir. Eğitimle ilgili özelliklerden ise; lisans ve meslek lisesi mezunu olmanın, Adalet önlisans uzaktan eğitim programında okumanın, uzaktan eğitimi 'üniversitede okumak' için seçmenin; ders çalışırken tercih edilen farklı yöntemlerin, derece almak için uzaktan eğitim görmenin akademik başarıyı yordadığı ortaya konulmuştur. Öğrencilerin bilgi ve iletişim teknolojileriyle ilgili özelliklerinin uzaktan eğitimde akademik başarının yordayıcılarından olmadığı görülmüştür.

Anahtar Kelimeler: akademik başarı, uzaktan eğitim, öğrenci özellikleri, ANKUZEM

Abstract
The aim of this study was to determine the factors predicting academic achievement of Ankara University Distance Education students. The study group of the research consists of 302 first grade students from Ankara University Distance Education Center (ANKUZEM) 2010-2011 associate degree programs. Data were formed in four groups respectively "individual characteristics", "characteristics related with family and working environment", "characteristics related with usage of information and communication technologies" and "characteristics related with education". Data were analyzed by stepwise multiple regression analysis. According to the results, individual characteristics as "age", "working in a full time job" and "being divorced" were found as predictors of students' academic achievement. Additionally, characteristics related with family and working place as "number of family members living with", "mothers' full time working" and "mothers' graduation type" were other variables predicting academic achievement of students. The characteristics related with education; graduation from "college" and "vocational high school", "studying with different methods", "choosing distance education as a university degree", "studying for getting a degree" are also predictors of academic achievement of students. The characteristics of students related with usage of information and communication technologies were not predictors of academic achievement of students.

Keywords: academic achievement, distance education, student characteristics, ANKUZEM

* Bu çalışma Selma ŞENEL'in Ankara Üniversitesi, Ölçme ve Değerlendirme Anabilim dalında Yrd. Doç. Dr. Ömer KUTLU danışmanlığındaki, aynı adlı yüksek lisans tez çalışmasından üretilmiştir.

**Uzman, Balıkesir Üniversitesi, Bilgi İşlem Araştırma ve Uygulama Merkezi, Balıkesir, selmasenel@balikesir.edu.tr
*** Yrd. Doç. Dr., Ankara Üniversitesi, Eğitim Bilimleri Fakültesi, Ankara, omerkutlu@ankara.edu.tr

Geliş Tarihi: 16.12.2014
Kabul Tarihi: 14.09.2015

GİRİŞ

Bilgi çağı olarak adlandırılan yirminci yüzyılda hızla çoğalan bilgi kaynakları ve bilginin yaygınlaşması ile birlikte, bireyin değişen bilgilere ve koşullara bağlı olarak kendisini yenileme gereği de zorunlu duruma gelmiştir. Bu zorunlulukla birlikte "yaşam boyu öğrenme" kavramı ortaya çıkmış, eğitimi; zaman, ortam, yaş gibi sınırlayıcı etkenlerden bağımsız kılabilen uzaktan eğitime olan ilgi de artmıştır. Uzaktan eğitim, geleneksel öğrenme-öğretme yöntemlerinin sınırlılıkları nedeniyle sınıf içi etkinliklerini yürütme olanağının bulunmadığı durumlarda, eğitim etkinliklerini planlayanlar ve uygulayıcılar ile öğrenciler arası iletişim ve etkileşimin özel olarak hazırlanmış öğretim üniteleri ve çeşitli ortamlar yoluyla belirli bir merkezden sağlandığı bir öğretim yöntemidir (Alkan, 1987).

Uzaktan eğitim; analiz, tasarım, geliştirme, uygulama ve değerlendirme basamaklarını içeren ve teknolojinin sürece bütünleştirilmesini gerektiren bir sistemdir (Driscoll, 2002). Yalnızca öğretim materyallerinin internet sitesine yerleştirilmesi uzaktan öğretimi temsil etmemektedir. Uzaktan eğitim sisteminin önemli bir basamağı da eğitimin amacına ulaşıp ulaşmadığının, öğrencilerin bilgileri ne ölçüde ve nasıl edindiğinin belirlenmesidir. Eğitim sistemindeki bilinen bu önemine ek olarak, uzaktan eğitimde öğrenciyle öğretmenin yüz yüze iletişimi olmaması nedeniyle, ölçme ve değerlendirme etkinlikleri daha fazla önem taşımaktadır (Seal ve Przasnyks, 2001). Uzaktan eğitim sistemlerinin de örgün eğitimde verilen sertifika ve diplomalara eşdeğer sertifika ya da diploma verme yeterliğine sahip olmaları, öğrenci başarılarının değerlendirilmesinin önemini daha da arttırmaktadır. Amerikan Uzaktan Eğitim Dergisinde (AJDE) 1987-2006 yılları arasındaki yapılan yayınlarda, baskın başlığın "değerlendirme" olması bu önemin bir göstergesi olarak değerlendirilebilir (Neto ve Santos, 2010).

Eğitimde değerlendirme çalışmaları, eğitim programının sağlam olup olmadığını anlama, öğretimde başvurulan yöntemlerin etkililik derecesini saptama, öğrencileri başarılı olabilecekleri düşünülen alanlara yönlendirme, öğrenme güçlüklerini teşhis etme, öğrenci başarısını saptama gibi birçok amaçla yapılmaktadır (Baykul, 2000). Öğrenci başarısını saptama, en sık karşılaşılan değerlendirme amaçlarından sayılabilir. Her sistemde olduğu gibi eğitim sisteminde de dışarıdan gelen bazı faktörlerin etkisiyle, süreç içindeki eğitim etkinlikleri amaçlanan davranışların kazandırılmasında yetersiz kalabilir. Bütün bunlar beklenen davranışların elde edilemeyişine ya da elde edilen ürünlerde yetersizliklere, başarısızlığa neden olabilir. Bu durum eğitim sisteminde ölçme ve değerlendirme ögesinin iyi işlemediğinin bir göstergesi olarak kabul edilebilir (Baykul, 2005).

Uzaktan eğitim sisteminin; öğrenci, iletişim ortamı ve kaynaklar olmak üzere üç ana ögesi vardır. Öğrenci ögesi, sistemin varlığının nedenini oluşturmaktadır. Bu ögenin çeşitli yönleriyle bilinmesi sistemin ve öğrencinin başarısı için zorunludur (Alkan, 1987). Genel anlamda başarı, istenilen sonuca ulaşma yönünde ilerlemedir (Wolman, 1973).Başarı kavramı çoğu zaman öğrencilerin genel başarısı ya da genel akademik başarısı karşılığında kullanılmaktadır. Bu tür geniş kapsamlı tanımlamalar yapılabileceği gibi, eğitimde başarı denildiğinde genellikle, öğretmenler tarafından takdir edilen notlarla, sınav puanlarıyla ya da her ikisi ile belirlenen beceriler veya kazanılan bilgilerin ifadesi olan "akademik başarı" kastedilmektedir (Carter ve Good, 1973; akt: Keskin ve Sezgin, 2009).

Öğrenci başarısında çok sayıda faktörün etkili olduğu bilinmektedir. Yapılan araştırmalar, bu faktörlerin çeşitli boyutlarda incelenebileceğini ortaya koymaktadır (Billings, 1989; Coldeway, 1986; Kember, 1995; Kerr, Rynearson, Kerr, 2006; Memduhoğlu ve Tanhan, 2009). Başarıyı yordayan bireysel, çevresel ve kurumsal faktörler bilinirse başarısızlığı doğuran nedenlerin kontrol altına alınabileceği düşünülmektedir (Özgüven, 1974).

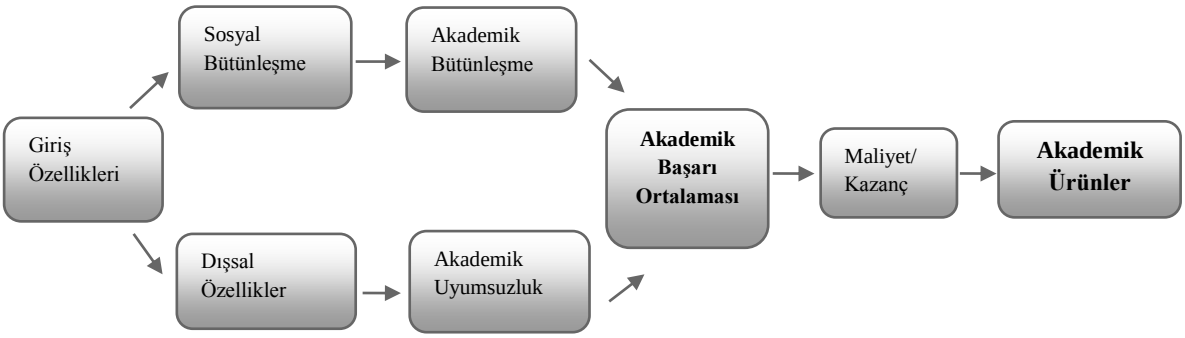
Uzaktan eğitim öğrencileri çok farklı özellikler taşıyabilmektedir. Hedef kitle; farklı meslek dallarından hizmet içi eğitim almak isteyen; ailevi sorumluluklar, örgün eğitimin devam zorunluluğu ya da fiziksel engeller nedeniyle örgün eğitim alamayan; bir eğitim kademesini tamamlamadan terk etmiş, farklı yaş gruplarından bireyler olabilmektedir (Sikora, 2002). Böyle geniş bir kitleye hitap

etmesi uzaktan eğitimde öğrenci özelliklerinin dikkate alınması gerekliliğini vurgulamaktadır. Bir başka anlatımla eğitim öğretim tasarımında, materyallerin seçiminde, hazırlanmasında ve iletilmesinde izlenen süreçte hedef kitlenin özelliği büyük önem taşımaktadır. Uzaktan eğitim, öğrenme ortamının yapısı nedeniyle daha çok kişisel bir çaba gerektirir. Uzaktan eğitim kurumunun görevi de; öğrencilerin yaş, deneyim, gelişim evresi, güdü, öğrenme isteği, öğrenme stili, aile ortamı, kendilerini yönlendirmeleri gibi özelliklerini dikkate alarak bu kişisel çabaya destek olmaktır (Koçer, 2001).

Alanyazında, uzaktan eğitimde başarıyı yordayan faktörlere ilişkin; öğrencinin uzaktan eğitimi seçme nedeni, daha önce herhangi bir uzaktan eğitim programına kayıt olup olmadığı, kursu belli bir süre içerisinde tamamlama niyetinin olması, programı bitirmek için hedeflerinin olması gibi uzaktan eğitimle ilgili özellikler; özgürlüğe düşkün olma, sosyal olma gibi kişisel özellikler; eğitimsel geçmiş, demografik özellikler, öğretim ortamları, öğretmen iletişimi ve başarı odağı gibi çok sayıda değişken listelenmektedir (Moore ve Kearsley, 2005; Oladejo, Ige, Fagunwa ve Arewa, 2010; Whittington, 1995).Bu çalışmaların yanında; uzaktan eğitimdeki akademik başarıyı yordayan faktörleri bir bütün halinde resmedebilmek adına oluşturulan modellerin de alan yazında dikkat çekici düzeyde tartışıldığı görülmektedir.

Tinto (1975)'nun başarı modelinde, öğrencilerin okula devam durumları "başarı" olarak kavramsallaştırılmıştır. Öğrenci özellikleri okula devamın en önemli yordayıcısı olmuştur. Tinto (1975)'nun başarı modelinde beş değişken nedensel sıra içerisinde sunulmuştur. Bunlar; geçmiş özellikleri; birincil amaç ve kurumsal üstlenme (taahhüt); akademik ve sosyal bütünleşme; sonraki amaç ve kurumsal taahhütler ve okulu bırakma kararlarıdır. Bean (1980) ise Tinto (1975)'dan farklı olarak üniversite öğrencilerinin okula devam etmelerini etkileyen dış faktörleri de içine alan Öğrenci Yıpranma Modeli'ni (Student Attrition Model) geliştirerek bu eksikliğe de değinmiştir. Öğrenci Yıpranma Modeli öğrencilerin okuldaki akademik ve sosyal deneyimleriyle oluşan inanç ve tutumlarının okula devam etme durumlarını etkilediğini önermektedir. Uzaktan eğitimdeki başarıda bireysel özellikler ve gereksinimlerin önemi Coldeway (1986) tarafından önerilen uzaktan eğitim başarısı modelinde de belirtilmiştir. Bu modelde, başarı dört faktörün etkisindedir. Bunlar; bireysel özellikler, okula kayıt için motivasyon, kurumsal faktörler (ödeme seçenekleri, iletişim materyali vb.) ve ders faktörleridir.

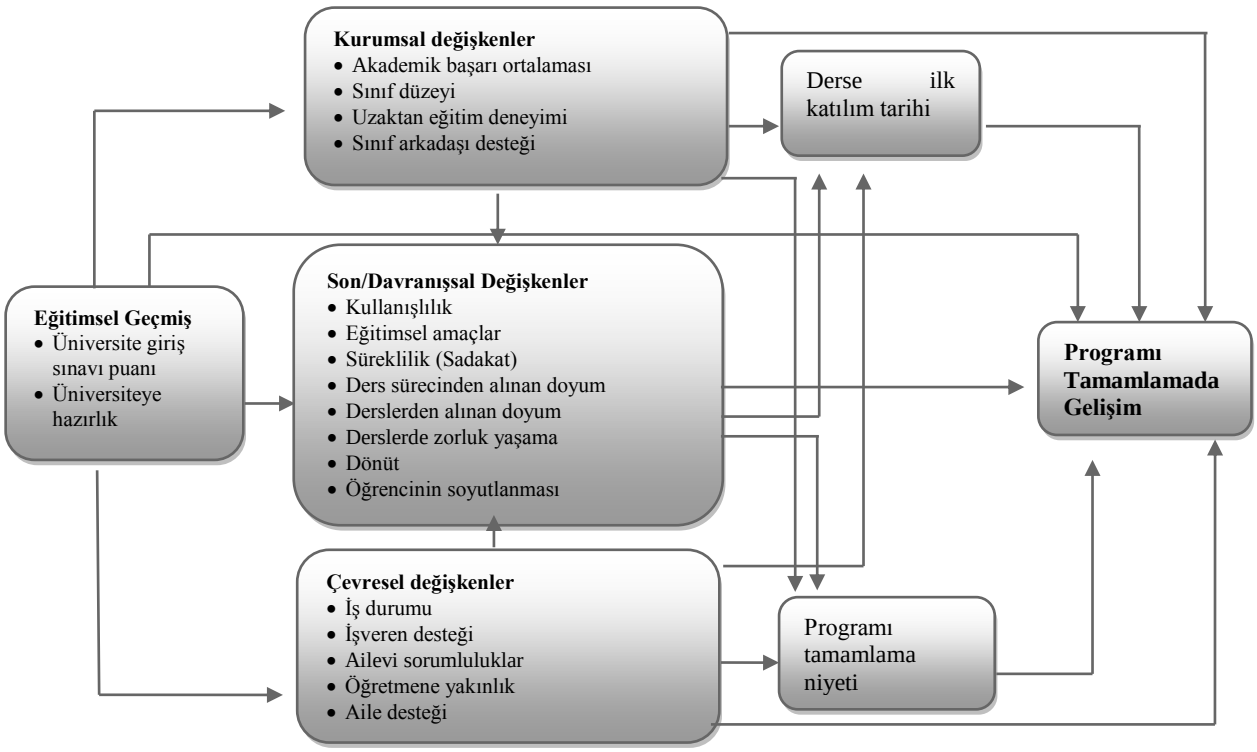
Kember (1995)'da uzaktan eğitimde okulu bırakma ve okula devam etme durumlarını öğrenci başarısı olarak ele almış; başarı ile öğrenci özellikleri arasında bir ilişki olduğunu düşünerek Uzaktan Eğitimde Akademik Başarı Modeli'ni (Şekil 1) geliştirmiştir. Modelin kuramsal çerçevesi oluşturulurken, öğrenci başarısını etkileyen faktörler üzerinde duran Tinto (1975) referans alınmıştır.



Şekil 1. Kember'in Uzaktan Eğitimde Öğrenci Başarısı Modeli

Bir diğer model ise Billings (1989)'in diğer modellere göre daha fazla değişken içeren Uzaktan Eğitimde Program Tamamlama Modeli'dir. Bu modelde programı tamamlamanın en iyi yordayıcısının "programı tamamlama niyeti" olduğu ifade edilmektedir. Derslerini belli bir sürede (üç ayda) tamamlama niyeti olan, ilk dersini nispeten erken (ilk 40 gün içinde) alan, giriş sınavı puanları ve akademik ortalamaları yüksek olan, aldığı dersleri tamamlamış, üniversiteye giriş için

hazırlanmış, ailesinden destek alan, öğretim elemanlarına daha yakın ve programı tamamlamak için amaçları olan öğrencilerin programı başarıyla tamamladıkları gözlenmiştir (Şekil 2).



Şekil 2. Billings'in Uzaktan Eğitimde Program Tamamlama Modeli

Tüm bu modeller ve çalışmalardan uzaktan eğitimde öğrenci özelliklerinin öğrencilerin akademik başarıları, okula devamları ya da dersi tamamlamaları üzerindeki etkisi görülebilir. Uzaktan eğitim sistemlerinde yeterli ve etkili sonuçların alınabilmesi için, hedef kitlenin demografik özelliklerinin, uzaktan eğitime yönelik ilgi ve tutumlarının iyi belirlenmesi, sistemin öğrencinin gereksinimlerine uygun olarak tasarlanması gerekmektedir. Uzaktan eğitimi tercih edenler genellikle yetişkinlerdir. Yetişkin öğrenciler; işleri, aileleri ve sosyal yaşamları olması ve farklı sorumluluklara sahip olmaları nedeniyle uzaktan öğrenme yöntemlerini kullanarak eğitim almak isteyebilmektedirler (Hughes ve Forest, 1997; Sikora, 2002). Bu öğrencilerin eğitime ilişkin yaşantılarının, yaşamlarındaki tek odakları olmadığı düşünülürse, akademik başarılarını etkileyebilecek faktörlere karşı daha savunmasız oldukları öngörülebilir. Uzaktan eğitim sürecinin daha en başında, öğrencilerin farklı sorumlulukları eğitimsel amaçlarının üzerine çıkabilmektedir. Uzaktan eğitimde başarıyı yordayan faktörlerin belirlenmesi, uzaktan eğitimde başarıyı artırmak için hangi değişkenler üzerinde çalışılabileceği konusunda bilgi sağlayacaktır. Uzaktan eğitim programlarını hazırlayan yetkililerin, bu faktörleri ve öğrenci başarısına etkilerini; eğitim programı hazırlama ve öğretim tasarımı süreçlerinde dikkate almalarını sağlamak, öğrencilerin akademik başarılarının artırılmasında önemli bir adım olabilir.

Araştırmanın Amacı

Bu çalışmanın amacı, Ankara Üniversitesi Uzaktan Eğitim Merkezi (ANKUZEM) önlisans programlarında okuyan öğrencilerin akademik başarılarını yordayan faktörleri belirlemektir. Öğrencilerin 1.sınıf sonundaki akademik not ortalamaları, akademik başarı olarak kavramsallaştırıldığından, bu amaç çerçevesinde aşağıdaki sorulara yanıt aranmıştır.

1. *Kişisel faktörler* (cinsiyet, yaş, iş durumu, medeni durum, çocuk sayısı) *akademik başarının yordayıcısı mıdır?*

2. Aileye ve çalışma ortamına ilişkin faktörler (anne ve babanın eğitim düzeyi, birlikte yaşanılan aile bireylerinin sayısı, anne ve babanın iş durumu, ailenin aylık geliri, evde sahip olunan olanaklar [bilgisayar, İnternet, kitaplık, çalışma masası, çalışma odası]) *akademik başarının yordayıcısı mıdır?*

3. Bilgi ve iletişim teknolojilerinin kullanımına ilişkin faktörler (Bilgisayarı ve interneti kaç yıldır kullandığı, bilgisayar ve internet kullanım düzeyi, haftada ortalama bilgisayar ve internet kullanım süresi, bilgisayar ve internete erişim ortamı) *akademik başarının yordayıcısı mıdır?*

4. Eğitimle ilgili faktörler (mezun olunan lise türü, daha önce bir lisans ya da önlisans programından mezun olup olmama, daha önce herhangi bir uzaktan eğitim programını tamamlayıp tamamlamama, uzaktan eğitimi seçme nedeni, günlük ortalama ders çalışma saati, derslere çalışmada izlenen yol/yollar, uzaktan eğitimi bitirerek ulaşmak istediği hedef) *akademik başarının yordayıcısı mıdır?*

YÖNTEM

Bu araştırma, öğrencilerin akademik başarıları ile "kişisel özellikler", "aile ve çalışma ortamı", "bilgi ve iletişim teknolojilerinin kullanımı" ve "eğitim" boyutlarında yer alan değişkenler arasındaki ilişkileri ve bu değişkenlerin öğrencilerin akademik başarılarını ne derece yordadığını belirlemeyi amaçlamaktadır. Bu nedenle araştırma yordayıcı korelasyonel bir araştırmadır. Korelasyonel araştırmalar, iki ya da daha çok değişken arasındaki ilişkinin, bu değişkenlere müdahale edilmeden incelendiği araştırmalardır. Bu tür araştırmalarda neden-sonuç ilişkisinden çok değişkenlerin birlikte değişimleri incelenir (Büyüköztürk, Kılıç Çakmak, Akgün, Karadeniz ve Demirel, 2009).

Çalışma Grubu

Çalışma grubu ANKUZEM önlisans programlarındaki 2010-2011 eğitim-öğretim yılında 1. sınıf derslerini alan 302 öğrenciden oluşmaktadır. Önlisans programlarında eğitim gören ve araştırmaya katılan öğrenci sayıları Tablo 2'te verilmiştir. Veri toplama aracı uygulanırken gönüllülük esası dikkate alındığından, 1.sınıf derslerini alan öğrencilerin %37'sine ulaşılabilmiştir.

Tablo 2. ANKUZEM Önlisans Programlarının 1. Sınıfında Okuyan ve Araştırmaya Katılan Öğrenci Sayıları

Önlisans Programları	1. Sınıf	1. Sınıf Tekrar	Toplam	Araştırmaya Katılan Öğrenci Sayısı
Adalet (ADUZEP)	121	65	186	107
Turizm ve Otel İşletmeciliği (TOİ)	66	12	78	24
Bankacılık ve Sigortacılık (BAS)	95	62	157	61
Bilgisayar Programcılığı (BİPRO)	136	88	224	105
Toplam	551	267	818	302

Veri Toplama Araçları

Çalışmaya ilişkin veriler, öğrenci özelliklerinin betimlendiği öğrenci anketi ile toplanmıştır. Anketin oluşturulmasında Kutlu, Büyüköztürk ve Doğan (2007) tarafından öğretmenlerin yeni değerlendirme yaklaşımlarına ilişkin tutumlarını etkileyen faktörleri belirlemek amacıyla kullanılan anketten ve ANKUZEM'in öğrencilerine uyguladığı öğrenci bilgileriyle ilgili anketin sonuçlarından yararlanılmıştır (ANKUZEM, 2011).

Öğrenci anketinin, yalnızca alan yazına dayanarak oluşturulması yeterli görülmediği için belirlenen faktörlerin seçilen örneklemi etkileyip etkilemediği, etkiliyorsa nasıl etkilediğini öğrenmek için önlisans öğrencilerine sorulmak üzere açık uçlu sorular oluşturulmuştur. Oluşturulan sorular öğrencilerin uzaktan eğitimde başarılarını hangi faktörlerin etkilediği ile ilgili düşüncelerini sorgulamaktadır. Sorular elektronik ortama e-anket olarak aktarılmış, anket bağlantısı önlisans programlarından 200 öğrenciye, e-posta yoluyla gönderilmiş ve 20 yanıt alınmıştır. Yanıtlayıcı sayısının e-posta gönderilen kişi sayısına oranla oldukça düşük olması e-anketlerde çoğunlukla rastlanan bir durumdur.

Öğrenci yanıtlarına yapılan içerik analizi sonuçları ve alanyazın dikkate alınarak, anket maddelerinde (1) kişisel özellikler, (2) aileye ve çalışma ortamına ilişkin özellikler, (3) eğitimsel özellikler ve (4) bilgi ve iletişim teknolojilerinin kullanımına ilişkin özellikler olmak üzere dört ana faktör üzerinde durulmuştur. Araştırmanın veri toplama aracı olarak geliştirilen anket bu ana faktörlerin altında yer alan yarı yapılandırılmış sorulardan oluşturulmuştur. Anket için eğitimde ölçme ve değerlendirme alanından bir öğretim üyesinden ve ANKUZEM'de görevli bir öğretim üyesinden uzman görüşü alınmıştır. Uzmanların bazı sorularda seçeneklerin eklenmesine ilişkin önerileri doğrultusunda ankete son hali verilmiştir.

İşlem

Öğrenci anketi, final sınavları için uzaktan eğitim merkezinin organize ettiği sınav merkezlerine gelen öğrencilere, sınav kâğıtlarıyla birlikte verilerek, öğrencilerin sınav sonrasında yanıtlamaları sağlanmıştır.

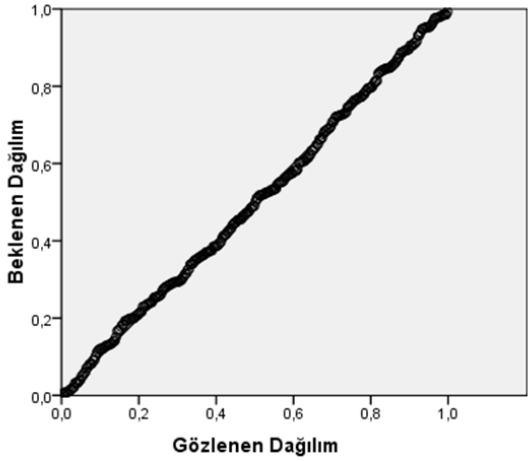
Çalışma grubundaki öğrencilerin önlisans akademik not ortalamaları, lise mezuniyet notu, mezun olduğu lise türü gibi eğitimsel geçmişleri ile ilgili bilgiler Ankara Üniversitesi Öğrenci İşleri Dairesi Başkanlığı'ndan elde edilmiştir.

Verilerin Çözümlenmesi

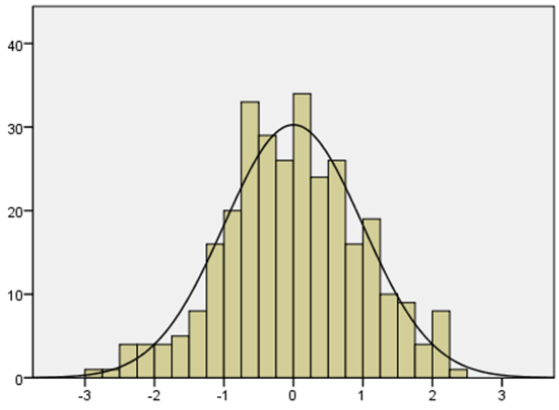
Verilerin çözümlenmesinde yordama çalışmaları için uygun bir istatistiksel teknik olan aşamalı çoklu regresyon analizi kullanılmıştır. Araştırmada aşamalı çoklu regresyon analizinin tercih edilmesinin nedeni, akademik başarıya önemli katkı getiren değişkenlerin işleme alınmak istenmesidir. Bir değişkenin herhangi bir aşamada regresyon eşitliğine alınabilmesi için $\alpha = .05$ düzeyi esas alınmıştır. Regresyon eşitliğine alınan bir değişkenin daha sonraki aşamalarda analiz dışı bırakılabilmesi için $\alpha = .10$ düzeyi esas alınmıştır. Çalışmanın yordanan (bağımlı) değişkeni; "öğrencilerin akademik not ortalamaları"dır. Yordayıcı (bağımsız) değişkenler ise ağırlıklı olarak süreksiz değişkenlerden oluşmaktadır. Çalışmada yer alan süreksiz değişkenler regresyon analizine "dummy değişken" olarak kodlanarak dâhil edilirken, sürekli değişkenler orijinal değerleri ile analize alınmıştır. Analize sokulan tüm değişkenlerin dummy kodlamalarına ilişkin bilgi Ek-1'te verilmiştir.

Her bir faktör için çoklu doğrusal regresyon analizinin; normal dağılım, doğrusallık, sabit varyans, otokorelasyonun olmaması, bağımsız değişkenler arasında çoklu bağlantı olmaması varsayımları (Kalaycı, 2009) test edilmiştir.

Kişisel faktörler olarak belirlenen değişkenlerin normallik ve doğrusallık varsayımları standartlaştırılmış tahmini değerler ile standartlaştırılmış hata (sapma) değerleri arasındaki grafiklerle (Şekil 3 ve Şekil 4) incelenmiştir (Büyüköztürk, 2009). Şekil 3'e göre değişkenler arasında doğrusal ve pozitif yönde bir ilişki vardır. Şekil 4'e göre standardize edilmiş yordanan değerler için oluşturulan histogram ve normal dağılım eğrilerinin normale yakın bir dağılımı gösterdiği ileri sürülebilir.



Şekil 3. Kişisel Faktörlere İlişkin Doğrusallık Dağılımı



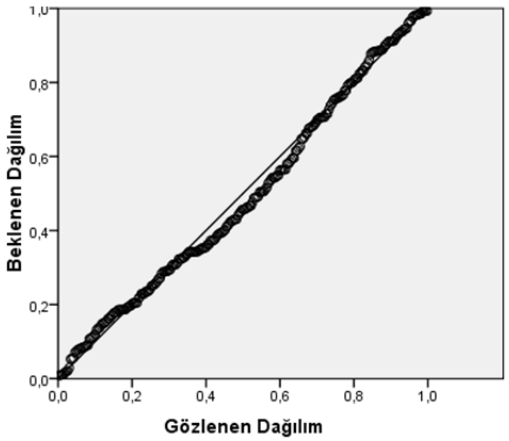
Şekil 4. Kişisel Faktörlere İlişkin Normallik Dağılımı

Çoklu regresyon analizlerinde yordayıcı değişkenler arasında çoklu bağlantılar (multi-collinearity) olarak tanımlanan bir sorunla karşılaşılabilir. Analizde aşağıda verilen durumlardan herhangi birinin ortaya çıkması bağımsız değişkenler arasında çoklu bağlantının olmasına işaret etmektedir (Büyüköztürk, 2009):

a. Tolerans değerinin ($1\text{-}R^2$) 0,20'den daha düşük çıkması
b. Varyans büyütme faktörü (VIF) değerinin, 10'dan yüksek çıkması
c. Durum indeks (CI) değerinin, 30'dan yüksek çıkması

Kişisel faktörlere ilişkin yordayıcı değişkenler arasında çoklu bağlantılılık göstergelerine bakıldığında tolerans değerleri 0,67 ile 0,76 arasında, varyans büyütme faktörü (VIF) değerleri 1,17 ile 1,50 arasında bulunurken, en yüksek durum indeks (CI) değeri ise 13,94 olarak bulunmuştur. Bu durumda yordayıcı değişkenler arasında çoklu bağlantılılık bulunmadığı ifade edilebilir.

Hata terimleri arasında ilişki olması anlamına gelen otokorelasyonu test etmede kullanılan Durbin Watson değeri 0 ile 4 arasında değişir. 0'a yakın değerler aşırı pozitif korelasyonu, 4'e yakın değerler aşırı negatif korelasyonu, 2'ye yakın değerler otokorelasyonun olmadığını gösterir. Bu nedenle benzer çalışmalarda Durbin Watson değerinin 1,5 ile 2,5 arasında olması arzulandığı görülmektedir (Kalaycı, 2009). Kişisel faktörlere ilişkin modelde Durbin Watson değerinin 1,77 olması birinci modelde otokorelasyonun olmadığını göstermektedir.

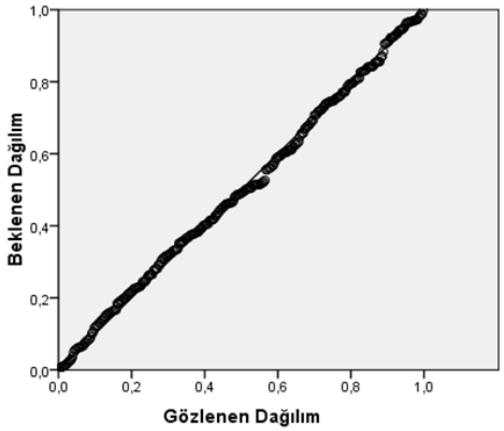


Şekil 5. Aile ve Çalışma Ortamı Faktörlerine İlişkin Doğrusallık Dağılımı

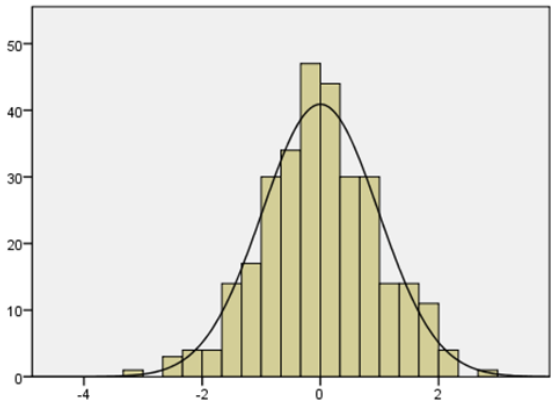
Şekil 6. Aile ve Çalışma Ortamı Faktörlerine İlişkin Normallik Dağılımı

Aileye ve çalışma ortamına ilişkin özellikler için doğrusallık ve normallik varsayımlarının incelenmesine ilişkin grafikler Şekil 5 ve 6'te verilmiştir. Şekil 5'e göre değişkenler arasında doğrusal ve pozitif yönde bir ilişki vardır. Şekil 6'ya göre standardize edilmiş yordanan değerler için oluşturulan histogram ve normal dağılım eğrilerinin normale yakın bir dağılımı gösterdiği ileri sürülebilir. Yordayıcı değişkenler arasında çoklu bağlantılılık göstergeleri incelendiğinde; tolerans değerleri 0,98; varyans büyütme faktörü (VIF) değerleri 1,00 ile 1,02 arasında, en yüksek durum indeks (CI) değeri ise 6,41 olarak bulunmuştur. Bu durumda yordayıcı değişkenler arasında çoklu bağlantılılık bulunmamaktadır. Durbin Watson değerinin 1,87 olması; ikinci modelde de otokorelasyonun olmadığını göstermektedir.

Bilgi ve iletişim teknolojilerinin kullanımına ilişkin özellikler grubundaki değişkenlerle öğrencilerin genel akademik not ortalamaları arasında varsayımların karşılanıp karşılanmadığıyla ilgili test yapılmak istenmiştir. Ancak belirlenen manidarlık düzeyinde hiçbir değişken regresyon eşitliğine giremediği için SPSS programı doğrusallık ve normallik varsayımlarını da test etmemiştir.



Şekil 7. Eğitimle İlgili Faktörlere İlişkin Doğrusallık Dağılımı



Şekil 8. Eğitimle İlgili Faktörlere İlişkin Normallik Dağılımı

Eğitimle ilgili özellikler için doğrusallık ve normallik varsayımlarına ilişkin grafikler Şekil 7 ve 8'de verilmiştir. Şekil 7'ye göre değişkenler arasında doğrusal ve pozitif yönde bir ilişki olduğu, Şekil 8'in normale yakın bir dağılımı gösterdiği ileri sürülebilir. Yordayıcı değişkenler arasında çoklu bağlantılılık göstergelerine bakıldığında; tolerans değerleri 0,76 ile 0,97 arasında, varyans büyütme faktörü (VIF) değerleri 1,04 ile 1,31 arasında, en yüksek durum indeks (CI) değeri ise 5,23 olarak bulunmuştur. Bu durumda yordayıcı değişkenler arasında çoklu bağlantılılık bulunmamaktadır. Durbin Watson değerinin 1,95 olması; üçüncü modelde de otokorelasyonun olmadığını göstermektedir.

Çoklu doğrusal regresyon analizinin varsayımlarından hata terimlerinin ortalamasının "0" olup olmadığını test etmek için, tahmini ortalama değerleri kaydedilmiş ve gerçek ortalama değerlerinden gözlenen ortalama değerlerinin çıkarılması ile hata terimlerine ulaşılmıştır. Hata terimlerinin ortalaması ise, "0,0064" bulunmuştur. Bu değerin sıfıra çok yakın bir değer olması hata terimlerinin ortalamasının sıfır olması varsayımının karşılandığını göstermektedir.

BULGULAR

Kişisel Faktörlere İlişkin Bulgular

Tablo 3'te verilen regresyon analizi sonuçlarına göre kişisel faktörlerin akademik başarıyı yordamasına ilişkin regresyon eşitliği aşağıdaki gibidir.

Akademik not ortalaması= $1,211+ 0,036X_1- 0,251X_2+0,275X_3$

Tablo 3. Kişisel Faktörlere İlişkin Aşamalı Çoklu Regresyon Analizi Sonuçları

Değişken	B	R	R^2	β	β^2	t	P	İkili r	Kısmi r
Sabit	1,211					5,24	,000		
Yaş [X_1]	,036	,239	,057	0,238	,057	3,53	,000	0,200	0,194
İş Durumu1 [X_2]	-0,251	,285	,024	-0,166	,027	-2,78	,006	-0,159	-0,153
Medeni Durum2 [X_3]	,275	,307	,013	0,132	,017	2,082	,038	0,120	0,115
R^2= 0,094 F=10,342 sd= 3;298 p=0,000									

Yordayıcı değişkenlerle yordanan değişken arasındaki ikili ve kısmi korelasyonlar incelendiğinde yaş ve boşanmış olma (MedeniDurum2) değişkeni ile akademik başarı arasında düşük düzeyde pozitif ilişkilerin (r=0,2 ve r=0,12) olduğu, diğer değişkenler kontrol altında tutulduğunda da bu ilişkilerin yaklaşık olarak aynı düzeyde kaldığı (r=0,194 ve r=0,115) ifade edilebilir. Tam günlü bir işte çalışma (İşDurumu1) değişkeni ile akademik başarı arasında düşük düzeyde negatif bir ilişkinin (r=-0,159) olduğu, benzer şekilde diğer değişkenler kontrol altında tutulduğunda da bu ilişkinin aynı düzeyde kaldığı (r=-0,153) görülmektedir.

Analiz sonuçlarına göre öğrencilerin kişisel özelliklerinden üç değişken, öğrencilerin akademik not ortalamaları ile düşük düzeyde, manidar bir ilişki göstermektedir ve birlikte akademik not ortalamalarındaki toplam varyansın yaklaşık %9'unu açıklamaktadır (R=0,307, R^2=0,094 ve p<.05). Akademik not ortalamalarının varyansına katkıları bakımından üç değişkenin önemli yordayıcılar olduğu görülmektedir. Regresyon katsayılarının karelerindeki değişim (ΔR^2) dikkate alındığında, akademik not ortalamalarının varyansına, yaş (Yaş) değişkeni %6, tam günlü bir işte çalışma (İşDurumu1) değişkeni %2, boşanmış olma (MedeniDurum2) değişkeni ise yaklaşık %1'lik katkı sağlamaktadır.

Aile ve Çalışma Ortamı Faktörlerine İlişkin Bulgular

Tablo 4. Aile ve Çalışma Ortamı Faktörlerine İlişkin Aşamalı Çoklu Regresyon Analizi Sonuçları

Değişken	B	R	ΔR^2	β	β^2	t	P	İkili r	Kısmi r
Sabit	2,488					17,010	,000		
AileBireySay [X_1]	-0,100	0,173	0,030	-0,191	,036	-3,114	,002	-,192	-,190
Aİşdurumu2 [X_2]	-1,245	0,223	0,020	-0,144	,021	-2,369	,019	-,148	-,144
AEğitdüzeyi3 [X_3]	-0,276	0,255	0,015	0,125	,016	-2,029	,043	-,127	-,124
R^2= 0,065 F=5,822 sd= 3;252 p=0,001									

Tablo 4'te verilen regresyon analizi sonuçlarına göre aile ve çalışma ortamı faktörlerinin akademik başarıyı yordamasına ilişkin regresyon eşitliği aşağıdaki gibidir:

Akademik not ortalaması= $2,488-0,100X_1- 1,245X_2-0,276X_3$

Yordayıcı değişkenlerle yordanan değişken arasındaki ikili ve kısmi korelasyonlar incelendiğinde birlikte yaşanılan aile bireyi sayısı (AileBireySay), annenin tam günlü bir işte çalışması (Aİşdurumu2) ve annenin lise mezunu olması (AEğitdüzeyi3) değişkenlerinin her biri ile akademik başarı arasında düşük düzeyde negatif ilişkilerin (r=-0,192, r=-0,148, r=-0,127) olduğu, diğer değişkenler kontrol altında tutulduğunda da bu ilişkilerin aynı düzeyde kaldığı (r=-0,153) gözlenmiştir.

Analiz sonuçlarına göre öğrencilerin aile ve çalışma ortamına ilişkin özelliklerinden üç değişken, öğrencilerin akademik not ortalamaları ile düşük düzeyde, manidar bir ilişki göstermektedir ve birlikte akademik not ortalamalarındaki toplam varyansın yaklaşık %7'sini açıklamaktadır (R=0,255, R^2=0,065 ve p<.05). Akademik not ortalamalarının varyansına katkıları bakımından üç değişkenin önemli yordayıcılar olduğu görülmektedir. ΔR^2 dikkate alındığında, birlikte yaşanılan aile bireyi sayısı (AileBireySay) değişkeni akademik not ortalamalarının varyansına%3 katkı sağlamaktadır. Annenin tam günlü bir işte çalışması (AİşdurumU2) ve annenin lise mezunu olması (AEğitdüzeyi3) değişkenlerinin her biri ise akademik not ortalamalarının varyansına yaklaşık %2'lik katkı getirmektedirler.

Bilgi ve İletişim Teknolojilerinin Kullanımına İlişkin Bulgular

Bilgi ve iletişim teknolojilerinin kullanımı ile ilgili değişkenlerin hiçbiri bağımlı değişkenin varyansına önemli bir katkı getirmediği için aşamalı çoklu regresyon analizi yapılamamıştır. Öğrencilerin bilgi ve iletişim teknolojileri ile ilgili faktörler konusundaki algılarının akademik not ortalamalarını manidar bir şekilde yordamadığı ifade edilebilir.

Eğitimle İlgili Faktörlere İlişkin Bulgular

Tablo 5'te verilen regresyon analizi sonuçlarına göre eğitimle ilgili faktörlerin akademik başarıyı yordamasına ilişkin regresyon eşitliği aşağıdaki gibidir:

Akademik not ortalaması= $1{,}545+0{,}485X_1+0{,}289X_2+0{,}281X_3-0{,}286X_4+0{,}267X_5+0{,}328X_6-0{,}236X_7+0{,}228X_8-0{,}333X_9-0{,}197X_{10}$

Yordayıcı değişkenlerle yordanan değişken arasındaki ikili ve kısmi korelasyonlar incelendiğinde denkleme giren değişkenlerinin her biri ile akademik başarı arasında düşük düzeyde ilişkilerin olduğu, diğer değişkenler kontrol altında tutulduğunda ise bu ilişkilerin düştüğü gözlenmiştir.

Analiz sonuçlarına göre öğrencilerin eğitimle ilgili özelliklerinden on değişken, öğrencilerin akademik not ortalamaları ile düşük düzeyde ve manidar bir ilişki göstermektedir ve birlikte akademik not ortalamalarındaki toplam varyansın yaklaşık %30'unu açıklamaktadır (R=0,546, R^2=0,298 ve p<.05). Akademik not ortalamalarının varyansına katkıları bakımından on değişkenin önemli yordayıcılar olduğu görülmektedir. ΔR^2 dikkate alındığında, lisans mezunu olma (LisansMez) değişkeni varyansa %9 katkı sağlamaktadır. Bu değişkeni %4 ile Adalet Önlisans Programı'nda eğitim görme (Program1) ve videoları izleyerek ders çalışma (DersNasıl3) değişkenleri izlemektedir. Derslere sanal derslere katılarak çalışma değişkeni (DersNasıl2) %3, üniversite okumak istediği için uzaktan eğitimi seçme değişkeni (NedenUE9) akademik not ortalamaları varyansına yaklaşık olarak %2'lik katkı getirmektedir. Derslere kitaplar ve fasiküllerle çalışma (DersNasıl1), önlisans programından mezun olma hedefi çalışılan işte derece alma olma (HedefUE3), konu sonlarındaki değerlendirme sorularını çözerek ders çalışma (DersNasıl5) değişkenleri varyansa yaklaşık olarak %2'lik katkı sağlamaktadırlar. Son olarak farklı üniversitelerin kaynaklarından yararlanarak ders çalışma (DersNasıl4) ve meslek lisesi mezunu olma (Lise2) değişkenleri varyansa yaklaşık %1'lik katkı getirmektedirler.

Eğitime ilişkin yordayıcı değişkenler incelendiğinde beş yordayıcı değişkenin derse nasıl çalışıldığı ile ilgili olduğu dikkat çekmektedir. Bu bulgudan yola çıkarak derse nasıl çalışıldığının akademik not ortalamalarını yordadığı yorumu yapılabilir.

Standardize edilmiş regresyon katsayılarına (β) göre, yordayıcı değişkenlerin akademik not ortalamaları üzerindeki göreli önem sırası; lisans mezunu olma, Adalet Önlisans Programı'nda eğitim görme, uzaktan eğitimi seçme nedeni "üniversitede okumak" olması, videoları izleyerek ders çalışma, kitaplar ve fasiküllerle ders çalışma, sanal derslere katılarak ders çalışma, önlisans programından mezun olma, hedefi çalışılan işte derece almak olması, konu sonlarındaki

değerlendirme sorularını çözerek ders çalışma, farklı üniversitelerin kaynaklarından yararlanarak ders çalışma ve meslek lisesi mezunu olma şeklindedir.

Tablo 5. Eğitimle İlgili Faktörlere İlişkin Aşamalı Çoklu Regresyon Analizi Sonuçları

Değişken	B	R	ΔR^2	β	β^2	t	P	İkili r	Kısmi r
Sabit	1,545					14,633	,000		
LisansMez [X_1]	0,485	,302	,091	,204	0,042	3,808	,000	,218	,187
Program1 [X_2]	0,289	,364	,041	,184	0,034	3,297	,001	,190	,162
DersNasıl3 [X_3]	0,281	,416	,041	,177	0,031	3,142	,002	,181	,155
DersNasıl2 [X_5]	0,267	,470	,025	,158	0,025	2,900	,004	,168	,143
NedenUE9 [X_4]	-0,286	,443	,023	-,179	0,032	-3,367	,001	-,194	-,166
DersNasıl1 [X_6]	0,328	,493	,022	,168	0,028	3,348	,001	,193	,165
HedefUE3 [X_7]	-0,236	,508	,015	-,155	0,024	-2,960	,003	-,171	-,146
DersNasıl5 [X_8]	0,228	,522	,015	,148	0,022	2,781	,006	,161	,137
DersNasıl4 [X_9]	-0,333	,536	,014	-,112	0,013	-2,234	,026	-,130	-,110
Lise2 [X_{10}]	-0,197	,546	,011	-,111	0,012	-2,123	,035	-,124	-,104
R^2= 0,298 F=12,306 sd= 10;290 p=0,000									

SONUÇLAR ve TARTIŞMA

Araştırma bulgularına bakıldığında eğitimsel özelliklerin, öğrencilerin akademik başarılarındaki değişkenliğin %30'unu açıklaması; belirlenen dört ana faktör içerisinden akademik başarıdaki en önemli faktör olduğunu göstermektedir. Regresyon analizinden elde edilen bulgulara bakıldığında; daha önce bir lisans bölümünden mezun olan, adalet önlisans uzaktan eğitim programında eğitim gören, eğitsel videoları izleyerek ders çalışan, uzaktan eğitimi seçme nedeni 'üniversitede okumak' olmayan, sanal derslere katılarak ders çalışan, kitaplar ve fasiküllerle ders çalışan, hedefi çalıştığı işte derece almak olmayan, konu sonlarındaki değerlendirme sorularını çözerek ders çalışan, farklı üniversitelerin kaynaklarından yararlanarak ders çalışmayan, meslek lisesi mezunu olmayan öğrencilerin akademik başarılarının diğerlerine göre daha yüksek olduğu görülmektedir.

Alanyazın eğitim hayatında hedefler belirlemenin başarıyı artırdığını göstermektedir (Papert, 1980; Gee, 1990; Garcia ve Pintrich, 1996; Lawson ve Johnson, 2002; akt: Kerr, Rynearson ve Kerr, 2006). Araştırma bulgularında 'çalışılan işte derece almak' hedefinin akademik başarıyı olumsuz yönde yordaması; işte derece almanın diploma notuyla ilişkili olmayıp, yalnızca bireyin diplomaya sahip olup olunmadığıyla ilişkili olabilmesi bu olumsuz etkinin bir açıklaması olabilir. Benzer şekilde bazı hedeflerin regresyon eşitliğine girmemesi, anketteki hedef ifadelerinin genel olması ve nispeten uzak hedeflere yer verilmesiyle açıklanabilir.

Alanyazında öğrencilerin ön eğitim durumlarının akademik başarıyı yordamasına (Darwazeh, 1998; Price, 1993) paralel biçimde lisans mezunu olmanın akademik başarıyı yordadığı görülmüştür. Araştırma sonucunda ne kadar ders çalışıldığının akademik başarılarını manidar bir şekilde yordamadığı, nasıl ders çalışıldığının öğrencilerin akademik başarılarını manidar bir şekilde yordadığı görülmüştür. Araştırma bulgularını destekler biçimde Wang ve Newlin (2002) de araştırmalarında haftada çalışılan ders saatinin final sınavından alınan notlarla ilişki göstermediğini bulmuşlardır. Romero ve Barbera'nın (2011) araştırmasında ise ders çalışmaya ayrılan zaman ve akademik performans arasında düşük düzeyde pozitif bir ilişki gözlenmiş, bulgular derse ayrılan zamandan çok; ayrılan zamanın niteliğinin önemli olduğunu vurgulamıştır. Benzer şekilde araştırma sonucunda öğrencilerin eğitsel videoları izleyerek, kitap ve fasiküllere çalışarak, konu sonlarındaki değerlendirme sorularını çözerek ders çalıştıklarında daha başarılı oldukları gözlenmiştir. Eğitsel

videolar izleyerek ders çalışmanın akademik başarıyı diğer üç ders çalışma yönteminden daha çok yordaması, Savaş'ın (2007) video destekli öğretim materyalinin animasyon destekli öğretim materyaline göre öğrenci başarısını daha olumlu etkilediğini gözlemlediği araştırmasıyla benzer bir sonuç olarak değerlendirilebilir. Öğrencilerin ne kadar değil nasıl ders çalıştıklarının önemli olması, verimli ders çalışmanın akademik başarıdaki önemine işaret etmektedir.

Öğrencilerin kişisel özellikleri akademik başarıdaki değişkenliğin yaklaşık %9'unu açıklamaktadır. Regresyon analizinden elde edilen bulgulara bakıldığında; yaşı daha büyük olan, tam günlü bir işte çalışmayan, boşanmış olan öğrencilerin akademik başarılarının diğerlerine göre daha yüksek olduğu görülmektedir. Öğrencinin tam zamanlı bir işte çalışıyor olmasının akademik başarıyı yordaması, Ross ve Powell (1990)'in bulgularını desteklemektedir. Yaş değişkeninin akademik başarıda etkili bir değişken olarak bulunması, cinsiyetin ise manidar etki yaratmaması benzer araştırmalarca desteklenmektedir (Chinnanon, 1985; Oladejo ve ark., 2010; Wang ve Newlin 2002). Alanyazında kızların erkeklere göre daha başarılı olduğuna ilişkin (Darwazeh, 1998) veya cinsiyetin akademik başarılarıyla ilişki göstermediğine ilişkin (Oladejo, Ige, Fagunwa ve Arewa, 2010; Wang ve Newlin, 2002) bulgular olduğu düşünüldüğünde, yeni araştırmalarla bu bulguların zenginleştirilmesi gerektiği ifade edilebilir.

Aile ve çalışma ortamına ilişkin özellikler akademik başarıdaki değişkenliğin yaklaşık %7'sini açıklamaktadır. Bulgulara bakıldığında; birlikte yaşadığı aile birey sayısı daha düşük olan, annesi tam günlü bir işte çalışmayan ve annesi lise mezunu olmayan öğrencilerin akademik başarılarının diğerlerine göre daha yüksek olduğu görülmektedir. Bowa (2011) çalışmasında aile büyüklüğü ile akademik başarı arasında bulduğu negatif ilişki de bu bulguyu destekler niteliktedir. Farklı öğrenci gruplarında ve geleneksel yöntemlerde de benzer bulguların olduğu görülmektedir (Gelbal, 2008). Bununla birlikte benzer araştırmalarda daha çok uzaktan eğitim öğrencisinin bireysel özelliklerine odaklanıldığı (Billings, 1989; Coldeway, 1986; Kerr, Rynearson, Kerr, 2006), aile özelliklerinin geri planda kaldığı görülmektedir.

Sonuç olarak, belirlenen dört ana faktör içerisinden sırasıyla eğitimsel özelliklerin, kişisel özelliklerin ve aile ve çalışma ortamına ilişkin özelliklerin öğrencilerin akademik başarılarını yordadığı; bilgi ve iletişim teknolojilerinin kullanımına ilişkin özelliklerin akademik başarılarını manidar bir şekilde yordamadığı ifade edilebilir.

KAYNAKÇA

Alkan, C. (1987). *Açık öğretim: Uzaktan eğitim sistemlerinin karşılaştırmalı olarak incelenmesi*, Ankara: Ankara Üniversitesi Eğitim Bilimleri Fakültesi Yayınları, Yayın No: 157.

ANKUZEM. (2011).*2010-2011 Eğitim-öğretim yılı güz yarıyılı beklenti ve memnuniyet anketi sonuçları*.Web: http://uzem.ankara.edu.tr/index.php/yayinlar.htmladresinden indirilmiştir.

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması.* Ankara: ÖSYM Yayınları.

Baykul, Y. (2005). *İlköğretimde matematik öğretimi* (8. baskı). Ankara: Pegem Akademi Yayıncılık.

Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education*, *12*, 155-187.

Billings, D. M. (1989). A conceptual model of correspondence course completion. *The American Journal of Distance Education*. *2(2)*, 23-35.

Bowa, O. (2011). The Relationship Between Learner Characteristics and Academic Performance of Distance Learners: The Case of External Degree Programme Of The University Of Nairobi. *Journal of Continuing, Open and Distance Education*. 1(*2*), 26-37.

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş. ve Demirel, F. (2009). *Bilimsel araştırma yöntemleri*. Ankara: PegemA Yayıncılık.

Chinnanon, S. (1985). *Multi-media distance education: A study of factors affecting the educational achievement of adult participants in the radio correspondence project in Thailand.* Dissertation Abstracts International, *A (*Humanities and Social Sciences). 46(4). Retrieved from Web:http://www.cabdirect.org/abstracts/19871844331.html;jsessionid=17F56B974CB77D3CC9651C9DE78C964E

Coldeway, D.O. (1986). Learner characteristics and success. In I. Mugridge, D. Kaufman (Eds.), *Distance Education in Canada*, 81-87.London: Croom Helm.

Darwazeh, A.N. (1998). *Variables affecting university academic achievement in a distance- versus a conventional education setting*. Proceedings of Selected Research and Development Presentations at the National Convention of the Association for Educational Communications and Technology (AECT). Retrieved from Web: http://files.eric.ed.gov/fulltext/ED423833.pdf

Driscoll, M. (2002).*Web-based training: Creating e-learning experiences* (2. edition). San Francisco, CA:Jossey-Bass/Pfeiffer.

Gelbal, S. (2008). Sekizinci Sınıf Öğrencilerinin Sosyoekonomik Özelliklerinin Türkçe Başarısı Üzerinde Etkisi. *Eğitim ve Bilim*, 33 (150).

Hughes, M. & Forest, S. (1997). Distance education in early intervention personnel preparation. In P. J. Winton, J. A. McCollum, & C. Catlett (Eds.), *Reforming personnel preparation in early intervention: Issues, models, and practical strategies(pp. 475-494).* Baltimore: Paul H. Brookes.

Kalaycı, Ş. (2009). *SPSS uygulamalı çok değişkenli istatistik teknikleri*. Ankara: Asil Yayıncılık.

Kember, D. (1995). *Open learning courses for adults: A model of student progress.* Englewood Cliffs, NJ: Educational Technology Publications.

Kerr, M. S., Rynearson, K. & Kerr, M. C. (2006). Student characteristics for online learning success. *Internet and Higher Education, 9*, 91-105.

Keskin, G. ve Sezgin B. (2009). Bir grup ergende akademik başarı durumuna etki eden etmenlerin belirlenmesi. *Fırat Sağlık Hizmetleri Dergisi*, *4*(10), 3-18.

Koçer, H. E. (2001). *Web tabanlı uzaktan eğitim (Yayımlanmamış* Yüksek lisans tezi, Selçuk Üniversitesi, Konya). http://tez2.yok.gov.tr/ adresinden edinilmiştir.

Kutlu, Ö., Büyüköztürk, Ş. ve Doğan, C. (2007). *İlköğretim öğretmenlerinin yeni değerlendirme yöntemlerine yönelik tutumlarını etkileyen faktörler*. 16.Ulusal Eğitim Bilimleri Kongresi, Tokat, Türkiye.

Memduhoğlu, H. B. ve Tanhan, F. (2009, Mayıs). *Üniversite öğrencilerinin akademik başarılarını etkileyen örgütsel faktörler ölçeğinin geliştirilmesi*. I. Uluslararası Eğitim Araştırmaları Kongresi, Çanakkale, Türkiye.

Moore, M. G. and Kearsley, G. (2005).*Distance education: A systems view*. (2. edition). Belmont, CA: Wadsworth Publishing Company.

Neto, J. D. O. & Santos, E. M. (2010).Analysis of the methods and research topics in a sample of the Brazilian distance education publications 1992 to 2007.*AmericanJournal of Distance Education, 24*(3), 119-134.

Oladejo, M. A., Ige, N. A., Fagunwa, A. O. and Arewa, O. O. (2010). Socio-demographic variables and distance learners' academic performance at the University of Ibadan, Nigeria, *European Journal of Scientific Research*, *46*(4), 540-553.

Özgüven, İ. E. (1974). *Üniversite Öğrencilerinin akademik başarılarını etkileyen zihinsel olmayan faktörler*. Hacettepe Üniversitesi Yayınevi, Ankara.

Price, M. (1993). *Student satisfaction with distance education at the University of South Carolina as it correlates to medium of instruction, educational level, gender, working status, and reason for enrollment* (Unpublished PH.D. degree Dissertation).University of South Carolina,Columbia.Retrievedfromhttp://www.ntlf.com/html/lib/umi/92-93e.htm

Romero, M. and Barbera, E. (2011).Quality of e-learners' time and learning performance beyond quantitative time-on-task. *The International Review of Research in Open and Distributed Learning.12*(5).

Ross, L. R. and Powell, R. (1990). Relationships between gender and success in distance education courses: A preliminary investigation, Research in distance education, *Research in Distance Education*, 2(2), 10-11.

Savaş, S. (2007). *Web tabanlı uzaktan eğitimde iki farklı öğretim modelinin öğrenci başarısı üzerindeki etkilerinin incelenmesi* (Yayımlanmamış Yüksek lisans tezi, Gazi Üniversitesi, Ankara). http://tez2.yok.gov.tr/ adresinden edinilmiştir.

Seal, K. C. & Przasnyksi, Z. H. (2001). Using the world wide web for teaching improvement. *Computers and Education, 36*, 33-40.

Sikora, A.C. (2002). A Profile of Participation in Distance Education: 1999-2000. Post secondary Education Descriptive Analysis Reports. *National Center for Education Statistics (ED),*Washington, DC.; MPR Associates, Berkeley, CA.

Tinto, V. (1975).Dropout from higher education: A theoretical synthesis of recent research. *Review of EducationalResearch,45*(1), 89-125.

Wang, A. Y. and Newlin, M. H. (2002). Predictors of web student performance: The role of self-efficacy and reasons for taking an on-line class. *Computers in Human Behavior, 18*(2), 151-163.

Whittington, L. A. (1995).*Factors impacting on the success of distance education students of the University of the West Indies: A review of the literature*. University of the West Indies: Barbados. Retrievedfromhttp://eric.ed.gov/?id=ED453740
Wolman, B. (1973). *Dictionary of behavioral science*. New York: Van Nostrand Company.

EXTENDED ABSTRACT

Introduction

Students' demographic characteristics, perceptions and attitudes against distance education must be clearly identified for getting sufficient and effective results in distance education environments. Designing of distance education environment is another critical factor for satisfying student needs. Attendees of distance education are generally working adults. Since working adults have high-demanding working hours, families, social life and other responsibilities they are demanding educational options which have distance methods and materials used. Disadvantageous side of the working adults is that interactions or practices of education are not their prior or firstly-ranked objective in their life. In the first steps of their distance education expertise, their musts and other responsibilities may affect them negatively. From this perspective, it looks important for students, designers and implementers to consider and understand the factors that may influence the success of distance education environments. In this study individual factors, family and working environments factors, information and communication technology factors and educational background factors relation with academic achievement is examined. Results of the study will be used as a source for designers, implementers and students in designing distance courses, programs and generating support systems for distance education.

The purpose of this study was to determine the factors predicting academic achievement of Ankara University Distance Education associate degree programs' students. With this aim research question given below were used:
1. Are individual factors (gender, age, working status, marital status and children number) predictors of academic achievement?
2. Are the characteristics related with family and working environment (educational status of parents, number of family members living with, working status of parents, monthly income of family, utilities in house environment like computer, internet access, library) predictors of academic achievement?
3. Are the characteristics related with usage of information and communication technologies (year number of computer and internet expertise, computer and internet usage level, average usage hour of computer and internet, computer and internet usage environment) predictors of academic achievement?
4. Are the characteristics related with education (type of high school graduation, having a prior undergrad or associate degree, having a prior distance education program degree, reason for choosing to study from distance, average studying hour per day, studying methods, aim of graduation from distance education) predictors of academic achievement?

Method

Some questions were sent to participants and answers collected by e-mail for generating the main data collection tool. Answers of 20 participants were analyzed by using content analysis. Data collection tool were prepared based on the results of content analysis and the literature related with the study. 4 main factors are considered in the data collection tool according to content analysis and literature review. These are;

1. Individual factors
2. Characteristics related with family and working environment
3. Usage of information and communication technologies

4. Educational background

The study group of the research consists of 302 1st grade students from Ankara University Distance Education associate degree programs. Stepwise multiple repression analysis, which is a reasonable statistical method for prediction researches, is used for analyzing data. Dependent variable is cumulative point averages of students as academic achievements. Independent variables of the study are generally categorical ones. Categorical variables that are used in the study were coded as "dummy variables" when analyzing data but the continuous variables are coded with their original values.

In the data analyzes part, three different equations were formulated, for each of three factors of individual factors, factors related with family and working environment and factors related with educational background.

Results and Discussion

Results of the study indicated that, three variables of the individual factors as "age", "working status" and "marital status" are found to explain academic achievement of students with the %9 of total variance. Three variables of characteristics related with family and working environment as "number of family members living with", "working status of parents" and "educational status of parents" explained academic achievement of students with the %7 of total variance. None of the variables related with usage of information and communication technologies (ICT) predicts academic achievement of students. Variables related with educational background as "type of high school graduation", "studying methods", "having a prior distance education program degree", "reason for choosing study from distance" and "the aim of graduation from distance education" explained academic achievement of students with the %30 of total variance.

When the results of the research examined, it is possible to state that students' studying methods like watching educational videos, studying from books and printed materials and solving end of unit evaluation questions are meaningfully predictors of academic achievement in distance learning environments. Another important result of the study is that, studying hours per day is not a predictor of academic achievement. From this point of view we may summarize that; not the period of time students spent for studying but the method they used is meaningful for the academic achievement. Another founding is that there is no variance explained by ICT usage abilities of the students. It must be consider that ICT usage abilities of students is not measured by an academic achievement test but with a self-reporting tool and data actually represented the perceptions of the students about their ICT abilities.

EK 1

DUMMY DEĞİŞKENLERİN KODLANMASI

Süreksiz Değişkenler	Düzey	Dummy Değişken	Kodlama	Dışta Tutulan Kategori
Cinsiyet	1. Kız 2. Erkek	Cinsiyet	Kız:1 Erkek:0	Erkek
Medeni Durum	1. Bekar 2. Evli 3. Boşanmış	Medenidurum1 Medenidurum2	Evli:1 Diğer:0 Boşanmış: 1 Diğer:0	Bekar
Çocuk Sayısı	1. Çocuk yok 2. 1 çocuk 3. 2 çocuk 4. 3 ve üstü çocuk	Cocuk1 Cocuk2 Cocuk3	Bir Çocuk:1 Diğer:0 İki Çocuk:1 Diğer:0 3 ve üstü çocuk:1 Diğer:0	Çocuk Yok
Çalışma Durumu	1. İşsiz 2. Tam günlü bir işte 3. Yarı zamanlı bir işte 4. Haftanın bazı günleri 5. İş buldukça	İşdurumu1 İşdurumu2 İşdurumu3 İşdurumu4	Tam günlü:1 Diğer:0 Yarı zamanlı:1 Diğer:0 Bazı günler:1 Diğer: 0 İş buldukça:1 Diğer:0	İşsiz
Anne Eğitim	1. Okuryazar değil 2. İlkokul 3. Ortaokul 4. Ortaöğretim (lise) 5. Yükseköğretim	AEğitdüzeyi1 AEğitdüzeyi2 AEğitdüzeyi3 AEğitdüzeyi4	İlkokul:1 Diğer:0 Ortaokul:1 Diğer: 0 Ortaöğretim (lise):1 Diğer:0 Yükseköğretim:1 Diğer:0	Okuryazar değil
Baba Eğitim	1. Okuryazar değil 2. İlkokul 3. Ortaokul 4. Ortaöğretim (lise) 5. Yükseköğretim	BEğitdüzeyi1 BEğitdüzeyi2 BEğitdüzeyi3 BEğitdüzeyi4	İlkokul:1 Diğer:0 Ortaokul:1 Diğer: 0 Ortaöğretim (lise):1 Diğer:0 Yükseköğretim:1 Diğer:0	Okuryazar değil
Anne Çalışma Durumu	1. İşsiz 2. Tam günlü bir işte 3. Yarı zamanlı bir işte 4. Haftanın bazı günleri	AİşdurumuI AİşdurumuI2 AİşdurumuI3 AİşdurumuI4	Tam günlü:1 Diğer:0 Yarı zamanlı:1 Diğer:0 Bazı Günler:1 Diğer:0	İşsiz
Baba Çalışma Durumu	1. İşsiz 2. Tam günlü bir işte 3. Yarı zamanlı bir işte 4. Haftanın bazı günleri 5. İş buldukça	BİşdurumuI BİşdurumuI2 BİşdurumuI3 BİşdurumuI4	Tam günlü:1 Diğer:0 Yarı zamanlı:1 Diğer:0 Bazı Günler:1 Diğer: 0 İşbuldukça:1 Diğer:0	İşsiz
Aile Aylık Gelir Ortalaması	1. 1000 TL'den az 2. 1001–2000 TL arası 3. 2001–3000 TL arası 4. 3001–4000 TL arası 5. 4001–5000 TL arası 6. 5001 TL üstü	Gelir1 Gelir2 Gelir3 Gelir4 Gelir5	1000 TL'den az:1 Diğer:0 1001–2000 TL arası:1 Diğer:0 2001–3000 TL arası:1 Diğer:0 3001–4000 TL arası:1 Diğer:0 4001–5000 TL arası:1 Diğer:0	5001 TL üstü
Evdeki Olanaklar	1. Bilgisayar 2. İnternet bağlantısı 3. Kitaplık 4. Çalışma masası 5. Çalışma odası	Evde1 Evde2 Evde3 Evde4 Evde5	Var: 1 Yok:0	Yok
Bilgisayar Kullandığı Süre	1. 1-4 yıl arası 2. 5-8 yıl arası 3. 9-12 yıl arası 4. 13 yıl ve üstü	BilgisayarSüre1 BilgisayarSüre2 BilgisayarSüre3	1-4 yıl:1 Diğer:0 5-8 yıl:1 Diğer: 0 9-12 yıl:1 Diğer:0	13 yıl ve üstü
İnternet Kullandığı Süre	1. 1-4 yıl arası 2. 5-8 yıl arası 3. 9-12 yıl arası 4. 13 yıl ve üstü	İnternetSüre1 İnternetSüre2 İnternetSüre3	1-4 yıl:1 Diğer:0 5-8 yıl:1 Diğer:0 9-12 yıl:1 Diğer:0	13 yıl ve üstü
Bilgisayar Kullanım Düzeyi	1. Başlangıç 2. Orta 3. İyi	BilgDüzey1 BilgDüzey2 BilgDüzey3	Orta:1 Diğer:0 İyi:1 Diğer:0 Mükemmel:1 Diğer:0	Başlangıç

	4. Mükemmel			
İnternet Kullanım Düzeyi	1. Başlangıç 2. Orta 3. İyi 4. Mükemmel	İntDüzey1 İntDüzey2 İntDüzey3	Orta:1 Diğer:0 İyi:1 Diğer:0 Mükemmel:1 Diğer:0	Başlangıç
Bilgisayar Erişim Ortamları	1. Evden 2. İşyerinden 3. İnternet kafeden	BHangiOrtam1 BHangiOrtam2 BHangiOrtam3	Evet:1 Hayır:0	Hayır
İnternet Erişim Ortamları	1. Evden 2. İşyerinden 3. İnternet kafeden	İHangiOrtam1 İHangiOrtam2 İHangiOrtam3	Evet:1 Hayır:0	Hayır
Program	1. Adalet (ADUZEP) 2.Bankacılık ve Sigortacılık (BAS) 3.Bilgisayar Programcılığı (BİPRO) 4.Turizm ve Otel İşletmeciliği (TOİ)	Program1 Program2 Program3	ADUZEP:1 Diğer:0 BAS:1 Diğer:0 BIPRO:1 Diğer:0	TOI
Lisans Mezunu	1. Evet 2. Hayır	LisansMezun	Evet:1 Hayır:0	Hayır
Önlisans Mezunu	1. Evet 2. Hayır	ÖnLisansMezun	Evet:1 Hayır:0	Hayır
Daha Önce Uzaktan Eğitim Alma	1. Evet 2. Hayır	ÖnceUzaktan	Evet:1 Hayır:0	Hayır
Günde Çalışılan Ders Saati	1. Hiç 2. 1 saat 3. 2-3 saat arası 4. 4-5 saat arası 5. 6 saat ve üstü	DersSaat1 DersSaat2 DersSaat3 DersSaat4	Hiç:1 Diğer:0 1 saat:1 Diğer:0 2-3 saat:1 Diğer:0 4-5 saat:1 Diğer:0	6 saat ve üstü
Mezun Olunan Lise	1.Lise (Resmi ve Gündüz Öğretimi Yapan Liseler) 2. Meslek Liseleri 3. Anadolu Meslek Lisesi 4. Yabancı Dil Ağırlıklı Öğretim Yapan Liseler	Lise1 Lise2 Lise3	Lise:1 Diğer:0 Meslek Lisesi:1 Diğer:0 Anadolu Meslek Lisesi:1 Diğer:0	Yabancı Dil Ağırlıklı Öğretim Yapan Liseler

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2015; 6(2); 207-217

Tek ve Çok Boyutlu Madde Tepki Kuramına Dayalı Bir Veri Analizi Yazılımı: IRTPRO 2.1

A Data Analysis Software Based on Uni- and Multidimensional Item Response Theory: IRTPRO 2.1

Esin YILMAZ KOĞAR * Derya ÇAKICI ESER **

Öz
Bu çalışmada, farklı madde tepki kuramı (MTK) modellerine göre analiz yapmaya olanak sağlayan "IRTPRO 2.1 for Windows" adlı bilgisayar programını tanıtmak amaçlanmıştır. Bu doğrultuda çalışmada IRTPRO 2.1 yazılımı için gerekli donanım özellikleri, programa erişim, programın çalıştırılması, analizler, çıktı dosyaları, uyum iyiliği indeksleri ve betimsel istatistikler üzerinde durulmuştur. Ayrıca farklı yapılara ilişkin 20 madde ve 2000 kişilik simülatif verilerle örnek uygulama yapılmıştır. Elde edilen sonuçlardan yola çıkarak farklı IRT modellerini analiz etmek için gerekli kestirim süresi hakkında araştırmacılara bilgi vermek amaçlanmıştır. Yapılan derlemeler ve araştırmacıların deneyimleri sonucunda elde ettikleri bilgiler yardımıyla IRTPRO 2.1 hakkında bir ön bilgi verilmeye çalışılmış, programın avantajlı ve dezavantajlı yönleri tartışılmıştır. IRTPRO 2.1; MTK'ya dayalı veri analizi yapan programların yerine getirebildiği fonksiyonların tek bir programla yapılabilmesini sağlaması bakımından kullanışlı bir yazılımdır. Çalışmalarında bu yazılımdan yararlanmak isteyen araştırmacılara, üstünlükler ve sınırlılıklar açısından yazılımın deneme sürümünü incelemeleri ve bundan sonra lisansı satın alıp almamaya karar vermeleri önerilmektedir.

Anahtar Kelimeler: Anahtar Kelimeler: Madde tepki kuramı, veri analizi, IRTPRO 2.1, bilgisayar yazılımı

Abstract
The purpose of this study is to introduce "IRTPRO 2.1 for Windows" computer software which enables to perform analysis based on different item response theory (IRT) models. In this context, the study is focused on the necessary hardware features for IRTPRO 2.1 software, access to the program, running the program, analysis, output files, the goodness of fit index and descriptive statistics. In addition, sample applications are made with simulative data belonging different structures including 20 items and 2000 examinees. Based on the obtained results it is intended to provide information about the amount of time needed to analyze different IRT models. As a result of compilations and experiences of researchers, a prior knowledge have tried to give about IRTPRO 2.1, the advantages and disadvantages of the program were discussed. IRTPRO 2.1 is a practical software and allows to execute functions, which can be performed by several softwares for data analysis based on IRT, with one program. Researchers who wish to make use this program in their research are recommended to investigate the trial version of the software in terms of strengths and limitations and then decide whether to buy a license.

Keywords: Item response theory, data analyses, IRTPRO 2.1, computer software

GİRİŞ

Klasik Test Kuramı'nın (KTK) yeterli olamadığı noktalarda test uygulayıcılarına ve araştırmacılara çözümler sunmak amacıyla geliştirilen Madde Tepki Kuramı (MTK) test eşitleme, değişen madde

* Araş. Gör., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, esinyilmazz@gmail.com
** Dr., Kırıkkale Üniversitesi, Eğitim Fakültesi, Kırıkkale-Türkiye, deryacakicieser@gmail.com

Geliş Tarihi: 26.05.2015
Kabul Tarihi: 20.10.2015

fonksiyonu ve bilgisayara dayalı bireyselleştirilmiş test uygulamalarında da KTK'ya karşı güçlü bir kuramdır. Bu kurama dayalı yapılacak uygulamalar için farklı algoritmaları kullanan birçok bilgisayar programları geliştirilmiştir. Uzun yıllardır MTK'ya dayalı madde ve yetenek parametresi kestirimleri ve farklı MTK uygulamaları için BILOG-MG (Zimowski, Muraki, Mislevy ve Bock, 2003), MULTILOG (Thissen, 2003), PARSCALE (Muraki ve Bock, 2003), WINSTEPS (Linacre, 2011) ve Xcalibre gibi yazılımlar kullanılmaktadır. Ayrıca açık kaynaklı ücretsiz bir yazılım olan R yazılımının Itm, eRm, plRasch, plink ve mirt (Rusch, Mair ve Hatzinger, 2013; Rizopoulos, 2015) paketlerinde tanımlanmış kodlar aracılığıyla çeşitli MTK uygulamaları gerçekleştirilebilmektedir. Bu yazılımlardan BILOG-MG, R'ın eRm ve plRasch paketleri iki kategorili verileri, MULTILOG ve R'ın plink paketi çok kategorili verileri analiz etmek için kullanılırken; PARSCALE, WINSTEPS, Xcalibre ve R'ın Itm paketi ile hem çok kategorili hem iki kategorili veriler analiz edilebilmektedir (Rizopoulos, 2015).

MTK, tek boyutluluk varsayımına dayanmaktadır. Bu nedenle bir veri seti üzerinde MTK'ya dayalı bir uygulama yapmak için ölçme aracının baskın bir faktörü ölçtüğünün gösterilmesi gereklidir (Hambleton ve Swamington, 1985; Embretson ve Reise, 2000). Ancak yapılan ölçme işlemlerinde tek boyutluluk varsayımı her zaman sağlanamayabilir ya da yapılan ölçme işlemi ile birden fazla özelliğin ölçülmesi amaçlanabilir. Bu sebeplerle, ortaya atıldığı dönemden 1970'lerin sonuna kadar tek boyutlu testler için kullanılan MTK, 1970'lerin sonu ve 1980'lerin başından itibaren çok boyutlu testlere genişletilmiş ve çok boyutlu madde tepki kuramı (ÇBMTK) başlığı altında birden fazla yeteneği ölçen testlerde de uygulanmaya başlanmıştır (Ansley ve Forsyth, 1985; Reckase, 2009). Çok boyutlu testlerde madde ve birey parametreleri tek boyutlu MTK'dan farklı biçimde; ÇBMTK'ya özgü eşitliklerinden yararlanarak kestirilmektedir. Bu sebeple ortaya konan yeni kuram ile yeni yazılımların geliştirilmesi ve kullanılması ihtiyacı doğmuştur.

Verileri ÇBMTK'ya dayalı olarak analiz etmek için kullanılan yazılımlardan TESTFACT (Bock vd., 2003), NOHARM (Fraser ve McDonald, 1988) BMIRT II (Yao, 2003), IRTPRO 2.1 (Cai Thissen ve du Toit, 2011) ve flexMIRT 2 (Cai, 2013) literatürde sıkça kullanılmaktadır. Ancak programlar çeşitli açılardan birbirinden farklılaşmaktadır. Buna göre NOHARM ve BMIRT ücretsiz, TESTFACT, IRTPRO 2.1 ve flexMIRT 2 ticari yazılımlardır. NOHARM ile madde parametrelerinden ayırt edicilik ve güçlük kestirilmekte, ancak yetenek parametresi kestirimi yapılamamaktadır. Buna karşılık TESTFACT, BMIRT, IRTPRO 2.1 ve flexMIRT 2'de hem madde hem de birey parametreleri analiz edilebilmektedir. Ayrıca programların parametre kestirimi için kullandığı yöntemler de farklılık göstermektedir. Çok boyutlu MTK'ya dayalı kestirim yapan programların IRTPRO 2.1 ile karşılaştırması Tablo 1'de özet olarak verilmiştir.

Tablo1. Çok Boyutlu MTK Kuramına Dayalı Veri Analizi Programlarının Karşılaştırılması

Özellikler	TESTFACT	NOHARM	BMIRT	FlexMIRT 2	IRTPRO 2.1
Şans parametresi kestirimi	✓	X	✓	✓	✓
Birey parametresi kestirimi	✓	X	✓	✓	✓
İki kategorili veriler	✓	✓	✓	✓	✓
Çok kategorili veriler	✓	X	✓	✓	✓
Profesyonel el kitabı	✓	X	X	✓	✓
Komut ile çalışma	✓	✓	✓	✓	✓
Tıklayarak (menüden seçim ile) çalıştırma	X	X	X	X	✓

Tablo 1'e göre NOHARM şans parametresi ile birey parametresini kestirememekte ve çok kategorili verileri analiz edememektedir. BMIRT'in ve NOHARM'ın profesyonel bir el kitabı bulunmamaktadır. Tabloda yer alan tüm yazılımlardan IRTPRO 2.1 dışında kalanlar yalnızca yazılan komut ile çalışmaktadır. Ancak IRTPRO 2.1 hem komut ile hem de menüden seçim yapılarak çalıştırılabilmektedir.

IRTPRO 2.1 (Item Response Theory for Patient-Reported Outcomes) programı Li Cai, David Thissen ve Stephen du Toit tarafından 2011 yılında geliştirilmiştir. IRTPRO madde kalibrasyonu ve test puanlama için MTK'yı kullanan istatistiksel bir yazılımdır. IRTPRO 2.1'de aşağıda maddeler halinde verilen MTK modelleri kullanılmaktadır (Cai, Thissen ve du Toit, 2011):

- İki parametreli lojistik model (2PL) [eşit ayırt edicilik parametresi kullanılması ile bir parametreli lojistik model haline gelir (1PL)]
- Üç parametreli lojistik model (3PL)
- Aşamalı Tepki Modeli
- Genelleştirilmiş Kısmi Puan Modeli
- Adlandırmalı Tepki Modeli

Bu çalışmada, yukarıda belirtilen farklı madde tepki kuramı modellerine göre analiz yapmaya olanak sağlayan "IRTPRO 2.1 for Windows" adlı bilgisayar programını tanıtmak amaçlanmıştır. Bu amaçla çalışmada programın kurulumu, analizler, çıktılar ve kestirim süreleri hakkında genel bilgilere yer verilmiştir. IRTPRO 2.1 ile belirtilen modellere ilişkin hem tek boyutlu hem çok boyutlu testler analiz edilebilmektedir. Ayrıca bu yazılım ile araştırmacının amacına ve ölçme aracının yapısına bağlı olarak bu modellerin farklı kombinasyonlarının yer aldığı bir testin analizini yapmak mümkündür.

IRTPRO 2.1 yazılımı yurt dışında oldukça sık karşımıza çıkan bir yazılım olmasına karşılık, yurt içinde yapılan MTK ve ÇBMTK çalışmasında oldukça az kullanılmıştır. Bu durumun nedeni olarak IRTPRO 2.1 yazılımının yurt içinde yeteri kadar tanınmaması görülmektedir. IRTPRO 2.1'in tanıtılmasını amaçlayan herhangi bir yurt içi çalışmanın olmaması ve bu çalışmada MTK tabanlı çalışma yapmak isteyen araştırmacılara yardımcı olacak bilgilerin sunulması sebebiyle çalışmanın önem taşıdığı düşünülmektedir. Kulanım kolaylığına sahip bu programın tanıtımı ile özellikle MTK tabanlı çalışma yapmak isteyen araştırmacılara yardımcı olacak bilgiler sunulmaktadır. Bu bilgiler ışığında da araştırmacılar gelecekte yapacakları çalışmaları planlayabileceklerdir. Araştırmanın bu bakımlardan alan yazına katkı getireceği düşünülmektedir.

PROGRAMIN KURULMASI, ANALİZLER, ÇIKTILAR ve KESTİRİM SÜRELERİ

Donanım Gereksinimleri ve Programa Ulaşım

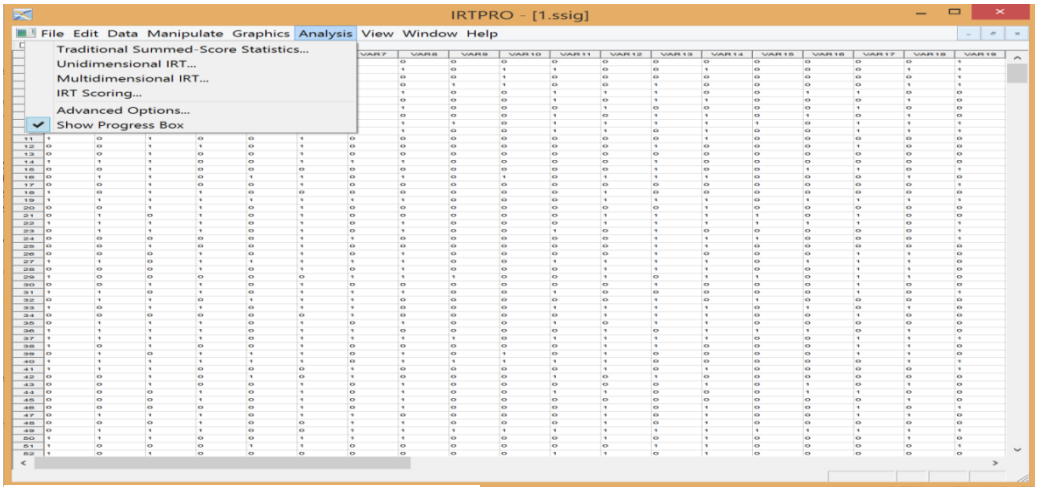
Program, Windows7, Vista ve XP işletim sistemleri ile kullanılabilmektedir. Programa http://www.ssicentral.com adresinden ulaşılabilir. Yazılımın en fazla 25 madde, 1000 kişi ve 3 boyut ile çalışabilen öğrenci versiyonuna ve 15 günlük deneme sürümüne aynı adresten ücretsiz olarak erişilmektedir. Akademik olan ve akademik olmayan kullanıcılara yönelik altı aylık ve on iki aylık tam sürümlerin 2015 yılındaki ücreti 75$ ile 950$ arasında değişmektedir. Programın yazılım anahtarı, ücretin internet üzerinden ödenmesinden sonra elektronik posta yoluyla gönderilmektedir. Programın ücretli versiyonlarını kullanan kişilere teknik destek hizmeti de verilmektedir. Tam sürümde maksimum madde sayısı ile ilgili bir limit olmasa da bu durum sistemin kullanılabilir hafızası ile sınırlandırılmıştır ve teorik olarak 4GB yeterli görülmektedir (Han ve Paek, 2014).

Analizlerin Yapılması

IRTPRO 2.1 programı bilgisayara yüklendikten sonra IRTPRO ikonu tıklandığında açılan ana menüde "File", "View" ve "Help" seçenekleri bulunmaktadır. Programa "File" sekmesinden elle veri girişi yapılabildiği gibi farklı formatlardaki verilerin aktarımı da sağlanabilmektedir. Farklı formatlı verileri kullanmak için "Import Data" sekmesi seçilir. Bu sekme ile fiks formatlı (.fixed), virgül ile ayrılmış (.csv), boşluk ile ayrılmış (.txt), Excel (.xls) ve SPSS (.sav) gibi dosyalarından veri alınabilir. Program analiz edilecek veriyi, .ssig uzantılı olarak kaydeder. Veri setinde yer alan kayıp veriler de program üzerinden tanımlanabilir. Programın ana ekranı Şekil 1'deki gibidir.

Veri girişi yapıldıktan sonra program, "Analysis" sekmesi altında yer alan (a) geleneksel toplam-puan istatistikleri, (b) tek boyutlu MTK, (c) çok boyutlu MTK, (d) MTK puanlama olmak üzere dört temel analiz türünden seçileni gerçekleştirmektedir. Ayrıca "Analysis" sekmesinde "Advanced Options" ve "Show Progress Box" adlı iki seçenek daha bulunmaktadır (Şekil 2). "Show Progress Box" seçeneği varsayılan olarak işaretlidir. Bu sayede açılan pencere ile programın analizin hangi aşamasında olduğu görülebilir. "Advanced Options" seçeneği analiz penceresindeki "Options" ile aynı içeriktedir ve bu seçenek, ilerleyen kısımda daha ayrıntılı olarak anlatılacaktır.

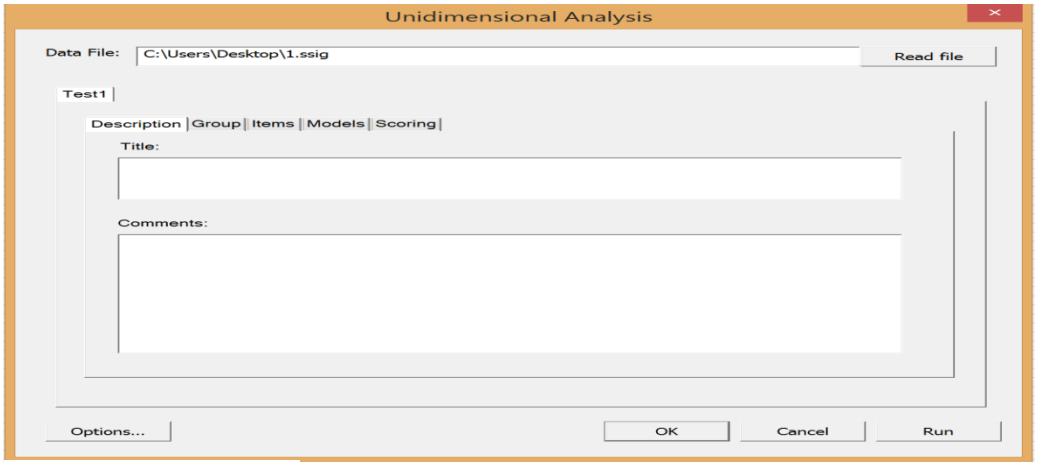
Şekil 1. IRTPRO Ana Ekranı



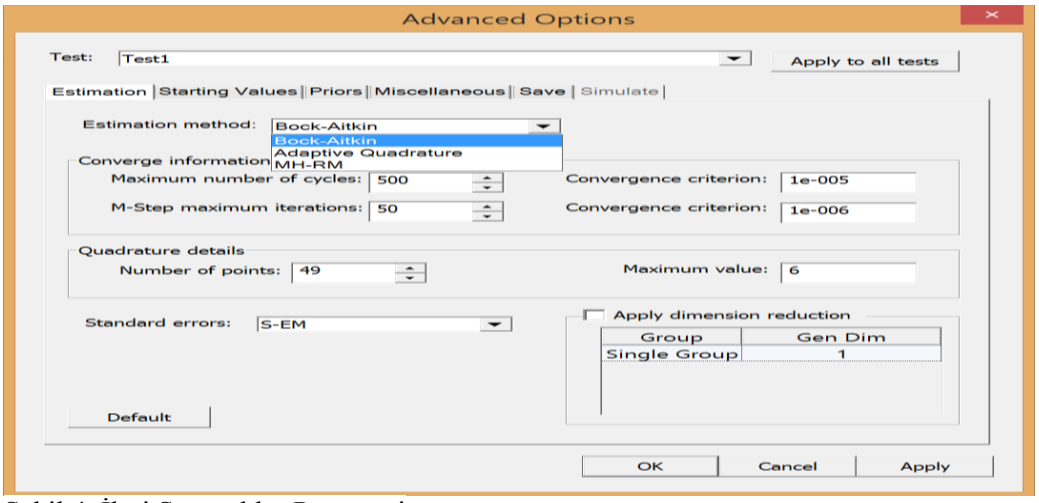
Şekil 2. IRTPRO 2.1'de Analiz Sekmesi

Analiz sekmesinden yapılacak analiz seçildiğinde "Tanımlama" (Description), "Grup" (Group), "Maddeler" (Items), "Modeller" (Models) ve "Puanlama" (Scoring) sekmelerinin yer aldığı pencere açılır (Şekil 3). Bu penceredeki "Tanımlama" kısmından analize isim verilebilir ve analiz ile ilgili yorumlar yazılabilir. Eğer analizler farklı gruplar üzerinden yapılacaksa "Grup" sekmesinden verilerin neye göre gruplandırıldığı ve çoklu grup söz konusu ise referans grubun hangisi olduğu belirtilmelidir. Ancak yapılacak analizde grup değişkeni söz konusu değilse bu sekme atlanabilir. "Maddeler" sekmesinde ise analize hangi maddelerin alınacağı ve çok boyutlu analiz yapıldığı durumda boyut sayısı girilir. "Modeller" sekmesinden 2 parametreli lojistik (2PL), 3 parametreli lojistik (3PL), Aşamalı (Graded), Genelleştirilmiş Kısmi Puan (GPCredit), Adlandırmalı Tepki (Nominal) modellerinden biri seçilir. Eğer analiz edilen veri iki kategorili bir veri ise Aşamalı (Graded), Genelleştirilmiş Kısmi Puan (GPCredit), Adlandırmalı Tepki (Nominal) seçenekleri kendiliğinden inaktif hale gelecektir. Ayrıca tek boyutlu analizlerde parametre sınırlandırmaları ve değişen madde fonksiyonu için maddelerin tanımlanmasına ilişkin seçimler de buradan yapılabilir. Çok boyutlu analizlerde parametre sınırlandırmaları yoluyla doğrulayıcı faktör analizi, açımlayıcı faktör analizi ve iki-faktör analizlerine ilişkin seçimler de bu sekmeden yapılmaktadır. "Puanlama" sekmesinden yanıt örüntülerini hesaplama yöntemi (EAP ya da MAP), ölçeklemede kullanılacak ortalama, standart sapma, minimum, maksimum, ölçekleme yapılacak dağılım (evren-örneklem)

tanımlanır. Son olarak "Seçenekler" (Options) sekmesi tıklanarak kestirim yöntemleri, başlangıç parametreleri, önseller (prior), kaydedilecek çıktılara ilişkin varsayılan değerler değiştirilebilir (Şekil 4).



Şekil 3. Analiz Penceresi



Şekil 4. İleri Seçenekler Penceresi

Çok boyutlu MTK'ya dayalı veri analizi yapan programlardan TESTFACT'da MAP ve Bayes EAP; NOHARM'da normal ogive modele dayalı olarak en küçük kareler yöntemi, BMIRT'ta MCMC (Markov Chain Monte Carlo), flexMIRT 2'de maksimum marjinal olabilirlik, BA-EM (Bock Aitkin Expectation Maximization) ve MH-RM (Metropolis Hastings Robbins Monro) ile kestirim yapılmaktadır. Bunlara karşılık IRTPRO 2.1 programında madde kalibrasyonu için maximum olabilirlik (Maximum Likelihood-ML) ya da madde parametrelerinin önsel dağılımları tanımlandığı takdirde Maximum a posteriori (MAP) yöntemi kullanılmaktadır. Bu yöntemlerle beraber boyutluluk ve modelin yapısının farklı kombinasyonlarına göre en iyi performansı sağlamak üzere Bock-Aitkin Expectation-Maximization (BA-EM), Adaptive Quadrature (ADQEM) ve Metropolis-Hastings Robbins-Monro (MH-RM) kestirim yöntemleri bulunmaktadır.

Program tarafından otomatik olarak seçili olan Bock-Aitkin EM, tek ve iki boyutlu analizler için kullanılması önerilen bir kestirim yöntemidir. Araştırmacılar, Bock-Aitkin EM'de bulunan quadrature noktaların sayısını, bu noktaların hangi aralık üzerinde yayıldığını, maksimum döngü sayısını (E-step), maksimum iterasyon sayısını (M-step) değiştirebilirler. Standart hata kestirimi içinse S-EM, M-Step, Xpd, Sandwich olmak üzere dört farklı seçenek bulunmaktadır. Bu kestirim

yöntemi birçok tek boyutlu ve iki faktör MTK modellerinde parametre kestirimi için etkilidir (Cai vd., 2011).

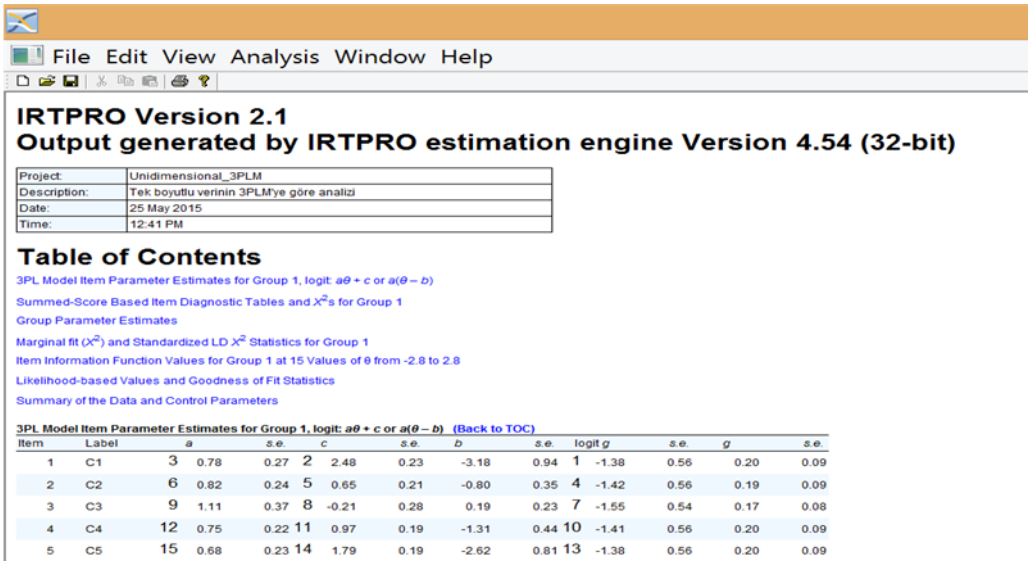
Bock-Aitkin EM algoritmasında bütün bireylerin önsel dağılımları için sabit bir quadrature node olması, kullanılan standart numerik quadraturelar ile ilgili bir problem oluşturmaktadır. Bu durum olabilirlik değerlerini hesaplamada çok sayıda quadrature noktasının kullanımını gerektirir. Bu problemle başa çıkmak için IRTPRO adaptive quadrature denilen sayısal bir integrasyon süreci önermektedir. Bu süreç deneysel Bayes ortalamalarını ve kovaryanslarını kullanmaktadır (Cai vd., 2011). MH-RM algoritması ise boyut sayısı iki ya da üçü aşan çok boyutlu MTK uygulamalarında kullanılmak üzere geliştirilmiştir.

Program ile ölçek puanlarını hesaplamak üzere üç tür MTK kestirim yöntemi önerilmektedir. Buna göre kullanılabilecek kestirim yöntemleri sırasıyla Bayes kestirimi (Bayes estimation-EAP), toplam puan EAP (Summed Score EAP-SSEAP) ve Bayes model kestirimi (Bayes model estimation-MAP) şeklindedir.

Çıktı Dosyaları, Uyum İyiliği İndeksleri ve Betimsel İstatistikler

IRTPRO 2.1; geleneksel toplam puana dayalı analizlerde –sss.htm ve MTK'ya dayalı analizlerde -ssc.htm ve -irt.htm uzantılı çıktı dosyaları oluşturmaktadır. –irt.htm uzantısı ile kaydedilmiş bir çıktı dosyasının görünümü şekil 5'teki gibidir. Bu çıktılarda sonuçlar tablolarla ve açıklamalarıyla birlikte yer almaktadır. Bu çıktı dosyalarında kestirim sonuçları, uyum indeksleri gibi bilgiler bulunmaktadır. Ayrıca -irt.irtplot uzantılı dosyada madde karakteristik eğrisi, madde bilgisi, madde karakteristik eğrisi ile madde bilgisinin kombinasyonu, toplam test bilgisi, test karakteristik eğrisi grafikleri yer almaktadır. Birey parametrelerine ilişkin analiz sonuçları ise –sco.txt uzantılı birey parametreleri çıktı dosyası çalışılan klasöre otomatik olarak kaydedilmektedir. Ayrıca ileri seçenekler penceresindeki "Save" seçimi açılarak madde parametre kestirimleri, parametre kestirimlerinin kovaryans matrisi, tek boyutlu modeller için bilgi değerleri, maddeler arası polikorik korelasyonlar, faktör yükleri ve temel çıktılar istenirse ASCII formatında (.txt) çalışılan klasöre kaydedilmektedir. Ayrıca program, bütün komutları otomatik olarak .irtpro dosyasında kaydetmektedir. Kaydedilen bu syntax dosyası üzerinden değişiklikler yapılabileceği gibi syntax dosyası kullanıcı tarafından da oluşturulabilir.

IRTPRO 2.1 yazılımı madde sayısına, cevap kategorilerine ve cevaplara dayalı olarak, -2log olabilirlik (-2log-likelihood), Akaike Bilgi Ölçütü (ABÖ), Bayesian Bilgi Ölçütü (BBÖ) gibi birkaç farklı uyum iyiliği indeksi ile Chen ve Thissen (1997) tarafından geliştirilen yerel bağımlılık istatistiği (local dependence-LD) ve Orlando ile Thissen (2000, 2003) tarafından önerilen madde uyumu istatistikleri ($S\text{-}X^2$) gibi betimsel istatistikler hesaplanmaktadır. Yazılım, yukarıdaki indekslere ek olarak Maydeu – Olivares ve Joe (2005) tarafından önerilen ki-kare testi ile ilişkili ve yaklaşık hataların ortalama karekökü değerini (RMSEA) veren M_2 istatistiğini de hesaplamaktadır. Ancak bu indeks, yalnızca BA-EM kullanıldığında analiz çıktılarında yer almaktadır. BA-EM algoritması ile tüm bu uyum indeksleri hesaplanırken, ADQ ve MH-RM algoritmaları kullanıldığında daha az uyum indeksi elde edilmektedir.



File Edit View Analysis Window Help

IRTPRO Version 2.1
Output generated by IRTPRO estimation engine Version 4.54 (32-bit)

Project:	Unidimensional_3PLM
Description:	Tek boyutlu verinin 3PLM'ye göre analizi
Date:	25 May 2015
Time:	12:41 PM

Table of Contents

3PL Model Item Parameter Estimates for Group 1, logit: $a\theta + c$ or $a(\theta - b)$
Summed-Score Based Item Diagnostic Tables and X^2s for Group 1
Group Parameter Estimates
Marginal fit (X^2) and Standardized LD X^2 Statistics for Group 1
Item Information Function Values for Group 1 at 15 Values of θ from -2.8 to 2.8
Likelihood-based Values and Goodness of Fit Statistics
Summary of the Data and Control Parameters

3PL Model Item Parameter Estimates for Group 1, logit: $a\theta + c$ or $a(\theta - b)$ (Back to TOC)

Item	Label	a		s.e.	c		s.e.	b	s.e.	logit g		s.e.	g	s.e.
1	C1	3	0.78	0.27	2	2.48	0.23	-3.18	0.94	1	-1.38	0.56	0.20	0.09
2	C2	6	0.82	0.24	5	0.65	0.21	-0.80	0.35	4	-1.42	0.56	0.19	0.09
3	C3	9	1.11	0.37	8	-0.21	0.28	0.19	0.23	7	-1.55	0.54	0.17	0.08
4	C4	12	0.75	0.22	11	0.97	0.19	-1.31	0.44	10	-1.41	0.56	0.20	0.09
5	C5	15	0.68	0.23	14	1.79	0.19	-2.62	0.81	13	-1.38	0.56	0.20	0.09

Şekil 5. IRTPRO 2.1 madde parametreleri çıktı dosyası örneği (-irt.htm)

Şekil 5'te yer alan madde parametreleri çıktısında görüldüğü gibi program ile her bir maddeye ilişkin "a", "c", "b" ve "g" parametreleri kestirilmektedir. IRTPRO 2.1'de "c" parametresi, "b" eşik parametresi ve "a" ayırt edicilik parametresinin etkileşimi (b = -c / a) ile elde edilen ve çok boyutlu MTK literatüründe "d" parametresi olarak geçen kesişim (intercept) parametresidir. Eğim-eşik formu çok boyutlu modellerde doğru genellemeler yapamadığı için IRTPRO 2.1'de hem tek boyutlu hem de çok boyutlu bütün modellerde eğim-kesişim formu kullanılmaktadır (a (Θ -b) = aΘ + c) (Cai vd., 2011). Ancak tek boyutlu modellerde "b" eşik parametresi de raporlanmaktadır. IRTPRO 2.1 çıktılarında şans parametresinin gösterimi de dikkat çekicidir. Buna göre literatürde "c" parametresi şeklinde gösterilen şans parametresi ise IRTPRO 2.1 çıktılarında "g" parametresi olarak yer almaktadır (Reckase, 2009; Ackerman, 1994; Way, Ansley ve Fosyth, 1988).

Kestirim Süresi

IRTPRO 2.1 ile yapılan analizlerde kestirilecek parametre sayısına, hesaplanacak istatistiklere, veri setinde yer alan madde/birey sayısına ve veri setinin yapısına bağlı olarak analiz süresi değişiklik göstermektedir. Intel® Core™ i7-4500CPU @1.80GHz 2.40 GHz 8.00GB RAM 64 bit işletim sistemi olan bir kişisel bilgisayarda tek boyutlu ve iki boyutlu yapılarda 20 madde içeren 2000 kişilik veri setleri üretilmiş ve bu veri setlerinin analizi için geçen süre Tablo 2'de verilmiştir.

Tablo 2. IRTPRO 2.1'de Örnek Kestirim Süreleri

Yapı	Model	Geçen Süre (saniye)
Tek boyutlu	1 PLM	0,87
	2 PLM	1,47
	3 PLM	2,85
	Aşamalı Tepki Modeli	4,33
	Genelleştirilmiş Kısmi Kredi M.	4,19
İki boyutlu	1 PLM	19
	2 PLM	56,41
	3 PLM	299,72
	Aşamalı Tepki Modeli	246,99
	Genelleştirilmiş Kısmi Kredi M.	551,16

Tablo 2'de yer alan kestirimlerde işlemci sayısı (Processors) iki olarak ayarlanmıştır ve birey parametreleri için EAP puanlama kullanılmıştır. Diğer seçimlerde varsayılan değerler kullanılmıştır. İki boyutlu kestirimlerde karmaşık yapılı veri setleri kullanılmış ve "Constraints" seçimi tıklanarak "a" parametrelerinde sınırlandırma yapılmıştır. Tablo 2'de yer alan kestirim süreleri incelendiğinde; model sabit kalmak üzere boyut sayısı arttığında birey ve madde parametrelerini kestirmek için gereken sürenin de önemli ölçüde arttığı görülmektedir.

SONUÇLAR ve TARTIŞMA

IRTPRO 2.1; hem tek boyutlu hem de çok boyutlu MTK'ya dayalı veri analizi yapan programların yerine getirebildiği fonksiyonların tek bir programla yapılabilmesini sağlamaktadır. Program ile iki-faktör ve madde takımı analizleri; hem iki kategorili hem de çok kategorili puanlanan maddelere sahip testlerin analizi de yapılabilmektedir. Yine MTK'ya dayalı olarak değişen madde fonksiyonu, çoklu grup ve çoklu tepki kategorilerinin analizlerinin yapılması mümkündür. Çok çeşitli uygulamaların tek bir programla yapılabilmesi sayesinde aynı veri setinde farklı uygulamalar yapan araştırmacıların kullandıkları veri setini farklı formatlara dönüştürmesine gerek kalmamaktadır. Kullanıcı dostu ara yüzü, araştırmacıların harf ve boşluk duyarlı kod yazmadan sekmelere tıklayarak analiz yapmasını sağlamaktadır. Bu açılardan ele alındığında IRTPRO 2.1, araştırmacıların yazılımı çalıştırmaya ilişkin olarak iş yükünü önemli ölçüde azaltmaktadır.

IRTPRO 2.1'in bu avantajlı yönlerinin yanında programın geliştirilmesine ihtiyaç duyulan yönleri de bulunmaktadır. Örneğin, program ile elde edilen parametre kestirimleri lojistik modele göre yapılmakta ve sonuçlar ölçekleme faktörünü (D = 1.7) içermemektedir. Bu sebeple ogive modelde kestirilen parametrelerle IRTPRO 2.1'in kestirdiği parametreleri karşılaştırabilmek için ayırt edicilik parametresi değerlerinin 1.7'ye bölünmesi gerekmektedir ve bu durum farklı program çıktılarıyla doğrudan karşılaştırma yapmanın önünde engel oluşturmaktadır. Program çıktılarına ogive model çıktısı eklenmesinin kullanıcıların için yararlı olacağı düşünülmektedir (Paek ve Han, 2012). Ölçek puanlarının hesaplanması için EAP ve MAP seçenekleri yer almakta; ML yöntemine ilişkin bir seçim bulunmamaktadır. Ayrıca program çıktılarında yer alan kesişim ve şans parametrelerinin simgeleri literatür ile tutarsızlık göstermektedir. İleri seçenekler seçimi ile analizde kullanılacak işlemci sayısı 1 - 8 arasında bir değerde belirlenebilmektedir. Ancak yüksek işlemci sayıları seçildiğinde program, bazı veri setlerinde hata vererek analizi sonlandıramamaktadır. Bu gibi durumlarda düşük işlemci sayısı ile çalışmak problemi ortadan kaldırsa da bu durum kestirim süresinin uzamasına neden olmaktadır. Programın kullanılan bilgisayarda yönetici modunda çalıştırılması gerektiği için bu durum iş bilgisayarlarında ya da ortak kullanımlı bilgisayarlarda problem oluşturabilmektedir. Son olarak yazılımın alandaki geçmişi fazla değildir ve kullanıcı sayısı sınırlıdır. Araştırmacıların yazılıma ilişkin kaynakları kullanıcı el kitabı ve teknik destekten oluşmaktadır ve farklı tartışmaların, soru cevapların ve uygulama örneklerinin yer aldığı forumlar gibi çok sayıda farklı kullanıcının oluşturduğu kaynakların geliştirilmesine ihtiyaç vardır (Han ve Paek, 2014). Sonuç olarak IRTPRO 2.1'in avantaj ve dezavantajları göz önünde bulundurulduğunda; yazılımın pek çok araştırma sorusunun tek ve çok boyutlu madde tepki kuramına dayalı olarak cevaplanmasında aracı olacağı ve kullanılan yazılıma dayalı olarak ortaya çıkan iş yükünü (veri çekme, kod yazma, çıktıların okunabilirliği) ciddi ölçüde azaltacağı düşünülmektedir. Araştırmacılara bu üstünlükler ve sınırlılıklar açısından yazılımın deneme sürümünü detaylı biçimde incelemeleri ve bundan sonra lisansı satın alıp almamaya karar vermeleri önerilmektedir. İleride yapılacak yazılım inceleme çalışmalarında MTK'ya dayalı analiz yapan yazılımlar ve IRTPRO 2.1 birlikte ele alınarak simülasyon verisi ve gerçek veri üzerinde yapılacak analizlerin çıktıları karşılaştırılabilir, kestirim süresi, uyum indeksleri vs. üzerinden üstünlükleri ve zayıflıkları tartışılabilir.

KAYNAKÇA

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement In Education, 7*(4), 255-278.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, *9*(1), 37-48.

Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). *TESTFACT 4.0* [Computer software and manual]. Lincolnwood, IL: Scientific Software International.

Cai, L. (2013). *flexMIRT Version 2: Flexible multilevel multidimensonal item analysis and test scoring* [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cai, L., du Toit, S. H. C., & Thissen, D. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling* [Computer software]. Lincolnwood, IL: Scientific Software International.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal o f Educational and Behavioral Statistics*, *22*(3), 265-289.

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

Fraser, C., & McDonald, R. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, *23*(2), 267-269.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory principles and applications.* Boston-USA: Kluwer-Nijhoff Publishing.

Han, K. T., & Paek, I. (2014). A review of commercial software packages for multidimensional IRT modeling. *Applied Psychological Measurement*, *38*(6), 486-498.

Linacre, J. M. (2011). *Winsteps Rasch measurement* [Computer program]. [Çevrim-içi: http://www.winsteps.com/index.htm], Erişim tarihi: 26 Nisan 2015.

Maydeu-Olivares, A., & Joe, H. (2005). Limited and full information estimation and testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association, 100*, 1009–1020.

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4: IRT item analysis and test scoring for rating-scale data* [Computer software]. Chicago, IL: Scientific Software International.

Orlando, M., & Thissen, D. (2000). Likelihood-based item fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50-64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of *S-X*2: An item fit index for use with dichotomous item response theory models. *Applied PsychologicalMeasurement*, 27(4), 289-298.

Paek, I., & Han, K.T. (2012). IRTPRO 2.1 for windows (Item response theory for patient-reported outcomes). *Applied Psychological Measurement*, *37*(3), 242-252.

Reckase, M. D. (2009). *Multidimensional item response theory (Statistics for social and behavioral sciences).* New York: Springer.

Rizopoulos, D. (2015). *Package "ltm", Latent Trait Models under IRT.* [Çevrim-içi: https://cran.r-project.org/web/packages/ltm/ltm.pdf], Erişim tarihi: 29 Nisan 2015.

Rusch, T., Mair, P., & Hatzinger, R. (2013). *Psychometrics with R: A review of CRAN packages for item response theory*. [Çevrim-içi: http://epub.wu.ac.at/4010/], Erişim tarihi: 10 Mayıs 2015.

Thissen, D. (2003). *MULTILOG 7: Multiple categorical item analysis and test scoring using item response theory* [Computer software]. Chicago, IL: Scientific Software International.

Way, W, D., Ansley, T. N., & Forsyth, R. A. (1988). Unidimensional IRT estimates the comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement, 12*(3), 239-252.

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

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items (Version 3)* [Computer software]. Chicago, IL: Scientific Software International.

EXTENDED ABSTRACT

Introduction

From the initial introduction to the end of 1970s, the Item Response Theory (IRT) used for unidimensional tests was extended to multi-dimensional tests from late 1970s and early 1980s and applied to tests measuring multiple abilities under Multidimensional Item Response Theory (MIRT) (Ansley & Forsyth, 1985; Reckase, 2009). Differently from unidimensional IRT, item and ability

parameters are estimated using MIRT equalities in multidimensional tests. The new theory therefore required to develop and use new software.

IRTPRO 2.1 (Item Response Theory for Patient-Reported Outcomes) program was developed by Li Cai, David Thissen and Stephen du Toit in 2011. IRTPRO 2.1 is a statistical software that uses item response theory for item calibration and test scoring. IRTPRO 2.1 uses IRT models that are specified below (Cai, Thissen & du Toit, 2011):

- Two-parameter logistic model (2PL) [becomes one-parameter logistic model with use of equal discrimination parameter (1PL)]
- Three-parameter logistic model (3PL)
- Graded Response Model
- Generalized Partial Scoring Model
- Nominal Response Model

Both unidimensional and multidimensional tests which belong to these models can be analyzed with IRTPRO 2.1. Also a test may also include different combinations of these models depending on the objective of the survey and the nature of measuring instrument.

Program Installation, Analyses, Outputs and Estimation Time

IRTPRO 2.1 can be used with Windows7, Vista and XP operating systems. Program can be accessed at http://www.ssicentral.com. The trial version of the software is also accessed at the same address with no cost. Users who use the paid version of the program are provided with technical support.

Data can be manually entered in IRTPRO 2.1 program via "File" tab as well as data can be extracted from files such as Excel (.xls) and SPSS (.sav) with fixed format (.fixed), comma separated (.csv) and space separated (.txt). Program saves the data that is extracted and analyzed with .ssig extension.

After data is entered, program executes the analysis selected from four basic types included under "Analysis" tab; (a) traditional summed-score statistics, (b) unidimensional IRT, (c) multidimensional IRT, (d) IRT scoring.

In IRTPRO 2.1 program, if maximum likelihood (ML) or prior distribution of item parameters is defined for item calibration, maximum a posteriori (MAP) method is used. With these methods, Bock-Aitkin Expectation-Maximization (BA-EM), Adaptive Quadrature (ADQEM) and Metropolis-Hastings Robbins-Monro (MH-RM) estimation methods are available to achieve the best performance based on the dimensionality and the different combinations of model structure.

Three types of IRT estimation methods are suggested with program to calculate scale scoring. So, the available estimation methods are Bayes estimation (EAP), Summed Score EAP (SSEAP) and Bayes model estimation (MAP), respectively.

IRTPRO 2.1 produces output files with –sss.htm extension in traditional summed score based analyses as well as output files with -ssc.htm and -irt.htm extension in analyses based on IRT. In addition, the file with -irt.irtplot extension contains item and test information curves. In analysis of individual parameters, individual parameters with –sco.txt extension will be automatically saved in the folder running the output files. Also, if "Save" option is opened in the next options window, the chosen outputs will be saved in the running folder as ASCII format (.txt).

Based on the number of items, response categories and responses, IRTPRO 2.1 software computes several different types of goodness of fit index such as -2log-likelihood, Akaike's Information

Criterion (AIC), Bayesian Information Criterion (BIC), and descriptive statistics such as local dependence-LD developed by Chen and Thissen (1997) and S-X2item compatibility statistics. In addition to indexes above, the software also computes the M2 statistics that provides chi-square test and root mean square error of approximation (RMSEA). However, this index is included analysis outputs only when BA-EM is used. While all these fit indexes are calculated with BA-EM algorithm, less fit index is achieved when ADQ and MH-RM algorithms are used.

In analyses performed with IRTPRO 2.1, the analysis time varies depending on the number of parameters to be estimated, statistics to be calculated, the number of items/individuals in the data set, and the structure of data. Simulated uni- and two-dimensional data sets with 20 items and 2000 examinees analyzed with a personal computer with Intel® Core™ i7-4500CPU @1.80GHz 2.40 GHz 8.00GB RAM 64 byte operating system. Table 1 presents the time for analyzing these data sets.

The number of processors was set at 2 for the estimations in Table 1, and EAP scoring was used for individual parameters. Default values were used for other options. Complex structured data sets were used for two-dimensional estimations and "a" parameters were constrained by clicking on the "Constraints" option. As can be seen in Table 2, when model remained constant, the time needed to estimate ability and item parameters was significantly increased as the number of dimensions was increased.

Table 1. Example Estimation Times in IRTPRO 2.1

Structure	Model	Estimation Time (second)
One-dimensional	1 PLM	0,87
	2 PLM	1,47
	3 PLM	2,85
	Graded Response Model	4,33
	Generalized Partial Credit Model	4,19
Two-dimensional	1 PLM	19
	2 PLM	56,41
	3 PLM	299,72
	Graded Response Model	246,99
	Generalized Partial Credit Model	551,16

Results and Discussion

IRTPRO 2.1 allows executing functions, which can be performed by several programs for data analysis based on IRT, with only one software. User-friendly interface allows researchers to perform analyses by clicking on tabs without typing case-sensitive and space-sensitive codes.

In addition to such advantageous of IRTPRO 2.1, it has aspects that require developing the program. For example, parameters obtained from the program are estimated based on logistic model and do not include scaling factor (D = 1.7). There is no option available for ML method to calculate scale scores. Furthermore, some parameter terms on the program outputs conflict with representations in the literature. When high number of processors are selected, program cannot terminate the analysis and gives errors in some data sets. Finally, the program can be run only on administrator mode.

As a result, when advantages and disadvantages of IRTPRO 2.1 considered; it can be expressed that the software appears to help answering many survey questions based on unidimensional and multidimensional item response theory and can reduce the problems due to the software which is being used. In addition, researchers are recommended to buy the software license after reviewing the trial version in detail.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2015; 6(2); 218-237

Bilgisayar Okuryazarlığı Testinin Bilgisayar Ortamında Bireye Uyarlanmış Test Olarak Geliştirilmesi*

Development of Computer Literacy Test as Computerized Adaptive Testing

Durmuş ÖZBAŞI ** Nükhet DEMİRTAŞLI ***

Öz
Bu araştırmanın amacı, Ankara Üniversitesi'nde tüm fakültelerde birinci sınıf öğrencilerine uygulanmakta olan Bilgi ve İletişim Teknolojileri dersi muafiyet sınavı testinin bilgisayar ortamında bireye uyarlanmış test (BOBUT) olarak uygulanabilirliğini araştırmaktır. Araştırma iki temel aşamada gerçekleştirilmiştir. İlk aşamada, hazırlanan maddeleri denemek ve simülatif BOBUT uygulamasında üzerinde çalışılacak verileri elde etmek üzere 1366 üniversite öğrencisiyle çalışılmıştır. İkinci aşamada ise, 142 üniversite birinci sınıf öğrencisiyle canlı (live) BOBUT ve kâğıt-kalem testi uygulaması gerçekleştirilmiştir. Araştırmada veri toplama aracı olarak Bilgisayar Okuryazarlık (BİLOKUR) testi kullanılmıştır. Araştırmada toplanan verilerin madde tepki kuramı'nın (MTK) 3 parametreli lojistik modeline uyum sağladığı tespit edilen 136 maddelik soru havuzu ile canlı BOBUT uygulaması yapılmıştır. Araştırmanın bulgularına göre, simülatif BOBUT uygulamasında kestirilen yeterlik kestirimlerine ilişkin en yüksek güvenirlik ölçüsü, sabit madde sayısına göre test sonlandırma koşulunda bulunmuştur. Ayrıca kullanılan madde sayısı bakımından, en az madde kullanımı, test sonlandırma koşulunun SH<0.50 olduğu durumda gerçekleşmiştir. En yüksek olabilirlik yöntemine (EYOY) göre yeterlik kestiriminin uygulandığı BOBUT testinde öğrencilerin yeterlik ölçüleri, kâğıt-kalem testinden elde edilenlere göre, özellikle uç değerlerde daha güvenilir ölçme sonuç vermiş, standart hata değeri açısından da BOBUT uygulamasıyla daha düşük hata kestirimleri elde edilmiştir. Canlı BOBUT uygulamasından elde edilen ortalama güvenirlik (test bilgi değeri), kâğıt-kalem testinden elde edilen güvenirlik değerinden daha yüksek bulunmuştur. Bu araştırmanın sonuçlarına göre, SH<0.30 test sonlandırma kuralı kullanıldığında EYOY; SH<0.50 ve sabit madde (30) test durdurma kuralı kullanıldığında ise, beklenen sonsal dağılıma (BSD) dayalı yeterlik kestirim değerlerinin daha güvenilir olduğu bulunmuştur. Ayrıca canlı BOBUT uygulamasında elde edilen test bilgi miktarı, kâğıt-kalem testinden elde edilen güvenirlikten anlamlı düzeyde yüksek bulunmuştur.

Anahtar Kelimeler: bilgisayar ortamında bireye uyarlanmış test, madde tepki kuramı, bilgisayar yeterlik sınavı.

Abstract
The purpose of this study is to investigate the applicability of the Computer Adaptive Testing (CAT) of the exemption exam of Information and Communication Technology in computer environment (as CAT) given to the first year students at every faculty of Ankara University each year. The research was carried out in two basic stages. In the first stage, the researchers studied with 1366 university students to obtain the data to study on with CAT application and and to test the prepared items. In the second stage, a paper and pencil test was given to 142 university first year students with live CAT. The test of computer literacy was used as an instrument of data collection. It was also tested in the study if the collected data met the hypothesis of Item

* Bu araştırma, Prof. Dr. Nükhet Demirtaşlı danışmanlığında Durmuş Özbaşı tarafından hazırlanan doktora tezinin bir bölümünden hazırlanmıştır..

**Arş. Gör. Dr. Durmuş ÖZBAŞI Çanakkale Onsekiz Mart Üniversitesi, Eğitim Fakültesi, Çanakkale-Türkiye, dozbasi@gmail.com
***Prof. Dr. Nükhet DEMİRTAŞLI Ankara Üniversitesi, Eğitim Bilimleri Fakültesi, Ankara-Türkiye, rnukhet@yahoo.com

Geliş Tarihi: 06.09.2015
Kabul Tarihi: 10.11.2015

Response Theory (IRT). With this regard, a live CAT application was carried out with 136-item pool which was found to comply with the three-parameter logistic model. According to the findings of the study, the highest reliability estimate found in simulative CAT application was found in test termination condition depending on the fixed item number (with 30 items). Besides, with regards to the number of used item, the least item use happened when test termination condition is Standard Error (SE)<0.50. In the CAT test, students' ability estimates in CAT in which proficiency estimate is done depending on Maximum Likelihood Estimation (MLE) has come up with more reliable results in extreme values compared to those obtained in paper-pencil test, and lower standard error estimates were obtained with the use of CAT application with regards to standard error value. The average reliability obtained from live CAT application was found to be higher that of paper-pencil test. According to the findings of this study, when the SE<0.30 test termination rule is applied, MLE was found to be more reliable, when SE<0.50 and fixed item (30) test termination rule was applied, the proficiency estimate value based on Expected A Posteriori Method (EAP) was found to be more reliable. Besides, the test information amount obtained in live CAT application was significantly higher than that of paper and pencil test.

Key Words: computerized adaptive testing (cat), item response theory (irt), computer literacy test, paper-pencil test

GİRİŞ

Günümüzde bilgisayar teknolojisinin gelişimine paralel olarak testler ve testlerin uygulanma yöntemleri de gelişmektedir. Özellikle, bilgisayar teknolojisinin gelişmesi ile birlikte son yirmi yıldır, eğitimde çeşitli amaçlarla (seçme, yerleştirme, teşhis, vb.) testler *bilgisayar ortamında bireye uyarlanan testler* (Computerized Adaptive Test) olarak kullanılmaktadır. Bilgisayar ortamında bireye uyarlanan test (BOBUT), psikometrik özellikleri daha önceden kestirilmiş bir madde havuzundaki maddeler arasından uygun seçimlerle, yanıtlayıcıların yeterlik (proficiency) düzeylerine uygun maddeler seçilerek, her birey için tüm maddelerin aynı olmadığı test olarak tanımlanabilir (Weiss, 2004). Bunu başarabilmek için de, tüm yanıtlayıcılara aynı güçlük dağılımında aynı maddeleri vermek yerine, aşağı-yukarı yöntemi olarak da bilinen; yanıtlayıcı doğru yanıt verirse daha zor, yanlış yanıt verirse daha kolay bir maddenin sorulmasına dayanan bir yöntem kullanılır (Rudner, 1998). Bu nedenle BOBUT uygulamalarında bireyin karşısına yeterlik düzeyine en yakın madde getirildiğinden uygulanan madde sayısında önemli miktarda bir azalma da sağlanmış olmaktadır. Böylece daha az madde ile daha güvenilir ölçme sonuçlarının elde edilmesi de mümkün olabilmektedir. (Çıkrıkçı-Demirtaşlı, 1999; Embretson ve Reise, 2000; Kalender, 2009; Mcglohen ve Chang, 2008).

BOBUT uygulamalarında önemli bir nokta, uygulama süresince yapılacak yeterlik kestirimlerinde hangi kuramın dikkate alındığı ve güvenilir sonuçlara nasıl ulaşıldığıdır. Birçok BOBUT uygulamasının psikometrik temeli Madde Tepki Kuramı (MTK)'na dayalı olarak düzenlenmektedir. Bazı BOBUT uygulamaları Klasik Test Kuramına (KTK) dayalı olarak yapılsa da (Frick, 1992; Rudner, 2002), BOBUT uygulamasında MTK'nın da sağladığı bazı avantajlardan (bireye özgü yeterlik ve hata kestirimi yapılabilmesi, madde ve test parametrelerinin değişmezlik özelliği gibi) yararlanarak, gerek testi alan yanıtlayıcıya gerekse testi uygulayana birtakım kolaylıklar sağlaması, BOBUT uygulamalarında KTK yerine MTK'nın tercih edilmesine neden olmuştur. MTK, bireylerin testle ölçülen yeterlik düzeyi ile testteki herhangi bir maddeyi yanıtlama davranışı arasında bir ilişki olduğunu belirten ve bu ilişkiyi olasılıklı bir modelle açıklayan bir kuramdır (Embretson ve Reise, 2000; Hambleton, Swaminathan ve Rogers, 1991; Wainer ve diğerleri, 1990). Ayrıca test geliştirme, madde analizi ve puanlama gibi bazı avantajlara sahip güçlü bir psikometrik paradigma olması da (Thompson ve Weiss, 2011), BOBUT uygulamalarında MTK'nın tercih edilmesini arttırmıştır.

MTK'ya dayalı BOBUT uygulamasında aşağıdaki sıra izlenir (Lord ve Stocking, 1988) :

- Belli bir yöntemle bireyin yeterlik parametresi bir kestirimi elde edilir.

- Madde parametreleri daha önceden kestirilmiş maddelerden oluşan bir havuzdan, bireyin yeterliğini en iyi kestirecek maddeler seçilir.
- Seçilen maddeler uygulanır ve bir sonraki madde seçilmeden önce bireyin son yeterlik düzeyi kestirilerek, havuzdan bireyin son yeterlik düzeyine en uygun madde seçilir.
- Seçilen sonlandırma kuralına göre test uygulaması sonlandırılır.

BOBUT uygulamalarının önemli bir boyutu da test uygulamasını başlatma, sürdürme ve sonlandırmada kullanılan ölçütlerdir. BOBUT uygulamalarındaki test başlatma ve test sürdürmede yeterlik kestirim yöntemlerini karşılaştıran çeşitli araştırmalar (Barrada, Olea, Ponsoda, Abad, 2010; Eggen, 2004; Keller, 2000; Kingsbury ve Zara, 1989; van Rijn, Eggen, Hemker ve Sanders, 2002) yapılmıştır. Bu araştırmalarda, test başlatma ve testi sürdürmede güvenilir kestirim veren yöntemin *En Yüksek Olabilirlik* -EYOY (Maximum Likelihood Estimation-MLE) Yöntemi olduğu bulunmuştur.

Yurt içinde ve yurt dışında yapılan araştırmaların çoğunda (Bulut ve Kan, 2012; Chae, Kang, Jeon ve Linarce, 2000; Frick, 1992; İşeri, 2002; Kalender, 2011; Kaptan, 1993; Keller, 2000; Kezer, 2013; Koşan-Aytuğ, 2013; McDonald, 2002; Miller, 2003; Mills ve Stocking, 1996; Mills ve Steffen, 2000; Öztuna, 2008; Rudner ve Guo, 2011; Scfhaer, Steffen, Golub-Smith, Mills ve Durso, 1995; Tian, Miao, Zhu ve Gong, 2007; Weiss ve Betz, 1973; Wainer, Dorans, Flaugher, Green, Mislevy ve Steinberg, 1990; Zitny, Halama, Jelinek, ve Kveton, 2012) simülatif BOBUT uygulamaları farklı uygulama stratejileri altında karşılaştırılarak, BOBUT uygulamasının kâğıt-kalem testine göre göstereceği psikometrik farklılıklar tespit edilmeye çalışılmıştır. BOBUT uygulaması ile kâğıt-kalem testi uygulamalarından kestirilen yeterlik /başarı düzeylerini karşılaştıran araştırmalar incelendiğinde, BOBUT uygulamalarının yeterlik kestirimlerinin güvenirliğini (yeterlik düzeyinde kestirilen standart hata değerinin düşük olması) artırdığı ve kullanılan madde sayısında önemli ölçüde tasarruf sağlandığı ulaşılan ortak bulgular arasındadır.

Türkiye'de merkezi olarak uygulanan geniş ölçekli sınav uygulamalarında (Yabancı Dil Sınavı, Temel Eğitime ve Orta Öğretime Geçiş Sınavı, Yükseköğretime Geçiş Sınavı, Lisans Yerleştirme Sınavı, vb.) çeşitli testler uygulanmakta ve bu uygulanan testlerin sonuçlarına dayanarak bireyler hakkında önemli kararlar alınmaktadır. Ancak bu testlerde, yanıtlayıcının kendi yeterlik düzeyine denk olan ve olmayan tüm soruları yanıtlaması beklenmektedir. Yanıtlayıcının tüm soruları yanıtlaması daha uzun zaman harcanmasına neden olmakta ayrıca yanıtlayıcının kendi yeterlik düzeyinin üstündeki çok zor ve altında kalan çok kolay birçok soruyu da yanıtlamasını gerektirmektedir. Bunların dışında bu tür sınavlarda test gizliliği konusunda da önemli sınırlılıklar bulunmaktadır. (Davis ve Dodd, 2005; French ve Thompson, 2003). Tüm bu sınırlılıkların önemli bir kısmı bu testler BOBUT uygulaması olarak geliştirildiğinde giderilebilir.

Türkiye'de BOBUT uygulamaları ile ilgili sınırlı sayıda görgül çalışma bulunmaktadır. İlk çalışmalar başarı testlerinin BOBUT olarak uygulanabilirliğini (Kaptan, 1993; Köklü, 1990; Yaşar, 1999) sonraki çalışmalar ortaöğretim ve yükseköğretim okullarına öğrenci seçmede kullanılan testlerle, farklı konularda yeterlik belirleme amaçlı testlerin BOBUT olarak uygulamasının geleneksel (kâğıt kalem testi) uygulamayla karşılaştırmasını konu edinmiştir (Aytuğ-Koşan, 2013; Bulut ve Kan, 2012; İşeri, 2002; Kalender, 2011; Kezer, 2013). Bir kısım çalışma ise, tıp alanında hasta beyanına dayalı teşhis araçlarının BOBUT olarak geliştirilmesiyle ilgilidir (Öztuna, 2008).

Bu çalışmada, üniversitelerde zorunlu temel derslerden biri olan "Temel bilgisayar" dersinden muaf tutulacak öğrencilere karar vermede kullanılan "Bilgisayar Okuryazarlığı (BİLOKUR) Muafiyet" testinin BOBUT olarak uygulanabilirliği araştırılmıştır. Her yıl uygulanmakta olan Bilgisayar muafiyet sınavı ile çok sayıda öğrencinin bu dersi alıp almayacağına ilişkin karar verilmektedir. Bu sınavlara yönelik olarak, her yıl birçok soru hazırlanmaktadır. Bu sınavlarda uygulanan testlerde soruların tamamını, yeterlik düzeyi ne olursa olsun tüm öğrenciler yanıtlamak durumundadır. Bir başka ifadeyle, öğrenciler kendi yeterliklerinin üstünde ve altında kalan gereğinden fazla sayıda soruyu da yanıtlamak zorunda kalmaktadırlar. Bu durum testlerin kullanışlılığının zayıflamasına ve test maliyetinin artmasına neden olmaktadır. BİLOKUR testinin bu sınırlıkları gidermek üzere,

kâğıt-kalem testi olarak olarak uygulanan BİLOKUR testinin BOBUT olarak uygulananabilirliğini sınamak bu araştırmanın problemini oluşturmaktadır.

Araştırmanın Amacı

Bu çalışmanın amacı, Bilgisayar Okuryazarlığı Testi (BİLOKUR)'nin BOBUT olarak uygulanabilirliğini farklı koşullar için araştırmaktır. Bu amaçla testin BOBUT uygulaması ile kâğıt-kalem uygulamasından elde edilen psikometrik nitelikleri karşılaştırılarak, en uygun BOBUT uygulaması stratejileri (testi sürdürme ve sonlandırma stratejileri) saptanmaya çalışılmıştır.

Bu genel amaç doğrultusunda BİLOKUR Testinin BOBUT olarak uygulanabilirliği aşağıdaki iki soru kapsamında sınanmıştır;

1) Simülatif BOBUT uygulamasında farklı yeterlik kestirim yöntemleri (En yüksek olabilirlik Yöntemi - EYOY ve Beklenen sonsal dağılım- BSD) ile farklı test sonlandırma kuralları (sabit test uzunluğu (k=30) ve ölçmenin standart hatası (SH<0.30 ve SH<0.50) kapsamında elde edilen birey yeterlik parametreleri ve test bilgi değeri (güvenirlik) arasında anlamlı bir fark var mıdır?
2) Canlı BOBUT uygulaması için sonlandırma kuralı olarak standart hata değeri ve sabit madde koşulu uygulandığında, birey yeterlik parametreleri, test güvenirliği ve madde sayıları kâğıt-kalem uygulamasından elde edilen değerlerden anlamlı bir farklılık göstermekte midir?

YÖNTEM

"Araştırmada, BİLOKUR testinin BOBUT olarak uygulanabilirliğini test etmek üzere, farklı yeterlik kestirim yöntemleri ve farklı sonlandırma kuralları karşılaştırılmıştır. Çalışmada, post-hoc simülasyon yöntemi ile farklı yeterlik kestirim yöntemleri EYOY, BSD ve test sonlandırma kurallarına (SH <0.30 ve SH<0.50), dayalı koşullara bağlı olarak yeterlik kestirimleri yapılmıştır. Bu amaçla simülatif BOBUT programı olan SimulCAT (Han, 2010) yazılımından yararlanılmıştır. Bu yönüyle araştırma mevcut kuramsal bilginin gelişmesine ve genişlemesine katkıda bulunan, aynı zamanda uygulamaya da katkı getiren temel araştırma modelindedir. Temel araştırmalar Karasar (2011)'e göre, kuramlara dayalı olarak teorilerin gelişmesine katkıda bulunmak, varsayımlar geliştirerek ve bunları test ederek, sonuçlarını bilimsel olarak yorumlayarak bilgilerin genişlemesini ve gelişmesini amaçlayan araştırmalardır. Daha sonra ise, canlı (live) BOBUT ve kağıt kalem testinden elde edilen yeterliklerin psikometrik özellikleri bir grup öğrencinin katıldığı uygulamada karşılaştırılmıştır. Bu amaçla, Kalender (2011) tarafından geliştirilmiş olan BOBUT uygulama yazılımı kullanılmıştır. Bu yönüyle de araştırma var olan bir uygulamayı geliştirme amacını taşıyan uygulamalı araştırma niteliğindedir.

Çalışma Grubu

Araştırmanın amacına yönelik olarak BİLOKUR testine ait veriler, Ankara Üniversitesi'nin çeşitli fakültelerinde 2012-2013 ve 2013-2014 eğitim öğrenim yılında birinci sınıfta öğrenim görmekte olan üniversite öğrencilerine uygulanarak elde edilmiştir. Çalışma grubunu, bu grupta yer alan farklı öğrencilerin oluşturmuş ve çalışmanın iki aşamasında yer almışlardır. İlk aşamada, soru bankası oluşturmak için, 2012-2013 eğitim öğretim yılında aynı bilgisayar dersinin okutulduğu birinci sınıf öğrencilerine uygulanmıştır. Uygulama; Eğitim, Hukuk, Mühendislik, Eczacılık, Dil-Tarih Coğrafya, Tıp fakületesi birinci sınıfta okuyan toplam 1452 üniversite öğrencisini kapsamıştır. Geliştirilen BİLOKUR testi maddeleri, 6 farklı madde grubuna bölünerek araştırma kapsamındaki öğrencilere uygulanmıştır. Bunun nedeni, aynı öğrencilerin bir oturumda 191 soruya yorgunluk, sıkılma faktörleri yüzünden güvenilir bir şekilde yanıt verememe olasılığıdır. Ancak ilk aşamada, öğrencilerin bazıları teste hiç yanıt vermediği veya birkaç maddeye yanıt verdiği için çalışma grubundan çıkarılmış ve sonuç olarak 1366 üniversite birinci sınıf öğrencisiyle pilot uygulamaya

ilişkin veriler toplanmıştır. İkinci aşamada, bu veriler simülatif BOBUT uygulamasında kullanılmıştır. Uygulamaya katılan öğrencilerin dağılımı Tablo 1'de verilmiştir.

Tablo 1. Soru Bankasının Pilot Uygulamasına Katılan Öğrencilerin Cinsiyetlerine Göre Dağılımı

Cinsiyet	F	%
Erkek	679	49
Kız	687	51
Toplam	1366	100

Araştırmanın son aşamasında, parametreleri kestirilen ve MTK'ya uygunluğu tespit edilen BİLOKUR testi maddeleri canlı BOBUT ve kağıt-kalem testi olarak, Ankara Üniversitesi Eğitim Bilimleri Fakültesi'nde farklı bölümlerde (sınıf öğretmenliği, sosyal bilgiler öğretmenliği, rehberlik ve psikolojik danışmanlık, bilgisayar eğitimi ve teknolojileri öğretmenliği ve okul öncesi öğretmenliği) öğrenim görmekte olan 2013-2014 eğitim öğretim yılında 142 üniversite birinci sınıf öğrencisine uygulanmıştır. Bu uygulamaya katılan öğrencilerin cinsiyetlerine ilişkin dağılım Tablo 2'de verilmiştir.

Tablo 2. Kağıt-Kalem Testi ile Canlı (Live) BOBUT Uygulamasına Katılan Öğrencilerin Cinsiyetlerine Göre Dağılımı

Cinsiyet	f	%
Erkek	32	23
Kadın	110	77
Toplam	142	100

Veri Toplama Araçları

Araştırmada veri toplama aracı olarak, BİLOKUR testi kullanılmıştır. Bu test aynı öğrencilere hem kağıt-kalem ortamında hem de BOBUT olarak uygulanmıştır.

Bilgisayar Okuryazarlığı Testi

Çalışmada kullanılan ölçme aracı, bilgisayar okuryazarlığıyla ilgili temel bilgi ve becerileri ölçmeyi amaçlayan BİLOKUR testidir. Bu test, her yıl üniversiteye yeni başlayan tüm öğrencilere Ankara Üniversitesi Enformatik Bölümü tarafından uygulanarak, bu dersten muaf olup olmayacak öğrenciler saptanır. Bu test, Avrupa Bilgisayar Yetkinlik Belgesi (European Computer Driving Licence-ECDL) programında tanımlanan bilgi ve becerileri ölçmeyi amaçlar. Bilgisayar okuryazarlığı testinin kapsamı "Bilgi ve İletişim Teknolojisi Kavramları" modülündeki yeterliklerle ile sınırlıdır. Bu modül; bir kişisel bilgisayarın fiziksel yapısı, veri saklama, bellek, toplumda çok kullanılan yazılım uygulamaları ve bilgisayar ağlarının kullanımı ile ilgili temel kavramların bilinmesini içermektedir. Aday ayrıca, bilgi teknolojilerinin günlük kullanımı ve bilgisayarların insan sağlığına etkileri ile bilgisayarlarla ilgili bazı güvenlik ve hukuk konuları hakkında bilgi sahibi olmalıdır (http://enformatik.ankara.edu.tr/?page_id=174).

Bilgi ve İletişim Teknolojisi Kavramları modülü temel alınarak hazırlanan BİLOKUR testi kapsamında adaylarda yoklanan beceriler sıralanmıştır (ECDL, 2007):

- Donanımı ve bilgisayar performansını nasıl etkilediğini bilir,

- Yazılımın ve uygulama yazılımlarının ne olduğunu anlar ve örneklendirir,
- Bilgisayarlar arasındaki ağın nasıl olduğunu anlar ve internete bağlanmanın değişik yollarını bilir,
- Bilgi ve İletişim Teknolojisi Kavramlarını anlar günlük yaşamımızda kullanımını örneklendirir,
- Bilgisayar kullanımında güvenlik ve sağlık faktörlerini anlar,
- Telif hakları, veri koruma ve yasal kullanım hakkındaki temel bilgileri bilir.

BİLOKUR testi, teknoloji okuryazarlığı alanındaki taksonomilere (Tomei, 2005) uygun olarak ve bilgisayar okuryazarlığı ile ilgili temel bilgi ve becerileri başat faktör olarak ölçmek üzere hazırlanmıştır. Bu süreçte hazırlanan maddeler havuzu belirtke tablosunda belirtilen becerilere dayalı olarak, iki ölçme ve değerlendirme uzmanı ile bir bilgisayar eğitimi ve teknolojileri alanında uzman olan üç kişilik bir uzman grup tarafından i) maddenin beceriyi temsil durumu ii) maddenin çoktan seçmeli madde tekniğine uygun yazılma durumu iii) bilimsel doğruluk iv) dil ve anlatım bakımından uygunluk ölçütleri bakımından; "uygun", "uygun değil" ve "düzeltilmeli" kategorileri kapsamında incelenmiştir. Bu ölçütler bakımından onaylanan ve düzeltme önerisi verilen maddeler (191 madde) gözden geçirilerek testin deneme uygulamasına alınmıştır.

BİLOKUR testini oluşturan maddeler, 2012-2013 eğitim-öğretim yılında toplam 1366 öğrenciye yaklaşık iki ay gibi bir sürede uygulamıştır. Madde havuzunun geniş olmasından dolayı, maddeler 6 farklı soru grubu/test olarak uygulanmıştır. Uygulama sonrası madde istatistikleri KTK'ya dayalı olarak hesaplanmış ve özet sonuçlar Tablo 3'te verilmiştir.

Tablo 3. BİLOKUR Testi'nde Kullanılan Maddelere İlişkin Betimsel İstatistikler

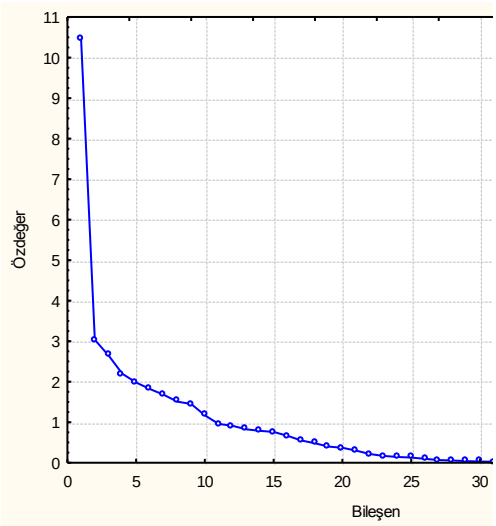
	Madde grupları (k: madde sayısı)					
Test İstatistikleri	Grup1 (k= 50)	Grup2 (k=50)	Grup3 (k=46)	Grup4 (k=15)	Grup5 (k=20)	Grup6 (k=10)
Ortalama	29.62	21.32	23.03	7.70	17.43	6.26
Ortanca	30	20	21	8	18	6
Ortalama $\bar{p}$	0.59	0.43	0.50	0.51	0.87	0.62
Ortalama ayırt edicilik	0.57	0.41	0.38	0.33	0.69	0.57
Tepe Değeri	31	16	21	8	20	7
En Küçük	11	6	7	0	0	0
En Büyük	44	43	45	14	20	10
Standart Sapma	6.21	6.77	7.86	2.99	3.25	1.82
Varyans	38.52	45.85	61.82	8.95	10.59	3.33
Basıklık	0.43	0.30	-0.12	-0.34	8.49	0.08
Çarpıklık	-0.62	0.68	0.47	-0.34	-0.92	-0.58
Çarpıklığın Standart Hatası	0.14	0.16	0.17	0.15	0.07	0.08

Tablo 3 incelendiğinde, maddelerin çoğunun ortalama güçlük düzeyinde olduğu, az sayıda maddenin kolay madde olduğu saptanmıştır. Maddelerin ortalama madde ayırt edicilik değerleri incelendiğinde ise, en küçük 0.33 en yüksek 0.69 arasında değiştiği bulunmuştur.

BOBUT uygulamasında MTK'ya göre ölçeklenmiş maddelerin kullanılması, MTK'nın birey ve madde parametrelerinde değişmez (invariant) kestirimler vermesini sağlar. Ancak bu avantajların elde edilmesi, büyük ölçüde kullanılan verilerin model ile uyumlu olmasına (Fan, 1998; Hambleton ve Swaminathan, 1989) ve ölçmeye konu olan psikolojik özelliğin, eldeki veriler tarafından ölçülebildiğinin kanıtlarının ortaya konmasına (Stark ve Chernyshenko, 2001) bağlıdır. Bu amaçla, soru bankasında kalan maddeler MTK'ya dayalı ön analizlere tabi tutulmuştur. Bu analizlerde MTK'nın tek boyutluluk, yerel bağımsızlık varsayımlarının karşılanma durumu, hız testi olup olmadığının kontrolü ile madde ve birey yeterlik parametrelerinin değişmezliği sınanmıştır.

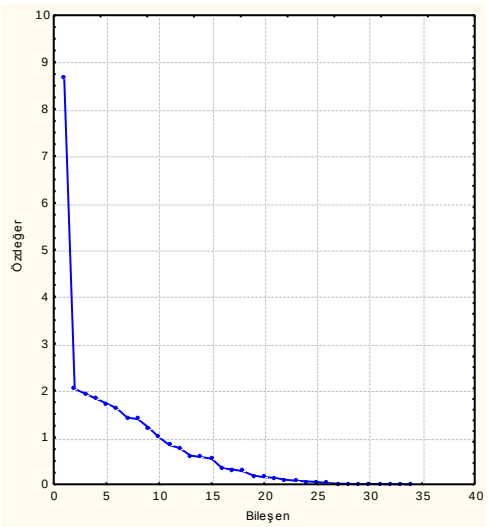
MTK'nın en önemli varsayımlarından biri olan tek boyutluluk, testi oluşturan maddelerin başat bir özelliği ölçmesi ve madde faktör yüklerinin bu boyut altında yük değerleri (faktör yükü> 0.30) vermesi olarak tanımlanmaktadır (Embretson ve Reise, 2000; Hambleton ve Swaminathan, 1985). Tek boyutluluğun sınanmasında açımlayıcı faktör analizi (AFA) kullanılmıştır. AFA değişken azaltma ve ortaya çıkan faktörleri isimlendirmenin dışında, davranışın anlaşılmasına olanak veren kuramsal yapı (gözlenemeyen gizil/örtük değişkenler) ile benzer olup olmadığını ortaya koyar (Kline, 2000). AFA'nın 1-0 şeklinde kategorik olarak puanlanan verilerde uygulanabilmesi için tetrakorik korelasyon matrisinin oluşturulması gerekmektedir (Baykul, 2010; Sheskin, 2004). Bu çalışmada her madde grubu ayrı yanıtlayıcı gruplarında uygulandığından, her grupta uygulanan madde grubunda maddeler-arası "tetrakorik korelasyon" matrisine dayalı temel eksenler analizine göre madde gruplarının (testin) kendi içerisinde tek boyutlu olup olmadığı incelenmiştir. Buna göre, tek boyutluluk varsayımının sınanmasında boyut sayısına karar verilirken kullanılabilecek en pratik yol olarak tetrakorik maddeler-arası korelasyon matrisinin örtük kökleri (λ_r) dikkate alınmıştır (Lord, 1980).

Buna göre, birinci özdeğer ikinci özdeğere göre büyükse ve ikinci öz değer kendinden sonra gelen özdeğerle arasındaki fark büyük değilse, maddelerin tek boyutlu bir yapıyı temsil ettiğinden söz edilebilir. Buna ilişkin madde gruplarına ait yamaç birinti grafikleri ve açıklanan varyans değerleri aşağıda verilmiştir.



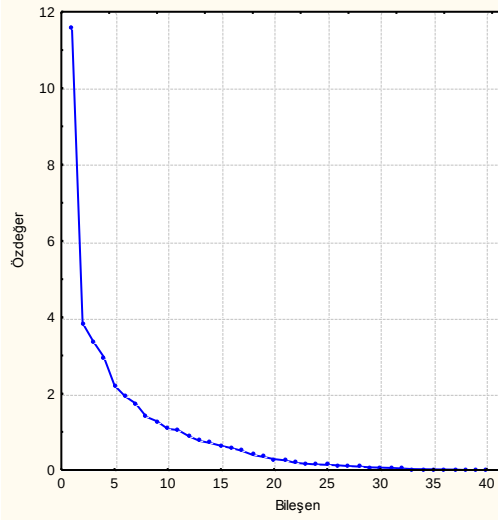
Şekil 1. Grup 1 maddeleri Yamaç Birikinti grafiği

(1. Boyutun açıkladığı toplam varyans: %28.97)

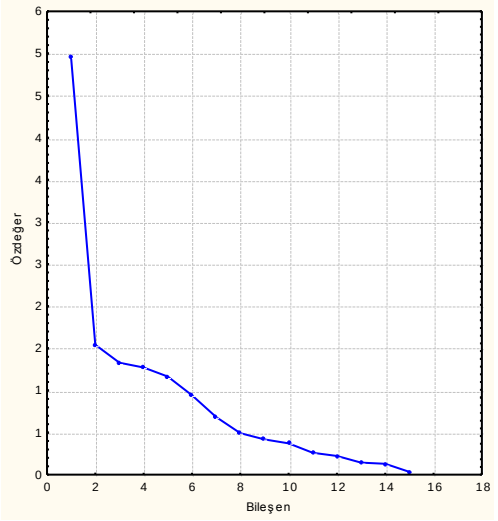


Şekil 2. Grup 2 maddeleri yamaç birikinti grafiği

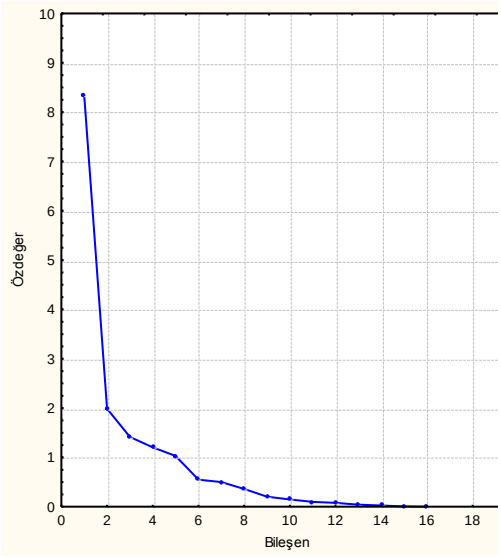
(1.Boyutun açıkladığı toplam varyans: %25.53)



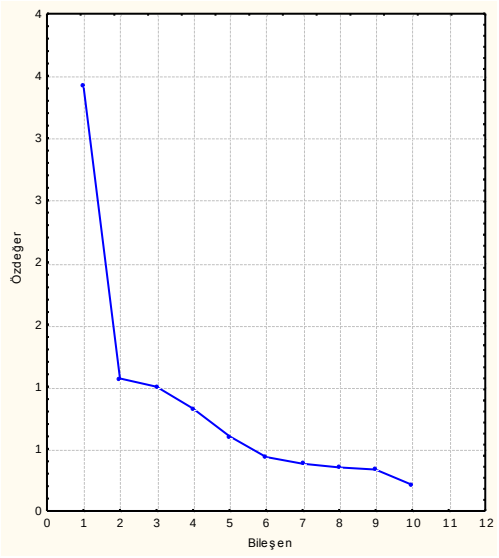
Şekil 3**.** Grup 3 maddeleri yamaç birikinti grafiği (1.Boyutun açıkladığı toplam varyans:% 37.78)



Şekil 4. Grup 4 maddeleri yamaç birikinti grafiği (1.Boyutun açıkladığı toplam varyans: 33.10)



Şekil 5. Grup 5 maddeleri yamaç birikinti grafiği (1.Boyutun açıkladığı toplam varyans: %52.16)



Şekil 6. Grup 6 maddeleri yamaç birikinti grafiği (1.Boyutun açıkladığı toplam varyans: %35.50)

MTK modellerinin avantajlarından yararlanmak için, maddelerin yerel bağımsızlık göstermesi de beklenir. Yanıtlayıcının yeterliği sabit olmak üzere test maddelerinden birine verdiği yanıtın, testin diğer maddelerine verdiği yanıtı etkilememesi olarak tanımlanan yerel bağımsızlık varsayımında; belli bir yeterlik düzeyindeki kişilerin maddelere vermiş oldukları yanıtlar arasındaki korelasyonun sıfır olduğu kabul edilir (Lord, 1980). Alan yazında birçok araştırmacı yerel bağımsızlık varsayımını, tek boyutluluk ile ilişkilendirerek açıklamıştır (Emretson ve Reise, 2000; Hambleton, Swaminathan ve Rogers, 1991). Bu kapsamda, araştırmada kullanılan tüm maddelerin uygulandığı gruplarda başat tek boyutluluğu sağladığı ölçüde yerel bağımsızlık varsayımını da karşıladığı sonucuna ulaşılmıştır. Bu bakımdan, bu araştırmada kullanılan BİLOKUR testi maddeleri, yerel bağımsızlık özelliğini göstermektedir.

MTK'da test edilmesi gereken önemli varsayımlardan biri de testin hız testi olup olmadığıdır. Çünkü MTK'daki modeller, hız sınırlaması olmadan uygulanan testlerdeki modellere uygundur.

Yanıtlayıcılar zamanı yetişemediği için maddeye yanlış yanıt vermemiş veya maddeyi boş bırakmamış olmalıdır (Hambleton ve Swaminathan, 1989). Bu bağlamda, araştırmada kullanılan testin hız testi gibi çalışıp çalışmadığı MTK modellerine uyum sağlaması açısından oldukça önemlidir. Hambleton, Swaminathan ve Rogers (1991), hız testi kontrolü için önerdikleri ölçüt, maddelerin hemen hemen tümünü tamamlayan öğrenci sayısının tüm öğrenci sayısına yakın olmasıdır. Bu yöndeki incelemeler sonucunda, her madde grubunda "son maddeye erişen öğrenci" oranının %95 ile %100 arasında değiştiği saptanmıştır. Bu sonuç, her grupta alınan testlerin bir hız tesi gibi çalışmadığını kanıtlar niteliktedir.

MTK ile ölçme aracından elde edilen verilerin, yorumlanması ve raporlanmasında sağladığı avantajların elde edilmesi, büyük ölçüde kullanılan veriler (maddeler) ve MTK modeli arasındaki uyuma bağlıdır (Fan, 1998; Hambleton ve Swaminathan, 1989). MTK'da model-veri uyumunun olup olmadığının kontrol edilmesi için öncelikle MTK'nın temel varsayımlarının karşılanıp karşılanmadığının test edilmesi gerekir. Bu bağlamda, araştırmada MTK varsayımlarının karşılanmasına ilişkin yukarıda açıklanan analizlerden elde edilen bulgular, verilerin model-veri uyumu için kanıt niteliğindedir. Ayrıca -2likelihood değerleri, modele ilişkin test düzeyinde gözlenen ve beklenen madde yanıt örüntülerine bağlı doğru ve yanlış yanıtlanma olasılıkları farklı yeterlik aralıklarında bulunan bireyler için karşılaştırılır. (Reise, 1990; Zimowski, Muraki, Mislevy ve Bock 2003). Araştırma kapsamında BİLOKUR testine ilişkin -2likelihood değerleri, modele ilişkin doğru karar verebilmek amacıyla her madde grubu için incelenmiş ve Tablo 4'te verilmiştir.

Tablo 4. Maddelerin model veri uyumuna ilişkin MTK modellerinden hesaplanan -2Loglikelihood değerleri

Madde Grubu	1PLM	2PLM	3PLM
Grup 1	10052.90	9974.30	9968.58
Grup 2	4219.67	4172.18	3837.42
Grup 3	4379.50	4256.12	4050.86
Grup 4	2241.53	2263.78	2233.88
Grup 5	7486.56	7238.04	5553.42
Grup 6	4586.15	4566.18	4537.55

Tablo 4'te verilen -2Loglikelihood değerleri incelendiğinde, 3PLM ile hesaplanan -2Loglikelihood değerleri diğer modellerden daha düşük olması, bu modele verilerin daha iyi uyum verdiğini göstermiştir. Ayrıca kullanılan testin maddelerinin çoktan seçmeli olması, şans başarısı faktörünü gözeten bir model (Crocker ve Algina, 1986; Hambleton, Swaminathan ve Rogers, 1991) olduğu için kestirimlerde 3PL model tercih edilmiştir.

3PL'ye göre kestirimleri yapılan bu maddelerin madde ayırıcılık gücü indeksi, a, ortalama değeri 1.132; ortalama madde güçlük indeksi, b, 0,542 ve ortalama şans parametresi c, değeri 0.363 olarak hesaplanmıştır. Model-veri uyumu testi sonucunda 15 madde model veri uyumu göstermediği için soru bankasından çıkarılmıştır.

BİLOKUR testi soru bankasından tek boyutluluk, model veri uyumu göstermeme gibi nedenlerden dolayı toplamda 55 madde çıkarılmış ve nihai olarak soru bankasında, bilgisayar okuryazarlığı becerilerini ölçebilecek 136 madde kalmıştır.

Bu sınamanın ardından, madde ve birey parametrelerinin değişmezliği incelenmiştir. MTK modeline göre kestirilen madde parametrelerinin değişmezliği, madde parametrelerinin testin uygulandığı gruptan bağımsız olarak kestirilebilmesidir (Hambleton, Swaminathan ve Rogers,1991). Bunu sınamak için, her madde grubunun uygulandığı öğrenci grubu tesadüfi olarak iki gruba ayrılarak her alt örneklemden madde parametreleri kestirilmiş ve kestirilen madde parametreleri arasındaki korelasyon "Pearson Momentler Çarpımı Korelasyon Katsayısı" ile incelenmiştir. Ayrıca, parametre değişmezliği incelemek için oluşturulan farklı öğrenci gruplarından kestirilen madde parametrelerinin büyüklük sırasının benzer olup olmadığının incelemek amacıyla da "Spearman Brown Sıra Farklı Korelasyonu" yapılarak incelenmiş ve sonuçlar Tablo 5'te verilmiştir.

Tablo 5. Madde Parametrelerinin Değişmezliğine ilişkin korelasyon değerleri

Madde grubu	Yeterlik Grubu		a	b	c
Grup 1 (k=33)	Grup1-Grup2	Pearson	0.36**	0.91**	0.68**
		Spearman	0.34**	0.90**	0.52**
Grup 2 (k=31)	Grup1-Grup2	Pearson	0.05	0.64**	0.68**
		Spearman	0.06	0.66**	0.66**
Grup 3 (k=38)	Grup1-Grup2	Pearson	-0.12	0.69**	0.76**
		Spearman	-0.07	0.62**	0.69**
Grup 4 (k=12)	Grup1-Grup2	Pearson	0.48**	0.75**	0.85**
		Spearman	0.59*	0.79**	0.68**
Grup 5 (k=14)	Grup1-Grup2	Pearson	0.62**	0.97**	0.76**
		Spearman	0.56**	0.94**	0.70**
Grup 6 (k=8)	Grup1-Grup2	Pearson	-0.05	0.63*	-0.23
		Spearman	-0.06	0.91**	-0.32

* p<0.05; **p<0.01; k: madde sayısı

Tablo 5'te görüldüğü gibi, her madde grubunun uygulandığı gruplarda, tesadüfi olarak iki gruba ayrılan bireylerden kestirilen a, b ve c parametreleri arasındaki korelasyonlarda, genellikle b ve c parametrelerinde orta ve iyi düzeyde manidar ilişki bulunmuştur. Ancak madde ayırt edicilik parametrelerine ilişkin korelasyonlar madde güçlük ve şans parametresine göre daha düşük düzeyde çıkmıştır. Bu durum, alan yazında yapılan çalışmalarla (Çıkrıkçı-Demirtaşlı, 2002; Fan, 1998; Kalender, 2011; Kelecioğlu, 2001; Kezer, 2013) paralellik göstermektedir. Stocking (1990) madde parametrelerinin kestirim yapıldığı grubun homojen olmasının, madde parametrelerinin değişmezliğini azalttığını belirtmiştir.

MTK'nın değişmezlikle ilgili diğer sayıltılarından birisi de yeterlik parametrelerinin değişmezliğidir. Araştırmada yanıtlayıcılara ait yeterlik parametrelerinin madde örnekleminden bağımsız kestirilip kestirilmediğini saptamak için maddeler tesadüfi olarak iki gruba ayrılmıştır. Buna gore, BİLOKUR testinden model-veri uyumunu sağlayan maddeler tesadüfi olarak iki gruba ayrılmış, iki madde seti oluşturulmuştur. Bu madde setlerinden kestirilen birey yeterlik parametreleri arasındaki korelasyonlar hesaplanmış ve sonuçlar Tablo 6'da özetlenmiştir.

Tablo 6**.** Farklı Madde Setlerine Ait Yeterlik Ölçüleri Arasındaki Korelasyonlar

Madde grubu	Madde Seti	Madde Seti 1	Madde Seti 2	Testin Tümü
Grup 1	Madde Seti 1	1.00		
	Madde Seti 2	0.69*	1.00	
	Testin Tümü	0.89*	0.93*	1.00
Grup 2	Madde Seti 1	1.00		
	Madde Seti 2	0.75*	1.00	
	Testin Tümü	0.92*	0.92*	1.00
Grup 3	Madde Seti 1	1.00		
	Madde Seti 2	0.78*	1.00	
	Testin Tümü	0.93*	0.94*	1.00
Grup 4	Madde Seti 1	1.00		
	Madde Seti 2	0.52*	1.00	
	Testin Tümü	0.88*	0.84*	1.00

Tablo 6 Devamı

Madde grubu	Madde Seti	Madde Seti 1	Madde Seti 2	Testin Tümü
Grup 5	Madde Seti 1	1.00		
	Madde Seti 2	0.54*	1.00	
	Testin Tümü	0.95*	0.77*	1.00
Grup 6	Madde Seti 1	1.00		
	Madde Seti 2	0.34*	1.00	
	Testin Tümü	0.86*	0.64*	1.00

**p<0.01

Her madde grubunda, alt madde setlerinden kestirilen yeterlik parametreleri arasındaki ilişkilerin pozitif yönde orta ve yüksek düzeyde olmak üzere anlamlı ilişkiler gösterdiği bulunmuştur. Buna göre BİLOKUR testini oluşturan soru bankasında kestirilen yeterlik parametrelerinin büyüklerine ilişkin sıraların madde alt setlerine göre tutarlı olduğu, değişmediği sonucuna ulaşılmıştır.

Verilerin Analizi

Yukarıda özetlenen ön analizlerden sonra, araştırma soruları çerçevesinde simülatif BOBUT uygulaması ile farklı yeterlik kestirim yöntemleri ve test sonlandırma koşullarında kestirimler yapılmıştır. Ardından simülatif BOBUT uygulamasından çıkan sonuçlara göre, kâğıt-kalem testinden elde edilen yeterlik kestirimlerine ilişkin karşılaştırmalar yapmak üzere canlı BOBUT uygulaması yapılmıştır. Farklı test sonlandırma (sabit madde ve sabit hata) ve farklı yeterlik kestirimi koşullarında tekrarlanan BOBUT uygulamalarından kestirilen yeterlik ölçülerinin farklılığı tek yönlü varyans analizi kullanılarak incelenmiştir. Varyans analizinde, varyansların homojen olmaması durumunda gruplar arası farklar için Dunnet C testi kullanılmıştır. Ayrıca simülatif ve canlı BOBUT uygulamalarından elde edilen test bilgi miktarları arasında anlamlı bir fark olup olmadığını incelemek için bağımsız örneklemler için t-testi kullanılmıştır.

Simülatif BOBUT Uygulaması

Simülatif BOBUT uygulaması, SimulCAT (Han, 2010) programının yardımıyla yapılmıştır. Bu uygulama için gereken yeterlik parametreleri, madde havuzunun kağıt-kalem testi olarak uygulanmasından elde edilen veriler 3PLM göre ve yetenek kestirim yöntemi olarak, EYOY kullanılarak elde edilmiştir. SimulCAT programının varsayılan ayarları; madde kullanım sıklığının kontrol edilmemesi, kapsam dengelemesinde (content balancing) herhangi bir kontrol yapılmaması ve başlangıç maddesi için yeterlik düzeyi -0.5 ile 0.5 arasında herhangi bir madde ile başlaması gibi özellikleri değiştirilmemiştir. Simülatif BOBUT uygulamasında, EYOY, Bayes ve farklı sonlandırma kuralları (sabit soru sayısı ve sabit standart hata değeri) kullanılarak yeterlik kestirimleri yapılmış, elde edilen yeterlik kestirimleri ve standart hata değerleri incelenmiştir. Güvenirlik ve SH arasındaki ilişki (Wang, Hanson ve Lau, 1999):

$$r^2 = 1 - SH(\theta)^2 \quad (1)$$

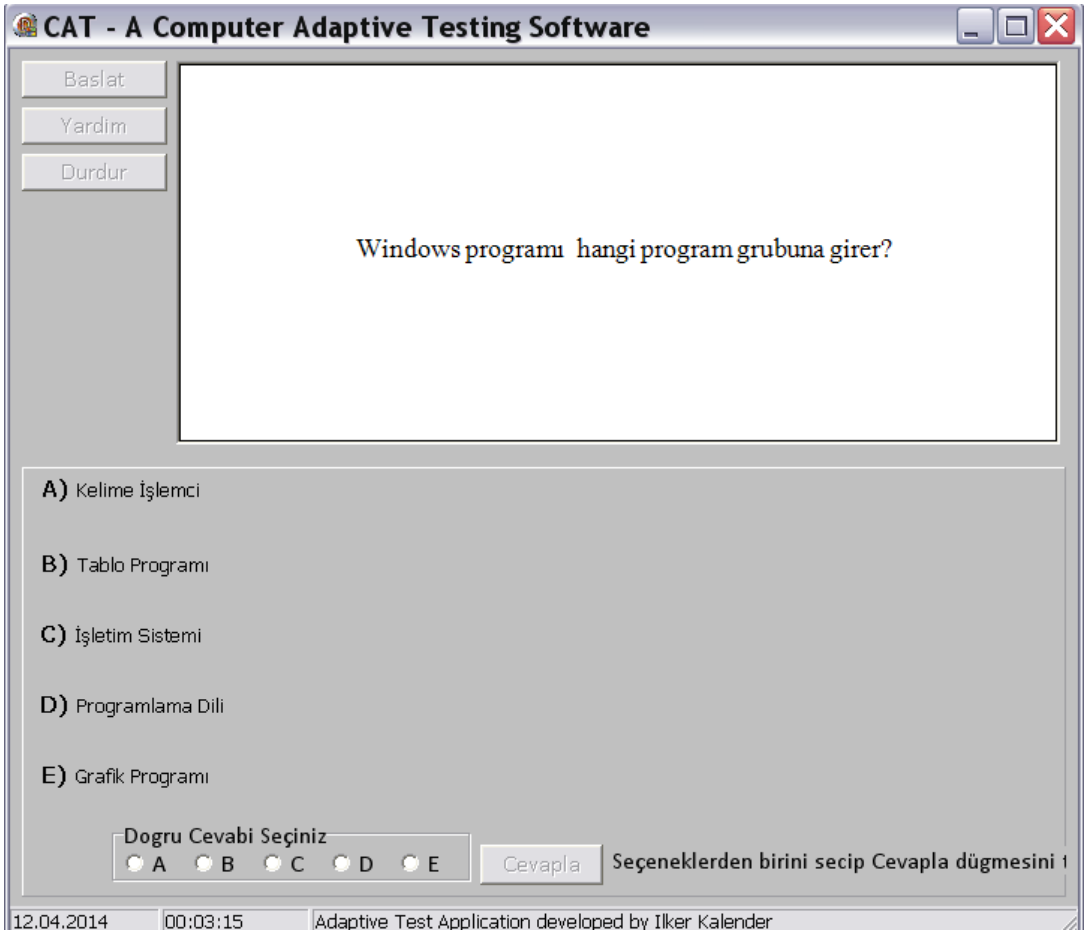
şeklinde tanımlanmaktadır. Babcock ve Weiss (2002) çalışmalarında test durdurma koşulunu belirlerken, güvenirliğin karesini göz önünde bulundurmuşlardır. Bu araştırma kapsamında, klasik test kuramındaki r^2= 0.91 güvenirlik değerine karşılık gelen standart hata değeri 0.30 ve r^2= 0.75 güvenirlik değerine karşılık gelen 0.50 standart hata, test sonlandırma koşullarında kesme değeri olarak alınmıştır. Simülasyon çalışmasında karşılaştırılan BOBUT stratejileri Tablo 7'de özetlenmiştir.

Tablo 7. Araştırmada Uygulanan Simülatif BOBUT Stratejileri

Teste Başlama Kuralı	Yeterlik Kestirim Yöntemi	Sonlandırma Kuralı
-0.5<θ<0.5	EYOY	Sabit test uzunluğu (k=30 madde) Ölçmenin Standart Hatası<0.30 Ölçmenin Standart Hatası<0.50
-0.5<θ<0.5	BSD	Sabit test uzunluğu (k=30 madde) Ölçmenin Standart Hatası<0.30 Ölçmenin Standart Hatası<0.50

Canlı BOBUT Uygulaması

Bilgisayar ortamında bireye uyarlanmış testin gerçek ortamda uygulanabilmesi için Kalender (2011) tarafından geliştirilen BOBUT yazılımı kullanılmıştır. Kağıt-kalem testi araştırmacı tarafından geliştirilen soru bankasına bağlı olarak, belirtke tablosundaki kazanımların dağılımlarına eşit oranda toplamda 30 soru olacak şekilde oluşturulmuştur. Soru yükleme işlemi bittikten sonra bilgisayar ortamında öğrencilerin soruları yanıtlamaları için uygulama yazılımı kullanılmıştır. Kullanılan yazlımın ekran görüntüsü Şekil 7'de verilmiştir.



Şekil 7. Canlı BOBUT Uygulamasında Bir Maddenin Ekran Görüntüsü

Öğrencinin sorulara verdiği bir doğru ve bir yanlış yanıttan sonra yeterlik düzeyi ve standart ölçme hatası hesaplanmaya başlamaktadır. Testin başlangıcında madde havuzunun en kolay beş maddesi arasından tesadüfi olarak biri başlangıç maddesi olarak ekrana gelmektedir. Sonlandırma kuralı ise, sabit soru ve sabit hata koşullarında çalışılmıştır. Sabit hata koşuluna göre testi durdurmada, yeterlik düzeyine bağlı olarak hesaplanan standart hata değeri 0.30 ve bir diğer koşul olarak da 0.50 olarak belirlenmiştir. Eğer 30 madde boyunca standart hata değeri 0.30/0.50 değerine ulaşmazsa oturum sona ermektedir. Yeterlik kestirimleri EYOY yöntemi kullanılarak yapılmıştır.

BULGULAR

Araştırmanın ilk sorusu olan, BOBUT uygulamalarında kullanılan yeterlik kestirim yöntemi ve test sonlandırma kurallarına göre yeterlik kestirimleri yapılmış ve Tablo 8'de bu kestirimlere ait betimsel istatistikler verilmiştir.

Tablo 8. Simülatif BOBUT Uygulamasında Farklı Sonlandırma Kuralı (ölçme hatası ve sabit madde sayısı) ve Yeterlik Kestirim Yöntemlerinden Elde Edilen Yeterlik Kestirimlerine ve Madde Sayılarına Ait Betimsel İstatistikler

Sonlandırma Kuralı	Yeterlik Kestirim Yöntemi	N	Ortalama Yeterlik	Standart sapma	Güvenirlik Ölçüleri	Ortalama Madde Sayısı
SMS (k=30)	BSD	1366	0.01	0.93	14.94	SMS
	EYOY	1366	-0.18	1.05	13.97	SMS
SH<0.30	BSD	1366	-0.00	0.95	12.91	59.8
	EYOY	1366	0.10	0.62	17.83	56.22
SH<0.50	BSD	1366	-0.55	1.03	5.73	11.11
	EYOY	1366	0.01	0.89	7.50	12.21

SH: Standart Hata SMS: Sabit Madde Sayısı

Tablo 8'de görüldüğü gibi, simülatif BOBUT'a dayalı karşılaştırmalar yapmak üzere, ilk olarak madde sayısı 30 olacak şekilde sabitlenmiş ve yeterlik kestirimleri EYOY ve BSD ile ayrı ayrı tekrarlanmıştır. Daha sonra SH<0.30 ve SH<0.50 olarak sabitlendiği koşulda EYOY ve BSD yöntemlerine göre yeterlik kestirimleri ayrı ayrı hesaplanmıştır. Simülatif BOBUT uygulamasında kestirilen yeterlik ölçülerine ilişkin güvenirlik ölçüleri, sabit madde test sonlandırma koşulunda, diğer test sonlandırma koşullarına göre daha yüksek bulunmuştur. Ayrıca kullanılan madde sayısı bakımından, en az madde kullanımı, test sonlandırma koşulunun SH<0.50 olduğu durumda gerçekleşmiştir (simülatif BOBUT uygulaması yeterlik kestiriminde kullanılan madde sayısı EYOY ve SH<50 koşulunda en fazla 101, en az 10; diğer tüm simülatif BOBUT yeterlik kestirim koşullarında kullanılan madde sayısı en az 10, en fazla 136 olarak gerçekleşmiştir).

Simülatif BOBUT uygulamasında elde edilen yeterlik kestirimleri için oluşturulan farklı durumlar sonucunda elde edilen yeterlik ölçüleri arasında anlamlı bir farklılığın olup olmadığını incelemek amacıyla gruplar arasında tek yönlü varyans analizi yapılmış ve Tablo 9'da verilmiştir.

Tablo 9. Farklı Test Sonlandırma Koşullarında (Sabit Madde ve Sabit Hata) Farklı Yeterlik Kestirimi koşullarında (BSD ve EYOY) Tekrarlanan BOBUT Uygulamalarından Kestirilen Yeterlik Ölçülerinin Farklılığına İlişkin Anova Sonuçları

Varyansın Kaynağı	Kareler Toplamı	sd	Kareler Ortalaması	F	p	Anlamlı Fark
Gruplararası	396.57	5	79.31	92.76	.00	1-2; 1-4; 2-3; 2-4; 2-5; 2-6; 3-5; 4-5; 5-6
Gruplariçi	7002.11	8190	0.85			
Toplam	7398.69	8195				

1: BSD ve sabit madde; 2: EYOY ve sabit madde; 3: BSD ve Sabit hata (SH<0.30); 4: EYOY ve sabit hata (SH<0.30); 5: EYOY ve sabit hata (SH<0.50); 6: BSD ve sabit hata (0.50)

Tablo 9'a göre, test sonlandırma kuralı olarak sabit madde ve sabit hata koşullarında farklı kestirim yöntemleriyle elde edilen yeterlik ölçüleri arasında anlamlı bir fark bulunmuştur [F(5-8190)=92.76; p<0.01]. Başka bir ifadeyle, kestirilen yeterlik parametreleri kullanılan BOBUT stratejisine göre değişmektedir. BOBUT uygulama koşullarına göre oluşan bu farklılığın, hangi yeterlik kestirim yönteminin lehine olduğunu tespit etmek amacıyla yeterlik ölçülerine Dunnet C testi uygulanmıştır. Buna göre, BSD ve sabit madde koşulunda uygulanan BOBUT uygulamasında kestirilen yeterlik puanları ortalaması ($\bar{X} = 0.01$), EYOY ve sabit madde stratejine göre kestirilen yeterlik puanları ortalamasından ($\bar{X} = -0.18$) daha yüksek bulunmuştur. Yine BSD ve sabit madde koşulu ile kestirilen yeterlik puanları ortalaması ile EYOY ve sabit hata (SH<0.30) koşuluna göre kestirilen ortalama yeterlik ($\bar{X} = 0.10$) değerine göre daha küçük olduğu bulunmuştur. Yapılan bazı araştırmalarda da (Bulut ve Kan, 2012; Wang, Kuo, Tsai ve Laio, 2012) BSD stratejisine göre kestirilen yeterlik puanlarının, kağıt-kalem testinden elde edilen yeterlik puanları ile arasındaki ilişkiler incelenmiş ve oldukça yüksek korelasyonlar elde edilmiştir. Bu çalışmada da BSD stratejisi ile kestirilen yeterlik puanları ortalamasının kağıt-kalem testi puanları ortalamasına ($\bar{X} = 0.00$) yakın olması, bu bulgunun yapılan çalışmalar (Bock ve Misley, 1982; Raiche ve Blais, 2002) ile paralellik gösterdiğini desteklemektedir.

Araştırmanın ikinci alt amacı doğrultusunda, simülatif BOBUT uygulamasına ilişkin farklı koşullarda elde edilen kestirimler sonucunda, EYOY ve SH<0.30 koşulunda elde edilen kestirimlerin güvenirlik ölçülerinin daha yüksek olmasından hareketle, canlı BOBUT uygulamasında test sonlandırma kuralı olarak ölçmenin standart hatasına (SH<0.30) koşulu belirlenmiştir. Ayrıca madde havuzunda yeteri kadar madde bulunmaması durumunda bireye en fazla 30 madde uygulama koşulu da canlı BOBUT uygulama yazılımında sabitlenmiştir. Yeterlik kestirimi olarak da, EYOY seçilmiştir. Böylece 142 öğrenci BİLOKUR testini hem kağıt-kalem hem de BOBUT olarak almıştır. Katılımcıların her iki uygulamadan teste verdikleri yanıtlardan kestirilen birey yeterlik parametreleri karşılaştırıldığında Tablo 10' da özetlenen bulgular elde edilmiştir.

Tablo 10. BOBUT Uygulaması ve Kağıt-Kalem Testinden Elde Edilen Yeterlik Ölçülerine Ait Betimsel Sonuçlar

	Kağıt-Kalem				Canlı BOBUT			
	Ortalama	Ss	En küçük	En yüksek	Ortalama	Ss	En küçük	En yüksek
Yeterlik ölçüleri	-0.06	0.73	-1.76	1.83	-0.42	1.32	-2.85	2.14
Hata	0.61	0.08	0.39	0.94	0.32	0.09	0.14	0.63
Güvenirlik	2.79	0.83	1.13	6.58	12.39	8.54	2.52	51.02

Tablo 10 incelendiğinde, her iki test uygulamasındaki standart hatalar ve güvenirlik değerleri karşılaştırıldığında, BOBUT uygulamasından elde edilen yeterlik kestirimlerine ait ortalama standart hata değeri ile kağıt-kalem ortamında uygulanan testen elde edilen yeterlik değerlerine ilişkin ortalama standart hata değerleri arasında oldukça büyük bir farkın oluştuğu Tablo 10'dan anlaşılmaktadır. Bu durumda, EYOY kestirim yönteminin uygulandığı BOBUT testinde öğrencilerin yeterlik ölçüleri, kâğıt-kalem testinden elde edilenlere göre, özellikle uç değerlerde daha iyi ölçme sonucu vermiş, bir diğer ifadeyle, standart hata değeri açısından da BOBUT uygulamasıyla daha güvenilir yeterlik kestirimleri elde edilmiştir

Canlı BOBUT ve kağıt-kalem testinden elde edilen yeterlik ölçülerinin sıraları arasındaki tutarlılık "Spearman Brown Sıra Farkları" korelasyonu incelenmiştir. Buna göre, her iki uygulamadan elde edilen yeterlik ölçülerinin sıraları arasında pozitif yönde iyi düzeyde (r= 0.70; p<.01) bir korelasyon bulunmuştur. Bu bulguya göre, canlı BOBUT uygulaması ile kağıt-kalem testi uygulamasından elde edilen yeterlik ölçüleri arasında daha yüksek tutarlılığın bulunmamasının nedeni öğrencilerin test

alma davranışlarından (sınav olma motivasyonlarının gerçek sınav ortamındaki kadar yüksek olmaması vb.) kaynaklanmış olabilir. Buna göre, canlı BOBUT uygulaması ile kağıt-kalem testinden elde edilen yeterlik ölçüleri arasında çok büyük bir değişiklik olmamakla birlikte, BOBUT uygulaması ile daha az sayıda madde ile daha güvenilir yeterlik parametreleri elde etmek mümkündür. Araştırmanın bu bulgusu, alan yazında yapılan diğer araştırmalarla (Bulut ve Kan, 2012; Chae ve Diğerleri, 2000; Frick, 1992; İşeri, 2002; Kalender, 2011; Kaptan, 1993; Kezer, 2013; Koşan-Aytuğ, 2013; McDonald, 2002; Miller, 2003; Mills ve Stocking, 1996; Mills ve Steffen, 2000; Öztuna, 2008; Rudner ve Guo, 2011; Scfhaer ve Diğerleri, 1995; Tian ve Diğerleri, 2007; Zitny ve Diğerleri, 2012) tutarlık göstermektedir.

Canlı BOBUT uygulaması ile kağıt-kalem testlerine ilişkin güvenirlik ölçülerini karşılaştırmak için her iki uygulamada da elde yeterlik ölçülerine ilişkin test bilgi (test information) miktarları hesaplanmıştır. Test bilgi miktarları arasında anlamlı bir fark olup olmadığını bulmak amacıyla ilişkili örneklemler için t-testi kullanılmıştır. Yapılan t-testi sonuçları Tablo 11'de verilmiştir.

Tablo 11. Canlı BOBUT ve Kağıt-Kalem Testlerine İlişkin Test Bilgi Miktarlarına Ait t-testi Sonuçları

	N	$\bar{X}$	Ss	Sd	t	p	En Küçük	En Yüksek
Kağıt-Kalem Testi	142	2.79	0.839	143.72	13.31	0.000	1.13	6.58
Canlı BOBUT	142	12.39	8.55				2.52	51.02

Tablo 11'e göre, test bilgi miktarları ortalamalarına ilişkin olarak, Canlı BOBUT uygulaması ile kağıt-kalem testinden elde edilen değerler arasında anlamlı bir farklılık göstermektedir [$t_{(143.72)}$=13.31; p<.05]. Canlı BOBUT uygulamasından elde edilen ortalama test bilgi miktarı ($\bar{X}$=12.39), kağıt-kalem testinden elde edilen ortalama test bilgi miktarından ($\bar{X}$=2.79) oldukça yüksek bulunmuştur. Buna göre, Canlı BOBUT uygulaması sonuçlarından elde edilen test bilgi miktarları, kağıt-kalem testinden elde edilen değerlerden daha yüksektir.

SONUÇLAR ve TARTIŞMA

Bu araştırmada, Ankara Üniversitesi Bilgi ve İletişim Teknolojileri dersinden muaf tutulacakları belirlemek üzere uygulanan BİLOKUR testinin Bilgi ve İletişim Teknolojisi Kavramları modülünün BOBUT olarak uygulanabilirliği çeşitli koşullarda test edilmiştir. Araştırmada bu amaçla, gerçek veriler kullanılarak kurgulanan sabit soru, sabit hata durdurma kurallarına ile iki farklı yeterlik kestirim yöntemine (EYOY ve BSD) göre, post hoc simülasyon çalışmaları gerçekleştirilmiştir.

Araştırmada elde edilen bulgulardan şu sonuçlara ulaşılmıştır: (1)Simülatif BOBUT uygulamasında farklı test sonlandırma koşullarında (sabit madde, SH<0.30 ve SH<0.50) farklı yeterlik kestirim yöntemlerinden (EYOY ve BSD) elde edilen yeterlik ölçüleri karşılaştırıldığında, sabit madde (30) ve SH<0.50 test sonlandırma koşulunda kestirilen yeterlik ölçülerinde BSD daha büyük değerler alırken; SH<0.30 test sonlandırma koşulunda elde edilen yeterlik ölçülerinden EYOY ile kestirilen yeterlik ölçülerinin daha güvenilir olduğu bulunmuştur. Bu bağlamda, farklı test sonlandırma koşullarında EYOY ve BSD yaklaşımlarının farklı güvenirlikte yeterlik kestirimlerinde bulunduğu söylenebilir. Kezer (2014) tarafından yapılan bir araştırmada ise, EYOY ve BSD yaklaşımları arasında bir fark bulmazken, Wang (1997) tarafından yapılan bir araştırmada BSD'nin EYOY'a göre nispeten daha yüksek standart hata barındırabileceği belirtilmiştir. Eggen (2004) ise test başlatma ve sürdürme işleminde en iyi yöntemin EYOY olduğu belirtmiştir. Bu durumun daha çok kullanılan soru bankasının büyüklüğü ve soru bankasındaki sorularının kalitesi ile ilgili olabileceği düşünülmektedir.

Araştırmanın ikinci bölümünde, ilk bölümden elde edilen BOBUT uygulama stratejilerine dayalı olarak, canlı BOBUT uygulaması ile kağıt kalem testinden elde edilen yeterlik ölçümleri karşılaştırılmıştır. Bu iki uygulama formatı arasındaki farklar incelendiğinde ise, özellikle testle ölçülen yeterlik düzeyi $\theta<0$ olduğu durumlarda bu farkların büyüdüğü, $0<\theta<2$ arasında bu farkın azaldığı bulunmuştur.(2) BİLOKUR testi kağıt-kalem testi olarak uygulanması ile BOBUT olarak uygulanması arasında yeterlik kestirimi açısından çok büyük farkın olmadığı sonucuna ulaşılmıştır. (3)Ayrıca, aynı testi canlı BOBUT ve kağıt-kalem alan katılımcıların iki uygulamadan elde edilen yeterlik düzeyleri arasında büyük ölçüde tutarlık bulunmuştur. Bu sonuç, canlı BOBUT uygulaması ile kağıt-kalem testinin kestirimlerinin birbirine yakın olduğunu, öğrencilerin yeterlik ölçümlerindeki sıralarının her iki uygulamada da çok değişmediğini göstermiştir. Bir başka deyişle, BİLOKUR testinin ilgili modülünün BOBUT olarak uygulanabileceğine ilişkin önemli bir kanıt elde edilmiştir. Benzer sonuçlar Kalender (2011) ve Kezer (2014) tarafından yapılan araştırmalarda da bulunmuştur. Adı geçen araştırmalar üniversite öğrencilerinin katılımıyla gerçekleşmiş; kağıt-kalem testi ile canlı BOBUT uygulamalarına ilişkin test puanları arasında arasında pozitif yönde ilişkiler bulmuşlardır. (4) Canlı BOBUT uygulaması ile kağıt kalem testinden elde edilen yeterlik parametrelerine ait test bilgi miktarları karşılaştırıldığında ise, canlı BOBUT uygulamasında, kağıt-kalem testinden elde edilen test bilgi miktarının anlamlı düzeyde yüksek olduğu ve bu nedenle de daha güvenilir yeterlik kestirimleri yaptığı sonucuna ulaşılmıştır. Yurt içi ve yurt dışında BOBUT ile ilgili yapılan gerek simülatif gerekse canlı BOBUT uygulamlarının birçoğunda (Bulut ve Kan, 2012; Chae ve Diğerleri, 2000; Frick, 1992; Kalender, 2011; Kaptan, 1993; Kezer, 2013; Koşan-Aytuğ, 2013; McDonald, 2002; Miller, 2003; Mills ve Steffen, 2000; Öztuna, 2008; Rudner ve Guo, 2011; Thompson & Weiss, 2011;Weiss ve Betz, 1973) benzer sonuçlar elde edilmiştir. Buna göre, BOBUT uygulaması ile elde edilen kestirimlerin kağıt-kalem testi kestirimlerine göre daha güvenilir olduğu söylenebilir.

Bu araştırma elde edilen sonuçlar doğrultusunda yapılacak araştırmalar için çeşitli önerilerde bulunulabilinir: Soru havuzunun büyük olmadığı durumlarda, yeterlik kestirim yöntemi olarak EYOY'un tercih edilmesi daha uygun olacaktır. Çünkü BSD yeterlik kestirim yöntemi, EYOY'a göre daha fazla madde ile kestirim yapmaktadır. Güvenilir yeterlik kestirimi yapmak için, test sonlandırma kuralı olarak standart hata ve yeterlik kestirim yöntemi olarak EYOY'un tercih edilebilir. Araştırmada canlı BOBUT uygulamasında BİLOKUR testinin sadece bir modülü kullanılmıştır. Daha sonra yapılacak çalışmalarda, bilgi ve iletişim teknolojilerine ilişkin modüller için de BOBUT çalışmaları yapılarak, kâğıt-kalem uygulaması ile sonuçları karşılaştırılabilir. Bu araştırmada maddenin kullanım sıklığı (item exposure) kontrol altına alınmamıştır. Yapılacak araştırmalarda daha geniş bir madde havuzu ile bu faktörün kontrol edildiği koşullarda uygulamanın psikometrik özellikleri test edilebilir.

KAYNAKÇA

Babcock, B., & Weiss, J. (2009, June 2). *Termination criteria in computerized adaptive tests: variable-length cats are not biased.* Kasım 12, 2011 tarihinde Realities of CAT Paper Session: http://www.psych.umn.edu/psylabs/catcentral/ adresinden alındı

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2010). A Method for the comparison of item selection rules in computerized adaptive testing. *Applied Psychological Measurement*(34), 438-452.

Baykul, Y. (2010). *Eğitimde ve psikolojide ölçme: klasik test teorisi ve uygulaması* (2 b.). Ankara: Pegem Akademi yayınevi.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement, 6, 431–444.

Bulut, O., & Kan, A. (2012). Application of computerized adaptive testing to Entrance Examination for Graduate Studies in Turkey. *Eurasian Journal of Educational Research*(49), 61-80.

Chae, S., Kang, U., Jeon, E., & Linarce, J. M. (2000). *Development of computerized middle school achievement test.* Seuol, South Korea: Komesa Press.

Çıkrıkçı- Demirtaşlı, N. (2002). A study of Raven Standard Progressive Matrices Test's item measures under classical and item response models: an empirical comparison. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 35*(1-2), 71-79.

Çıkrıkçı-Demirtaşlı, N. (1999). Psikometride yeni ufuklar: Bilgisayar ortamında bireye uyarlanmış test. *Türk Psikoloji Bülteni, 5*(13), 31-36.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Philadelphia: Harcourt Brace Jovanovich College Publishers.

Davis, L.L. & Dodd, B.G. (2005). *Strategies for controlling item exposure in computerized adaptive testing with partial credit model.* Pearson Education Measurement.

De Ayala, R. (2009). *The Theory and practise of item response theory.* New York Kondon: The Guilford Press.

ECDL. (2007). *European Computer Driving Licence 5.0 Müfredat Versiyonu.* Ocak 1, 2012 tarihinde European Computer Driving Licence Foundation: www.ecdl.com adresinden alındı

Eggen, , T. J. (2004). *Contributions to the theory and practice of computerized adaptive testing.* Arnhem, NL.: Omslag: Roel Ottow / Harold Kainama, Druk: Print Partners Ipskamp B.V., Enschede, Citogroep.

Embretson, E., & Reise, P. (2000). *Item response theory for psychologists.* New Jersey: Lawrence ErlbaumAssociates, Publishers Mahwah.

Fan, X. (1998). Item response theory and classical test theory: An Empirical comparison of their ıtem-person statistics. *Educational and Pscyhological Measurement, 58*(3), 357-382.

Frick, T. (1992). Computerized adaptive mastery tests as a expert systems. *Journal of Educational Computing Research, 8*(2), 182-213.

French, B.F. & Thompson, T.D. (2003). The Evaluation of exposure control prosedures for an operational CAT. The annual meeting of the American educational Research Assıciation, Chiccago.

Frick, T. (1992). Computerized adaptive mastery tests as a expert systems. Journal of Educational Computing Research, 8(2), 182-213.

Hambleton, K., & Swaminathan, H. (1985). *Item response theory : Principles and applications.* Hingham, MA, U.S.A. : Distributors for North America, Kluwer Boston.: Kluwer-Nijhoff Pub.

Hambleton, K., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory* (volume 2 b.). Newbury, London.

Han, K. T. (2010, Mart 20). *SimulCAT: Simulation software for computerized adaptive testing [computer program].* Eylül 17, 2013 tarihinde SimulCat: http://www.hantest.net/ adresinden alındı

İşeri, İ. (2002). *Assessment of students' mathematics achievement through computer adaptive testing procedures.* Yayınlanmamış Doktora Tezi. Ankara: ODTÜ Ortaöğretim Fen ve Matematik Alanları Eğitimi Bölümü.

Kalender, İ. (2009). Başarı ve yetenek kestirimlerinde yeni bir yaklaşım: Bilgisayar ortamında bireyselleştirilmiş testler (computerized adaptive tests-CAT). *CITO Egitim Kuram ve Uygulama*(5), 39-48.

Kalender, İ. (2011). *Effects of different computerized adaptive testing strategies on recovery of ability.* Yayınlanmamış Doktora Tezi. Ankara: ODTU.

Kaptan, F. (1993). *Yetenek Kestiriminde adaptive (bireyselleştirilmiş) test uygulaması ile geleneksel kağıt-kalem testi uygulamasının karşılaştırılması.* Yayınlamamış Doktora Tezi. Ankara: Hacettepe Üniversitesi.

Karasar, N. (2011). *Bilimsel araştırma yöntemi* (22 b.). Ankara: Nobel yayın Dağıtım.

Kelecioğlu, H. (2001). Örtük özellikler teorisindeki b ve a parametreleri ile klasik test teorisindeki p ve r istatistikleri arasındaki ilişki. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*(20), 104-110.

Keller, A. (2000). *Ability estimation procedures in computerized adaptive testing.* USA: American Institute of Certified Public Accountants-AICPA Research Concortium-Examination Teams.

Kezer, F. (2013). *Bilgisayar ortamında bireye uyarlanmış test stratejilerinin karşılaştırılması.* Yayınlamamış Doktora Tezi. Ankara: Ankara Üniversitesi Eğitim Bilimleri Enstitüsü.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items forcomputerized adaptive tests. *Applied Measurement in Education,, 2*(4), 359-375.

Kline, P. (2000). *An easy guide to factor anaylsis.* Routledge, Taylor and Franchis groups.

Köklü, N. (1990). *Klasik test teorisinie gore gelistirilen tailored test ile grup testi arasında bir karşılaştırma.* Yayınlanmamış Doktora Tezi. Ankara: Hacettepe Üniversitesi.

Koşan-Aytuğ, M. (2013). *Tıp eğitiminde gelişim sınavı soru bankası oluşturulması ve benzetim verileri ile bilgisayar uyarlamalı test uygulaması.* Yayınlanmamış Doktora Tezi. Ankara: Ankara Üniversitesi Eğitim Bilimleri Enstitüsü.

Lord, F. (1980). *Applications of Item Response Theory to Practical Testing: Problems.* New Jersey: Lawrence Erlbaum.

Lord, F., & Stocking, M. (1988). Item response theory. J. Keeves içinde, *Educational Research, Methodology, And Measurement: An International Handbook* (s. 269-272). NewYork: Pergamon press.

Mcdonald, P. (2002). *Computer adaptive test for measureing personality factors using item response theory.* Unpublished Doctoral Dissertaion. London: The University Western of Ontario.

Mcglohen, M., & Chang, H. H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods, 40*(3), 808-821.

Miller, D. (2003). *Assessment of student achievement: A comparative study of student achievement using paper and pencil assessment and computerized adaptive testing (CAT).* Unpublished Dissertation. Mich: Graduate School of Wayne State University.

Mills, N., & Steffen, M. (2000). The GRE computer adaptive test: operational issues. V. d. Glass içinde, *Computerized adaptive testing: Theory and practice.* Netherlands: Kluwer Academic Publishers.

Mills, N., & Stocking, L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education, 9*(4), 287-304.

Öztuna, D. (2008). *Kas-İskelet sistemi sorunlarının özürlülük değerlendiriminde bilgisayar uyarlamalı test yönteminin uygulanması.* Yayınlanmamış Doktora Tezi. Ankara: Ankara Üniversitesi Sağlık Bilimleri Ensititüsü.

Raîche, G. & Blais, J.-G. (2002). Practical Considerations about expected a posteriori estimation in adaptive testing: Adaptive a priori, adaptive correction for bias, and adaptive integration interval. Communication at the 11th Biennal intenational objective measurement workshop. New Orleans, LO: IOMW. [ERIC DOCUMENT NO ED 464 110]

Reise, P. (1990). A Comparison of item and person fit methods of assessing model data fit in IRT. *Applied Psychological Measurement, 14*(2), 127-137.

Rudner, L. (2002). An Examination of decision-theory adaptive testing procedures. *Paper presented at the annual meeting of the American Educational Research Association, April 1-5.* New Orleans, LA.

Rudner, L. (1998, Kasım). *An On-line, Interactive, Computer Adaptive Testing Tutorial*. Ocak 12, 2012 tarihinde Measurement Resources from EdRes.org: http://edres.org/scripts/cat adresinden alındı

Rudner, L., & Guo, F. (2011). *Computerized adaptive testing for small scale programs and instructional systems.* Graduate Management Admission Council-GMAC.

Scfhaer, A., Steffen, M., Golub-Smith, L., Mills, N., & Durso, R. (1995). *The introduction and comparability of the computer adaptive GRE general Test.* GRE Board Professional Report No 88-08a.

Sheskin, D. (2004). *Handbook of parametric and nonparametric statistical Procedures.* Ewyork, Chapman&Hall/CRC.

Smits, N., Cuijer, P., & Straten, A. (2011). Applying computerized adaptive testing to the CES-D Scale: A simülation study. *Psychiatry Research*(188), 145-155.

Stark, S., & Chernyshenko, O. S. (2001). Examining model-data fit using graphical and statistical methods. *16th annual conference for the Society of Industrial and Organizational Psychology.* San Diego, CA.

Stocking, M. (1990). Specifying optimum examinees for item parameter estimation in item response theory. *Psychometrika*(55), 461-475.

Thompson, A., & Weiss, D. (2011). A Framework for the Development of Computerized Adaptive Tests. *Practical Assessment Research & Evaluation, 16*(1), 1-9.

Tian, J.-Q., Miao, D.-M., Zhu, X., & Gong, J.-J. (2007). An ıntroduction to the computerized adaptive testing. *US-China Education Review, 4*(1).

Tomei, A. (2005). *Taxonomy For The Technology Domain.* Hersbe: Information Science Publishing: ISBN 1-59140-526-2.

van Rijn, P., Eggen, T. J., Hemker, B. T., & Sanders, P. F. (2002). Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Applied Psychological Measurement*(26), 393-411.

Wainer, H., Dorans, J., Flaugher, R., Green, F. B., Mislevy, R., Steinberg, L., et al. (1990). *Computerized adaptive testing: A Primer.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Wang, T., Hanson, B., & Lau, C. (1999). Reducing bias in cat trait estimation: A Comparison of approaches. *Applied Psychological Measurement, 23*(3), 263-278.

Weis, J. D., & Betz, E. N. (1973). *Ability measurement: Conventional or adaptive.* Minnesota, USA: Psychometric Methods Program Department of Psychology University of Minnesota.

Weiss, D. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37*(2), 70-84.

Yaşar, M. (1999). *Bireyselleştirilmiş testler üzerine bir çalışma.* Yayınlanmamış Doktora Tezi. Ankara: Hacettepe Üniversitesi.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG (Version 3.0) [computer program]*. Mooresville, IN: Scientific Software.

Zitny, P., Halama, P., Jelinek, M., & Kveton, P. (2012). Validity of cognitive ability tests-comparison of computerized adaptive testing with paper and pencil computer-based forms of administrations. *Studia Psychologicia, 54*(3), 181-194.

EXTENDED ABSTRACT

Introduction

In parallel with the developments in computer Technologies in today's World, the tests and methods used in test application also develop accordingly. Thanks to the advancements in the World of computer technology, tests have been used as Computer Adaptive Testing (CAT) in education for various purposes (Selection, Placement, diagnosis and etc.) in last 20 years. CAT can be defined as tests which offer different test items for test-takers, which are prepared appropriately choosing items from pre-prepared test pool considering test takers' proficiency levels (Weiss, 2004). To achieve that, a new method is used rather than giving the same items with the same difficulty level to everybody, and in this method, if test-takers find the correct response, the next item gets more difficult, if test taker is wrong, the next item gets easier, which is known as more or less method (Rudner, 1998). Therefore, as the most appropriate item is directed to test takers considering their proficiency levels in CAT applications, a significant amount of decrease in the number of directed items is obtained. Thus, it becomes possible to have more reliable measuring results using less test items (Embretson & Reise, 2000; Mcglohen & Chang, 2008). There is very limited number of empirical studies in literature in Turkey. Relevant preliminary studies investigated the applicability of achievement test as CAT application (Kaptan, 1993; Köklü, 1990), and the following studies focused on a comparison between traditional applications (paper pencil test) and the application of the tests used to select students for higher education institutions and the test aiming to measure proficiency levels in CAT form (Aytuğ-Koşan, 2013; Bulut & Kan, 2012; İşeri, 2002; Kalender, 2011; Kezer, 2013).The purpose of the study is to investigate how applicable computer literacy test is in CAT form under various conditions. To reach that aim, psychometric qualifications of paper-pencil test form and CAT form will be compared, the most appropriate CAT application strategies (maintenance and termination of test) will be found out.

Method

Different proficiency estimate methods and different termination rules were compared to test how applicable computer literacy test was in CAT form. In this study, proficiency estimates were done considering post-hoc simulation method and different proficiency estimate methods (maximum likelihood estimation method-MLE) and expected A Posteriori Method- EAP and test termination rules (Standard error-SE) is equal to SE<0.30 and SE<0.50). For this purpose, SimulCAT software which is a CAT program developed by Han (2010), was used in the study. This research contributes to improvement and development of already existing hypothetic knowledge as well as contributing to application as a basic research model. According to Karasar (2011), fundamental research studies are the ones which contribute to the development of theories, hypothesis formation, test them and discuss the findings of the test. Then, the psychometric characteristics of proficiency levels obtained from live CAT and paper-pencil test were compared to those of a group of students. For this purpose, CAT application software which was developed by Kalender (2011) was used. With this regard, this study seems to be an applied research which aims to improve an already existing study.

The data of computer literacy test was collected from the first year students at various faculties of Ankara University in 2012-2013 and 2013-2014 education years. The study group participated in the research in two phases. In the first phase, 1452 university first year students were given a computer literacy test developed for this study dividing all the items in 6 groups in 2012-2013 academic years. In the last phase of the study, the items of the computer literacy test whose appropriacy to IRT was confirmed and whose parameters were estimated, were implemented on 142 university first year students at Ankara University, Faculty of Educational Science in 2013-2014 academic year in the form of paper-pencil and live CAT test. 142 students participated in the real CAT application, 32 (23%) of them were male, and 110 (77%) of them were female.

The items making up the computer literacy test were administered on 1366 students in about 2-month period, which is quite short in 2012-2013 academic year. As the item pool is very big, the

items were administered in 6 different item/test groups. After the application, item statistics were calculated based on CTT. According to that calculation, most of the items were found to be at average difficulty, few items were found to be easy. When the average item discrimination values of the items were examined, what was found was that it changed between the lowest was 0.33 and the highest was 0.69. In CAT application, using the items scaled based on IRT provides invariant estimates for the individual and item parameters of IRT. However, having these advantages is largely dependent on the compatibility of the used data with the model (Fan, 1998; Hambleton & Swaminathan, 1989) and revealing the evidence of the measurability of psychological feature by the available data (Stark & Chernyshenko, 2001). For this aim, the items left in the item bank were subject to preliminary analysis based on IRT. With this analysis, the following issues were tested, such as unidimensionality, meeting the local independence assumptions, the control of speed test and constancy of item and individual proficiency parameters.

Results and Discussion

According to research findings, a significant difference was found among the proficiency measurements obtained through various estimation methods under the condition of fixed item and fixed error as test termination rule. According to that, the means of proficiency scores estimated in CAT application applied under EAP and fixed item condition was found to be higher than that of proficiency score means estimated depending on MLE and fixed item strategy. The means of proficiency score average estimated under the condition of EAP and fixed item and average proficiency estimated under the condition of MLE and fixed error (SH<0.30) was found to be lower than ($\overline{X} = 0.10$). Some studies investigated the relationship between the proficiency score obtained from paper-pencil test and the proficiency score estimated according to EAP strategy, and quite high correlations were obtained (Bulut & Kan, 2012; Wang, Kuo, Tsai & Laio, 2012). In this study, the fact that the average proficiency score estimated through EAP strategy was close to that of paper-pencil test ($\overline{X} = 0.00$), demonstrates a parallelism with the findings of the studies carried out in the field.

When comparing the item quantities in competency prediction of the real CAT application and paper-and-pencil test, real CAT application provided huge savings. Also, when comparing the reliability values of real CAT and paper-and-pencil test, the reliability values gained from the real CAT application was found significantly higher. Therefore, it was concluded that Computer Literacy test carried out with the CAT application can predict the competency measures more reliably when compared with paper-and-pencil tests. This findings of the study seems to be consistant with those of the other studies in the literature (Bulut & Kan, 2012; Chae, Kang, Jeon and Linarce, 2000; Frick, 1992; İşeri, 2002; Kalender, 2011; Kaptan, 1993; Kezer, 2013; Koşan-Aytuğ, 2013; McDonald, 2002; Miller, 2003; Mills & Steffen, 2000; Mills & Stocking, 1996; Öztuna, 2008; Rudner & Guo, 2011; Scfhaer, Steffen, Golub-Smith, Mills & Durso, 1995; Tian, Miao, Zhu & Gong, 2007).

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2015; 6(2); 238-253

A Diagnostic Comparison of Turkish and Korean Students' Mathematics Performances on the TIMSS 2011 Assessment*

TIMSS 2011 Sınavına Kore ve Türkiye'den Katılan Öğrencilerin Matematik Performanslarının Tanısal Bir Karşılaştırılması

Sedat ŞEN** Muhammet ARICAN***

Abstract
The purpose of the present study was to analyze an international large-scale data set using a cognitive assessment approach. Although some researchers question the usefulness of international large-scale assessments (e.g., TIMSS), participating countries have continued to use the results from these large-scale assessments to improve their curricula and teaching methods. Despite the common reporting practice—single-score—in these large scale assessments gives useful insights about students' overall performances, they still lack diagnostic information. Cognitive diagnosis models (CDMs) were developed to provide more feedback on students' cognitive strengths and weaknesses. This study retrofitted the TIMSS 2011 eighth grade mathematics assessment by applying a specific CDM called the DINA (the deterministic, inputs, noisy, "and" gate) model to data from South Korea and Turkey. Results of the DINA model were used to make a detailed comparison between students of these two countries.

Key words: Mathematics assessment, cognitive diagnosis, DINA model, TIMSS

Öz
Bu çalışmanın amacı büyük ölçekli bir sınavın tanılayıcı değerlendirme yaklaşımlarından biriyle analiz edilmesidir. Bazı araştırmacılar büyük ölçekli sınavların (örn: TIMSS) kullanışlılığını sorguluyor olsa da katılımcı ülkeler bu sınavlardan alınan sonuçları kullanarak müfredatlarında ve öğretim metotlarında geliştirmeler yapmaya devam etmektedir. Bu sınavlarda yaygın olarak kullanılan ve tek bir puan sunmaya dayalı olan uygulamalar öğrencinin genel performansı hakkında bilgi sunsa da tanısal bilgi sunmada yeterli değildir. Öğrencilerin bilişsel olarak güçlü ve zayıf yanlarıyla ilgili daha detaylı bilgi sunabilmek için bilişsel tanı modelleri geliştirilmiştir. Bu çalışmada Kore ve Türkiye veri setleri kullanılarak TIMSS 2011 sekizinci sınıf matematik sorularının bilişsel tanı modellerinden DINA model ile tekrar analizi yapılmıştır. Bu modelden elde edilen sonuçlar kullanılarak iki ülke öğrencilerinin performanslarının karşılaştırılması yapılmıştır.

Anahtar kelimeler: Matematik eğitimi, bilişsel tanı, DINA model, TIMSS

INTRODUCTION

Since the first administration of the Trends in International Mathematics and Science Study (TIMSS) in 1995, the comparison of the relative performances of participating countries has become very helpful for finding out country-level success relative to other countries. Although some researchers question the relationship between student-level (or country-level) achievement and comparison studies based on such international large-scale assessments (Holliday & Holliday, 2003; Wang, 2001),

* This study was presented at the 2015 annual meeting of the American Educational Research Association Conference, Chicago, IL, USA.
** Asst. Prof. Dr., Harran University, Faculty of Education, Educational Sciences, Şanlıurfa-Turkey, e-mail: sedatsen@harran.edu.tr
*** Dr., Ahi Evran University, College of Education, Mathematics Education, Kırşehir-Turkey, e-mail: muhammetarican@gmail.com

Geliş Tarihi: 01.07.2015
Kabul Tarihi: 07.12.2015

participating countries have continued to use the results from these large-scale assessments (e.g., TIMSS) to improve their curricula and teaching methods to fill gaps or to reach excellence.

According to Toker and Green (2012), educational assessment is important since, by means of this assessment, educators evaluate the effects of educational programs and manage these programs. Because outcomes of education necessitate meeting a universally accepted criteria (Toker & Green, 2012), in mathematics education, researchers have been using international comparative studies (e.g., TIMSS, PISA) to evaluate students' achievements and their mastery of curricular instruction (Lee, Park, & Taylan, 2011). However, as discussed by Dogan and Tatsuoka (2008), since the reports on students' achievements and their mastery of curricular instruction rely on total scores and rankings of the participating countries, they do not provide enough information about students' strengths and weaknesses. The common reporting practice in these large scale assessments is to provide a single overall score for each student and report students' averages across their countries. Although the single test scores give useful insights about the overall performances in terms of subject areas, they still lack diagnostic information. The lack of the diagnosity of a single score based on test assessments has frustrated many researchers (Nichols, 2012). Hence, as Leighton and Gierl (2007) stated,

> There is increasing pressure to make assessments more informative about the mental processes they measure in students. In particular, there is increasing pressure to adapt costly large-scale assessments (Organization for Economic Co-operation and Development [OECD], 2004; U.S. Department of Education, 2004) to be informative about students' cognitive strengths and weaknesses. (p. 5)

In order to provide an example to show how diagnostic feedback can be given using real data, this study analyzes TIMSS 2011 data from a cognitive diagnostic assessment (CDA; Leighton & Gierl, 2007) perspective. Over the last two decades, the interest in CDA has increased in order to obtain more information about students' performances on a measurement. This type of assessment classifies students based on their degrees of mastery of specific skills. Thus, examiners and instructors can obtain more information relevant to classroom teaching and learning. Unlike a single-overall test score, CDA-based reports simply show what students know (master) and what they do not know (master) rather than how much they know.

The main purpose of this study is to examine Turkish eight graders' strengths and weaknesses on topics that were covered on the TIMSS 2011 mathematics achievement test. In order to do so, in this study, the relative performances of Turkish students in comparison with South Korean (Korea hereafter) students were assessed. Hence, this CDA-based study examines the following research questions:

1. How do Turkish and Korean eight graders' relative TIMSS 2011 mathematics performances differ?
2. What are the Turkish eight grade students' weaknesses and strengths on TIMSS 2011's mathematics topics in comparison to the Korean eight graders?

Literature Review

Several recent studies (e.g., Dogan & Tatsuoka, 2008; Im & Park, 2010; Lee et al., 2011; Toker & Green, 2012; Lee et al., 2013) have been conducted to compare students' achievements on international large-scale assessments (e.g., TIMSS, PIRLS) using DCMs. These studies have provided useful feedback on the students' performance and skills, the linkage between teachers' instruction and students' performances, and the countries' educational systems and their curricular instructions. For instance, Dogan and Tatsuoka (2008) compared Turkish and American eight-grade students' mathematics performances on the TIMMS-R 1999. Their results indicated that Turkish students were weak in algebra and probability/statistics in comparison to their American peers, and they also "demonstrated poor profiles in skills such as applying rules in algebra, approximation/estimation, solving open-ended problems, recognizing patterns and relationships, and quantitative reading" (Dogan & Tatsuoka, 2008, p. 263). Similarly, Im and Park (2010) compared Korean and American eight-grade students' mathematics performances on the TIMMS 2003. The results showed significant

differences in the performances of Korean and American students, especially in "problem restructuring and reasoning, measurement, and geometry" (p. 287). Their results suggested that encouraging students' independent problem solving was the most useful instructional strategy for both Korean and American students. Moreover, American students benefitted from reviewing, re-teaching, and clarifying as well. In addition to the above studies, Lee et al. (2011) compared the performances of fourth-grade students' in Massachusetts and Minnesota to the nationwide results (not including MA and MN) on the TIMSS 2007. Their results demonstrated that students in Massachusetts and Minnesota outperformed students in the US overall. Lee et al. (2011) also provided fine-grained diagnostic information on students' performances, which they suggest could be exactly applied to classroom instruction. For example, by analyzing item parameter estimates (e.g., slipping and guessing) they offered curricular suggestions to the classroom teachers on how to improve students' performances.

In this study, Korea was chosen as a reference country, because Korean eight graders have been regularly placed in the top three in TIMSS mathematics performance. As stated by Mullis, Martin, Foy, and Arrora (2012), 42 countries and 14 benchmarking entities participated in TIMSS 2011. In that assessment, the international TIMSS scale average was set to 500. Among 42 countries, Turkish students had an average score of 452 and were ranked in 24th place. Korean students had an average score of 613 and were ranked in first place on the TIMSS 2011. As explained by Im and Park (2010), several studies investigated which characteristics of Korean education have been contributing to such tremendous performance in mathematics. According to Im and Park (2010), the results of those studies pointed out that factors contributing to Korean students' high achievement could be grouped under social and instructional factors. Social factors included "competitive examination and selection, a regular and metric number system, the serious attitudes of students towards tests, meaningful repetitive learning, and the competence of mathematics teachers (Kim et al., 2008; Park, 2004)" (Im & Park, 2010, p. 288), and instructional factors included "cooperative learning activities (Chung & Son, 2000; House, 2009), the use of constructivist strategies (Fisher & Kim, 1999), and teachers' guidance (Oh, 2005)" (Im & Park, 2010, p. 288). These social and instructional factors also affected our decision to select Korea as the reference country.

Diagnostic Classification Models

A number of cognitive diagnosis models (CDMs), also known as diagnostic classification models (DCMs), have been developed (Rupp, Templin, & Henson, 2010) to apply the CDA approach. For an overview of DCMs, the reader is referred to DiBello, Roussos, and Stout (2007), Fu and Li (2007), Rupp and Templin (2008a), and Rupp et al. (2010). However, it should be noted de la Torre (2011) classified these psychometric models as either general or a specific type based on their characteristics. Specific DCMs include: d*eterministic inputs, noisy "and" gate* (DINA; Haertel, 1989; de la Torre, 2009; Junker & Sijtsma, 2001), *deterministic inputs, noisy "or" gate* (DINO; Templin & Henson, 2006), *noisy-input, deterministic "and" gate* (NIDA; Junker & Sijtsma, 2001), and the *reduced reparameterized unified model* (R-RUM; Hartz, 2002; Roussos et al., 2007). General DCMs include the log-linear cognitive diagnostic model (LCDM; Henson, Templin, & Willse, 2009), the general diagnostic model (GDM; von Davier, 2005), and the generalized DINA (G-DINA; de la Torre, 2011) model. This study focused on the DINA model. Thus, a brief description of the DINA model is presented below.

The DINA Model

The DINA model is a non-compensatory model with a conjunctive rule (Rupp et al., 2010). Based on the conjunctive nature of the DINA model, a respondent has to master all of the measured attributes of an item in order to get full credit for this item. Respondents get zero credit for an item if they did not master at least one of the measured attributes of this item. Thus, the DINA model divides respondents into two groups for each item: those who mastered all attributes and those who did not master all attributes. This is done with the conjunctive kernel of the DINA model, which is presented as a latent response vector (ξ_{ri}) below (Equation 1). Let X_{ri} be the response of examinee *r* to item *i*, and let

$\boldsymbol{\alpha}_r = \{\alpha_{rk}\}$ be the examinee's binary attributes vector, which is coded as 1 for presence or mastery of attribute *k* on the *k*th element and zero otherwise. Like most of the CDMs, the DINA model requires a Q-matrix (Tatsuoka, 1985) that shows the relationship among items (*i*,...,*I*) and attributes (*k*,...,*K*). A value of 1 for the Q-matrix entry (i.e., $q_{ik} = 1$) indicates that attribute *k* is measured for item *i*. For example, suppose we measure four attributes in an arithmetic test. Let addition, subtraction, division, and multiplication be four attributes coded as Attribute 1, Attribute 2, Attribute 3, and Attribute 4, respectively. Based on this attribute list and the DINA model specification, students have to master both Attribute 1 (addition) and Attribute 3 (division) in order to get full credit ($X_{ri} = 1$) for an item such as $\frac{4+8}{3}$. A student with mastery of addition or division cannot get full credit ($X_{ri} = 0$), as he/she would miss one of the required attributes for this item. The conjunctive kernel of the DINA model can be presented as below:

$$\xi_{ri} = \prod_{k=1}^{K} \alpha_{rk}^{q_{ik}} \qquad \text{(Equation 1)},$$

where ξ_{ri} is the latent variable which is coded as zero or one for respondent *r* and item *i*, and q_{ik} is the Q-matrix entry described above. α_{rk} represents the latent attribute variable indicating whether respondent *r* has mastered attribute *k* ($\alpha_{rk} = 1$) or not ($\alpha_{rk} = 0$). Thus, the latent response vector (ξ_{ri}) can have a value of 1 if respondent *r* masters all the attributes required for item *i* and a value of 0 if the respondent did not master at least one of the measured attributes for item *i*. It is possible that respondents who have mastered all attributes can give a wrong answer to item *i*, while respondents who have missed one of the required attributes can correctly answer item *i*. The former refers to slipping, and the latter refers to a guessing situation in the DINA model specifications. Thus, two parameters are obtained for each item in the DINA model regardless of the number of attributes. Item slipping (s_i) and guessing (g_i) parameters do not change across attributes, because they are item-specific. In the DINA model, these two item parameters are defined as follows:

$$s_i = P(X_{ri} = 0 | \xi_{ri} = 1) \qquad \text{(Equation 2)},$$

$$g_i = P(X_{ri} = 1 | \xi_{ri} = 0) \qquad \text{(Equation 3)}.$$

After defining slipping and guessing parameters, the probability of the correct response of a respondent in latent class *c* for item *i* can be computed as below:

$$P(X_{ri} = 1 | \xi_{ri}) = (1 - s_i)^{\xi_{ri}} {g_i}^{1-\xi_{ri}} \qquad \text{(Equation 4)}.$$

According to Equation 4, respondents need to master all attributes measured by an item in order to answer this item correctly. DINA model was used in this study, because the DINA model requires an estimation of two parameters for each item, and the number of attributes does not affect the number of estimated parameters in the DINA model. The DINA model is also an appropriate model for equally important items like TIMSS items. The DINA model has been used in analyses of the TIMSS data by several authors, including Lee et al., (2011) and Choi, Lee, and Park (2015).

METHOD

Subjects and Data

Data sets from the students of two countries (i.e., Korea and Turkey) were compared in this study. Data were taken from the TIMSS 2011 eighth grade mathematics test, which included 28 blocks (14 science and 14 mathematics) and 14 test booklets. Each booklet was composed of four blocks of items: two mathematics and two science blocks. Students responded to different types of questions including multiple-choice (four response options) and constructed responses assessing four content domains: Number (30%); Algebra (30%); Geometry (20%); and Data and Chance (20%). According to the

TIMSS 2011 design, only six of the 14 mathematics assessment blocks were made publicly available. Based on the pairs of released blocks, only four booklets (Booklets 1, 2, 5, and 6) can be obtained for an eighth grade mathematics assessment as administered in the real exam settings. Booklet sample sizes for Korea and Turkey and the number of items for different content domains are presented in Table 1. Each booklet showed different distributions for content domains. The administration of Booklet 2 to Korean and Turkish students was selected for the DINA model analyses in this study due to the following reasons: (a) there were relatively more topics—13—in Booklet 2; (b) the subject areas of the items were distributed evenly—nine items for Numbers, nine items for Algebra, seven items for Data and Chance, and seven items for Geometry; and (c) the cognitive domains among the items were also distributed evenly—10 items required knowing, 13 items required applying, and nine items required reasoning. Booklet 2 was composed of Block 2 and Block 3 with 32 items, including 15 multiple choice and 17 constructed response items. There were 368 Korean students and 488 Turkish students who had taken Booklet 2.

Table 1. Descriptive Characteristics of the TIMSS 2011Mathematics Booklets

Booklets	Blocks	Turkey (*N*)	Korea (*N*)	Number	Algebra	Geometry	Data and Science
Booklet1	M01-M02	503	410	8	9	5	4
Booklet2	M02-M03	488	368	9	9	7	7
Booklet5	M05-M06	490	369	7	9	10	6
Booklet6	M06-M07	494	361	5	12	8	8

Construction of Q-Matrix

Attributes, which are used to define skills required to solve a specific item, were adopted from the Common Core State Standards for Mathematics (CCSSM; Common Core State Standards Initiative, 2010). The CCSSM was developed as a result of recognizing the need for a more focused and coherent mathematics curriculum in the United States to improve the quality of mathematics education and to increase mathematics achievement to the level of high-performing countries (Common Core State Standards Initiative, 2010). Therefore, standards from high-performing countries played a significant role in the development of the CCSSM (Common Core State Standards Initiative, 2014). Thus, in this study, the CCSSM was used to determine our attributes. By means of carefully examining TIMSS items and the standards, a list of 13 attributes (see Table 2) was created. In order to generate attributes that cover all possible skills, some of the two related standards were combined and separated with semi-colons. Using the attribute list in Table 2, 32 items were coded independently by four doctoral students with advance degrees in mathematics education at one large public university in the Southeast. An attribute was included in our Q-matrix if at least two coders agreed that an item measured that attribute (see Table 3).

The attributes in the Q-matrix are independently generated by considering the required steps to solve each item. For example, in Item 6, students were given a picture of a rectangular garden that had a (*x* + *4*)-meter width and an *x*-meter height (see Figure 1). The garden consisted of two small rectangular gardens and one rectangular path. The path was 1 meter wide and was between the two small gardens. Students were asked to calculate the total area of the two small rectangular gardens, which were shaded, in m^2. In order to solve this problem, students need to master three attributes (Attributes 4, 5, and 11). First, they must understand the concept of area and relate area to multiplication—Attribute 11. Second, they need to multiply width and height for the big rectangular garden and for the rectangular path to calculate their areas. These two multiplication operations involve using algebraic expressions and require applying previous knowledge of arithmetic to algebra—Attribute 4. Third, they must know the distribution property, which also requires applying previous knowledge of arithmetic to algebra, and understand that the equivalent expressions of $x * (x + 4)$ and $x * 1$ are $x^2 + 4x$ and x—Attribute 5. In the last step, they can obtain the area of the shaded garden as $(x^2 + 3x)\,\mathrm{m}^2$ by subtracting *x* from $x^2 + 4x$. This last step also requires mastery of Attribute 4, since students who master Attribute 4 can apply arithmetic operations to algebraic equations. A student can solve this problem also by subtracting 1-meter from (*x* + *4*)-meter and multiplying (*x* + *3*)-meter by *x*-meter. Students also need to master Attributes 4, 5, and 11 to use this method. Note that one item

(Item M052503A) was dropped when constructing the Q-matrix, because Item M052503A and Item M052503B were identical in the original 32-item list. Thus, only 31 items were used to create our Q-matrix (see Table 3).

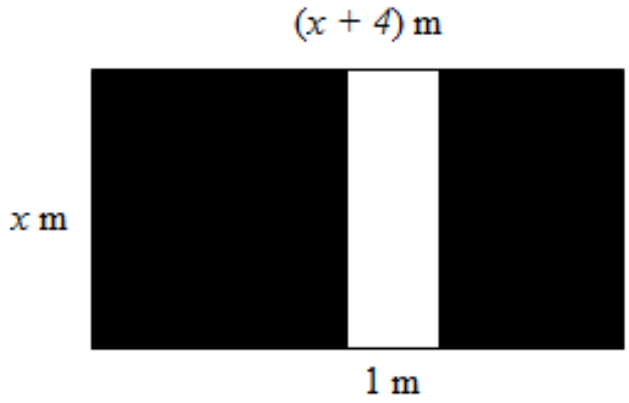


Figure 1. Item 6 (M052173) From the TIMSS 2011 Fourth Grade Mathematics

Table 2. Attributes Adopted from the Common Core State Standards Initiative (2010)

Content domain	Attribute description	Frequency
Numbers	A1-Possesses understanding of fraction equivalence and ordering; uses equivalent fractions as a strategy to add and subtract fractions.	5
	A2-Understands decimal notation for fractions, and compares decimal fractions; performs operations with decimals.	5
	A3-Understands ratio concepts, and uses ratio reasoning to solve problems; finds a percent of a quantity as a rate per 100.	4
Algebra	A4-Applies and extends previous understandings of arithmetic to algebraic expressions; solves real-life and mathematical problems using numerical and algebraic expressions and equations.	8
	A5-Reasons about and solves one-variable equations and inequalities; uses properties of operations to generate equivalent expressions.	4
	A6-Analyzes and solves linear equations and pairs of simultaneous linear equations.	1
	A7-Uses the four operations with whole numbers to solve problems; identifies and explains patterns in arithmetic.	3
Geometry	A8-Draws, constructs, and describes geometrical figures, and describes the relationships between them.	6
	A9-Solves real-life and mathematical problems involving angle measure, area, surface area, and volume.	5
	A10-Understands congruence and similarity using physical models, transparencies, or geometry software.	3
	A11-Recognizes perimeter, understands concepts of area, and relates area to multiplication and addition.	2
Data and Chance	A12-Represents and interprets data; draws informal comparative inferences about two populations.	3
	A13-Investigates chance processes and develops, uses, and evaluates probability models.	4

Data Analysis

As outlined in the TIMSS 2011 assessment framework, the TIMSS items were assessed using a three-parameter logistic item response theory (3PL IRT) model. This comparative study attempted to analyze TIMSS data sets for Korea and Turkey using a DINA model in order to present an application of a CDA-based analysis. As de la Torre and Lee (2008) showed, the results of the DINA model are consistent with that of the IRT models for the same data.

Table 3. Q-Matrix for the Eighth Grade TIMSS Mathematics Test

		Attributes												
Item	Item ID	1	2	3	4	5	6	7	8	9	10	11	12	13
1	M052216	1	1	0	0	0	0	0	0	0	0	0	0	0
2	M052231	0	1	0	0	0	0	0	0	0	0	0	0	0
3	M052061	0	0	0	1	0	0	1	0	0	0	0	0	0
4	M052228	1	0	0	0	0	0	0	0	0	0	0	0	0
5	M052214	1	1	1	0	0	0	0	0	0	0	0	0	0
6	M052173	0	0	0	1	1	0	0	0	0	0	1	0	0
7	M052302	0	0	0	1	0	0	0	0	0	0	0	0	0
8	M052002	0	0	0	1	1	0	0	0	0	0	0	0	0
9	M052362	0	0	0	0	0	0	0	1	1	1	0	0	0
10	M052408	0	0	0	0	0	0	0	1	1	1	0	0	0
11	M052084	0	0	0	0	0	0	0	0	1	0	1	0	0
12	M052206	0	0	0	0	0	0	0	1	1	0	0	0	0
13	M052429	0	0	0	0	0	0	0	0	0	0	0	0	1
14	M052503B	0	0	0	0	0	0	0	0	0	0	0	1	0
15	M042032	1	1	0	0	0	0	0	0	0	0	0	0	0
16	M042031	1	1	0	0	0	0	0	0	0	0	0	0	0
17	M042186	0	0	0	0	0	0	1	0	0	0	0	0	0
18	M042059	0	0	1	0	0	0	0	0	0	0	0	0	1
19	M042236	0	0	0	1	1	0	0	0	0	0	0	0	0
20	M042226	0	0	0	1	0	0	0	0	0	0	0	0	0
21	M042103	0	0	0	1	1	0	0	0	0	0	0	0	0
22	M042086	0	0	0	1	0	0	0	0	0	0	0	0	0
23	M042228	0	0	0	0	0	0	1	0	0	0	0	0	0
24	M042245	0	0	0	0	0	1	0	0	0	0	0	0	0
25	M042270	0	0	0	0	0	0	0	1	0	0	0	0	0
26	M042201	0	0	0	0	0	0	0	1	1	0	0	0	0
27	M042152	0	0	0	0	0	0	0	1	0	1	0	0	0
28	M042269	0	0	0	0	0	0	0	0	0	0	0	1	0
29	M042179	0	0	0	0	0	0	0	0	0	0	0	0	1
30	M042177	0	0	1	0	0	0	0	0	0	0	0	0	1
31	M042207	0	0	1	0	0	0	0	0	0	0	0	1	0

In addition to responses from Korean and Turkish students to the TIMSS eight grade mathematics assessment, the attributes (see Table 2) and Q-matrix (see Table 3) were also inputted into a DINA model. Since the TIMSS mathematics items included multiple choice and constructed responses, we dichotomized (0 = wrong answer, 1 = correct answer) those items for use with the dichotomous DINA model in this study. The DINA model parameters were estimated using maximum likelihood estimation with an expectation-maximization (EM) algorithm. All analyses were conducted using an object-oriented software package called OxEdit (Doornik, 2003) in order to obtain DINA model estimations using expectation-maximization (EM) algorithm. This program was chosen for analyses because it was a free software unlike other commercial software packages. The codes for the DINA model were requested from de la Torre (personal communication, February, 2014). The results of the two countries were compared in order to identify the weaknesses and strengths of the students of each country. Item parameter estimates and attribute mastery prevalence estimates are presented in the Results section. In addition, 3PL IRT model estimations were obtained using maximum likelihood estimation method for comparison purpose.

RESULTS

As presented above, the DINA model provides one slipping and one guessing parameter per item. These two parameters are equal across attributes. The DINA-based discrimination index (de la Torre, 2008) can also be calculated using slipping and guessing parameters for each item (i.e., $\delta = 1 - g - s$). The item discrimination index refers to the probability of correctly solving an item without the effect of guessing and slipping parameters. Put differently, it is the difference in probabilities of a correct response between $\xi = 0$ and $\xi = 1$. Slipping, guessing and discrimination parameter estimates for Korean and Turkish samples are presented in Table 6. Sixty-two item parameter estimates (31

guessing and 31 slipping parameters) were obtained for each sample. In addition to item parameters, in total 2^{13} = 8,192 attribute profile parameters were estimated for the 13 attributes listed in Table 2. Fit statistics for DINA model analyses are presented in Table 4. Since IEA (The International Association for the Evaluation of Educational Achievement) used 3PL IRT model for TIMSS analyses, results of 3PL IRT model were also provided for two samples before presenting main DINA model results (see Table 5).

Table 4. Fit Statistics for DINA Model Analyses

Country	Log-Likelihood	AIC	BIC
Korea	-4201.82	24909.65	57163.05
Turkey	-7552.31	31610.63	66193.30

Note. AIC = Akaike's Information Criterion; BIC = Bayesian Information Criterion.

Item Parameters

Table 6 presents item parameter estimates for slipping, guessing and the discrimination index for both countries. The small slipping and guessing parameter estimates indicate that examinees who master the measured attributes are able to apply the attributes correctly. As shown in Table 6, Items 4, 24, and 25 (the three with the lowest guessing and slipping parameter estimates) are the most informative items for Korean and Turkish samples. For example, for a Korean respondent who mastered Attribute 1, there is less than a 1% chance (s_4 = .009) that Item 4 is answered incorrectly. In contrast, a respondent who has not mastered Attribute 1 has no chance (g_4 = .000) of answering this item correctly. On the other hand, a Korean student has a 93% chance of answering Item 15 correctly even if he/she lacks at least one of two attributes (i.e., Attribute 1 or Attribute 2). It is desirable for a DINA model to have small guessing and slipping parameter estimates for a good model-data fit (Rupp et al., 2010). Higher values of item guessing and slipping parameters could be an indication of item-specific model misfit (Rupp & Templin, 2008b). DINA model item parameter estimates with high guessing values can be an indication item-specific misfit for Items 1, 7, 9, 15, 19, 29 and 31 in Korean data set while high slipping parameter estimates indicates possible misfits for Items 12, 14 and 21 in Turkey data set.

The mean values for item guessing, slipping parameters and the discrimination index are presented in the last row of Table 4. As can be seen in Table 6, Korean students had higher guessing parameter estimates and lower slipping parameter estimates than Turkish students for most of the items. The mean item discrimination index for the Korean sample was lower ($\bar{\delta}$ = .525) than that ($\bar{\delta}$ = .619) for the Turkish sample (see Table 6). Both samples had high discrimination indices for most of the items. A high discrimination index indicates a greater difference of probabilities of correct responses between $\xi = 0$ and $\xi = 1$. For most of the items, the item discrimination index was lower for the Korean sample than for the Turkish sample due to the higher guessing parameter estimates for the Korean sample. Among the 31 items, Items 1 (requires Attributes 1 and 2; Numbers), and 15 (requires Attributes 1 and 2; Numbers) had the lowest discrimination indices for the Korean sample due to their high guessing and low slipping parameter estimates. It should be noted that the item discrimination index for Item 24 was found to be very high (.999) for both the Korean and Turkish samples, indicating that Item 24 was very informative. This item appeared to discriminate probabilities of correct responses between $\xi = 0$ and $\xi = 1$ very well.

Attribute Probability and Attribute Prevalence

In addition to item parameter estimates, the DINA model provides respondent parameters estimates (attribute probability and attribute prevalence). Attribute probability assigns respondents to any of the C (2^A where A denotes the number of attributes) latent classes. As mentioned above, 8,192 classes exist for 13 attributes in our TIMSS example. The attribute prevalence estimate is obtained by summing the probabilities across all latent classes requiring that specific attribute. Attribute prevalence estimates are presented in Table 7 for the Korean and Turkish samples. For all of the 13 attributes, the

Korean sample had a higher attribute prevalence than the Turkish sample (see Table 7). These results indicate that Korean students are more likely to master all of the attributes. The probability of Turkish students mastering some attributes is also high (e.g., Attributes 3 and 11). Attribute 6 had the lowest probability value for the Turkish sample (.320) and the Korean sample (.609). Thus, Attribute 6, analyzes and solves linear equations and pairs of simultaneous linear equations, was difficult to master by eighth grade students. Besides Attribute 6, Turkish students also had difficulty in mastering Attributes 13, 7, 4, and 1.

Table 5. 3PL IRT Model Item Parameter Estimates for the Korean and Turkish Samples

	Korea			Turkey		
	Guessing	Difficulty	Discrimination	Guessing	Difficulty	Discrimination
Item 1	0.000	-2.560	1.928	0.252	0.888	2.629
Item 2	0.000	-2.350	1.092	0.002	0.222	1.120
Item 3	0.368	-0.240	2.378	0.027	1.111	1.690
Item 4	0.214	-1.349	1.654	0.111	0.882	3.239
Item 5	0.110	-0.949	0.878	0.269	1.436	4.200
Item 6	0.066	0.199	4.025	0.101	1.603	5.450
Item 7	0.000	-2.296	1.357	0.000	-0.365	1.602
Item 8	0.000	0.265	2.131	0.006	1.414	4.495
Item 9	0.000	-2.004	1.695	0.000	1.006	1.511
Item 10	0.356	-0.851	3.366	0.040	0.677	1.895
Item 11	0.000	-1.194	1.898	0.173	0.619	2.461
Item 12	0.109	-0.375	1.901	0.045	1.458	3.580
Item 13	0.000	-1.143	1.710	0.193	0.779	4.996
Item 14	0.275	-0.922	1.531	0.000	1.396	1.056
Item 15	0.950	-0.054	6.516	0.194	0.237	2.398
Item 16	0.478	-0.767	4.488	0.259	1.065	3.458
Item 17	0.000	-0.838	1.742	0.000	0.767	1.724
Item 18	0.000	-0.744	2.426	0.032	0.850	2.228
Item 19	0.400	-0.828	4.344	0.264	0.765	3.080
Item 20	0.000	-1.068	3.072	0.000	0.694	3.590
Item 21	0.000	-0.260	2.348	0.000	1.789	2.503
Item 22	0.000	-0.640	2.401	0.000	0.805	2.613
Item 23	0.386	-0.495	1.743	0.011	0.179	1.981
Item 24	0.000	-0.355	2.531	0.151	1.190	3.012
Item 25	0.000	-1.412	2.025	0.000	0.588	1.714
Item 26	0.000	-0.906	2.549	0.085	0.854	4.351
Item 27	0.441	-0.662	1.804	0.295	2.008	2.471
Item 28	0.000	-1.761	1.137	0.351	0.877	1.177
Item 29	0.624	-1.124	2.536	0.000	-0.335	1.359
Item 30	0.147	-0.972	2.064	0.112	0.276	1.626
Item 31	0.344	-1.085	1.979	0.000	0.539	1.180

By means of examining the attribute prevalence estimates (see Table 7), it was concluded that Turkish students were particularly weak in mastering Attributes 1, 8, and 13 when compared to their Korean peers. These three attributes had the highest prevalence estimate differences for Korea and Turkey. Hence, while most of the Korean students mastered these three attributes, many Turkish students had

difficulty in mastering them. On the contrary, Attributes 3, 11, and 5 had the lowest prevalence estimate differences for Korea and Turkey. That means the probability of Turkish students' mastery of those three attributes were close enough to the probability of Korean students' mastery. However, it should be noted that these three lowest prevalence estimate differences mainly occurred because of the increments on the probability of the Turkish students' mastery on those attributes, not because of the decrements on the probability of the Korean students' mastery.

Table 6. Item Parameter Estimates for the Korean and Turkish Samples

	Korea			Turkey		
Item	Guessing	Slipping	Discrimination	Guessing	Slipping	Discrimination
1	.878	.004	.118	.306	.121	.573
2	.586	.034	.381	.144	.134	.722
3	.470	.113	.417	.115	.347	.538
4	.000	.009	.991	.000	.107	.893
5	.517	.187	.297	.269	.323	.408
6	.087	.209	.704	.086	.411	.503
7	.778	.027	.195	.441	.077	.482
8	.031	.317	.652	.011	.384	.605
9	.768	.009	.224	.151	.234	.615
10	.581	.000	.419	.203	.018	.779
11	.470	.024	.507	.190	.031	.779
12	.232	.184	.583	.044	.526	.430
13	.186	.055	.759	.172	.115	.714
14	.365	.024	.611	.079	.540	.381
15	.928	.007	.065	.442	.078	.480
16	.527	.000	.473	.265	.146	.589
17	.218	.067	.715	.072	.210	.718
18	.248	.025	.727	.113	.210	.678
19	.682	.000	.319	.350	.000	.651
20	.462	.032	.506	.021	.197	.782
21	.177	.137	.685	.006	.557	.437
22	.138	.078	.784	.030	.265	.706
23	.475	.099	.426	.266	.116	.618
24	.000	.001	.999	.000	.001	.999
25	.003	.033	.964	.007	.160	.832
26	.243	.016	.741	.103	.137	.760
27	.419	.043	.538	.262	.421	.317
28	.521	.044	.435	.340	.161	.499
29	.719	.011	.270	.443	.088	.469
30	.504	.041	.455	.357	.101	.542
31	.668	.011	.321	.172	.127	.701
Mean	.415	.059	.525	.176	.204	.619

The top five highest attribute class profiles with the highest probability estimates are presented in Table 8. These classes with highest probabilities were selected from 8,192 possible latent classes. The probability estimates listed in Table 8 can be interpreted as percentages, as the sum of probabilities for 8,192 different latent class profiles are equal to unity. For example, a probability value of .013 for a

latent class profile indicates that only 13% of the respondents were assigned to this specific latent class. As shown in Table 8, 44% of Korean students mastered all of the attributes (attribute class of 1111111111111), while only almost 13% of Turkish students mastered all of the attributes. Other latent class profiles with the highest probability values showed that Attributes 5, 6, 7, and 12 were difficult to master for Korean students (see bolded zeros in Table 8). Less than one percent (p = .0016) of Korean respondents appeared to master none of the attributes (attribute class of 0000000000000). The second largest latent class for Turkish students was the mastery of all attributes except for Attributes 5 and 6. The posterior probability of this latent class profile was .026. Therefore, Attribute 6 appeared to be difficult to master by Turkish students as it was not mastered by most of the Turkish students (see bolded zeros for Attribute 6 in Table 8). Furthermore, 1.0% of the Turkish students could not master any of the attributes, while another 1.0% of Turkish sample only mastered Attribute 3 (understands ratio concepts and uses ratio reasoning to solve problems; finds a percent of a quantity as a rate per 100).

Table 7. Estimates of Attribute Prevalence

	Attribute Prevalence	
Attribute	Korea	Turkey
1	0.866	0.392
2	0.803	0.464
3	0.753	0.611
4	0.708	0.382
5	0.728	0.522
6	0.609	0.320
7	0.702	0.361
8	0.882	0.445
9	0.768	0.543
10	0.806	0.543
11	0.768	0.581
12	0.714	0.462
13	0.790	0.354

Table 8. *Top Five Attribute Class Profiles for the Korean and Turkish Samples*

Korea		Turkey	
Attribute Profile	Probability	Attribute Profile	Probability
1111111111111	0.443	1111111111111	0.128
11111**0**1111111	0.039	1111**00**1111111	0.026
111111111110**1**	0.019	11111**0**1111111	0.017
1111**000**111111	0.012	**001**0000000000	0.010
1111**000**1111**0**1	0.013	**0000000000000**	0.010

DISCUSSION

This study showed the application of a CDA-based assessment for a large-scale test data set, which has been originally analyzed with a traditional IRT model (3PL). CDM approach was selected, because it is possible to report a more detailed evaluation of students' performances on specific skills. Korea (the top performing country) and Turkey (the focus of the study) were selected for analyses in this study to show how a DINA model can be used to obtain fine-grained information about the performances of the students from these two countries. There are several advantages of the DINA model over traditional IRT models. For example, analyses based on IRT models provide a single overall score based on invariant item and ability parameters. Unlike IRT models, CDMs (e.g., the DINA model) are used to obtain qualitative information in addition to quantitative information. The qualitative part of the CDMs comes from a latent class based structure. Using this property, it was tried to show which

skill profiles both Korean and Turkish students were assigned. This specific respondent information could be very useful for instructors and educational policy makers for demonstrating the mastery of each student on each attribute, which is important feedback for instructors. In addition to attribute masteries, a number of item parameter estimates can be estimated with DCMs, like item guessing, slipping, and discrimination parameters.

The results of the DINA model for the Korea and Turkey data sets provided different patterns for the strengths and weaknesses of the two countries. As in the original 3PL IRT analysis, the Korean sample showed a higher performance than the Turkish sample in this study. As expected, the posterior probability of mastering all of the attributes (i.e., 1111111111111) for Korean students was higher than that of Turkish students. Additionally, one percent of the Turkish sample mastered none of the attributes, while this percentage was less than one percent for the Korean sample. In addition, six percent of Turkish respondents mastered only one of the thirteen attributes. These findings were very crucial for diagnosing the most problematic attributes (or skills) for the Turkish sample. Another attribute related finding showed that attribute prevalence estimates were higher than .70 for all items except for Attribute 6 in the Korean sample. However, all of the attribute prevalence estimates were less than .70 for the Turkish sample.

As a result of examining the estimates provided in Table 7, it was decided that Turkish students had difficulties mastering Attributes 4, 6, and 7. Because these three attributes were classified in the Algebra content domain, it was suggested that Turkish educators should pay more attention to eight graders' understanding of Algebra topics. They should especially focus on students' understanding of analyzing and solving linear equations and applying previous understandings of arithmetic to algebraic expressions. This result was consistent with the findings from Dogan and Tatsuoka (2008) who also stated Turkish students' weaknesses in algebra content domain when compared to American students. Furthermore, when compared to their Korean peers, Turkish students were particularly weak in mastering Attributes 1, 8, and 13. These three attributes had the highest prevalence estimate differences for Korea and Turkey. Hence, while most of the Korean students mastered these three attributes, many Turkish students had difficulties in mastering them. The items in which the mastery of Attribute 1 was required were all fractions and decimals items. Therefore, the results indicate Turkish students' weaknesses in the fractions and decimals subject area—especially with understanding fraction equivalence and ordering—compared to their Korean peers. Similarly, the mastery of Attribute 8 was required in solving geometry items; so, compared to their Korean peers, Turkish students did not perform well on the geometry items that involved drawing, constructing, and describing geometrical figures and the relationships between them. Additionally, except for one item, the mastery of Attribute 13 was necessitated in solving data and chance items. Hence, in comparison to Korean students, as in the Dogan and Tatsuoka (2008) study, Turkish students also did not perform well on the data and chance problems that investigated the chance process and using and evaluating probability models. On the contrary, the three lowest prevalence estimate differences between Korea and Turkey were obtained for Attributes 3, 11, and 5. Thus, it can be concluded that Turkish students performed relatively well on items that involved understanding ratio concepts and using ratio reasoning; recognizing perimeter and understanding concepts of area; and reasoning about and solving one-variable equations and inequalities.

It should be noted that model-data fit and item fit statistics may have an effect on the interpretations of item parameter estimates obtained from a DINA model. More appropriate conclusions can be made based on models with better fit. It is obvious that DINA models in this study did not show perfect fit to two data sets. Assuming that we have enough model-data fit, we can make several conclusions based on DINA model results. Under this condition, item parameter estimates from the DINA model can provide feedback for students from the two countries. Apparently, Korean students were less likely to slip and were more likely to guess correct answers. However, the Turkish sample yielded lower guessing parameter estimates and higher slipping parameter estimates, indicating possible problems with content knowledge or testing strategies. Item parameter estimates can also be used for improving measurement instruments. Results of item parameter estimates showed problems with several items. For example, Items 6 and 8 yielded higher slipping parameter estimates for both samples. Both items

were classified under the algebra content domain, and Item 6 was a multiple choice item, whereas Item 8 was a constructed response item. Turkish students were also most likely to slip on Items 21, 14, 12, and 27. Considering Items 21, 14, and 12 were also constructed response items, it can be concluded that Turkish students were more likely to slip on constructed response items. Dogan and Tatsuoka (2008) also stated a similar weakness of the Turkish students' in their study. They observed that Turkish students did not perform well on the open-ended items and had difficulty constructing answers in comparison to selecting an answer from given alternatives. Therefore, the findings of this study suggest that Turkish educators and policy makers should pay more attention to teaching students how to deal with constructed response items instead of teaching test skills to solve multiple choice items. To accomplish this, teachers should encourage students through verbal and written expressions of their mathematical understandings. In addition, item discrimination indices may also be useful for identifying poor items. For instance, Items 1 and 15 had the highest guessing parameters and lowest discrimination indices for the Korean sample. Hence, these two items were not very informative and required improvements.

In sum, various factors might have affected Korean and Turkish eight-grade students' performances on the TIMSS assessment. As previously discussed, Im and Park (2010) attributed Korean students' high achievement to the social and instructional factors. In a similar vein, when compared Chinese Taipei and Turkey on the TIMSS 2007 eight-grade science items, Ozturk and Ucar (2010) found that socio-economics, parents' education level, and quality of schooling contributed to Turkish students' relatively low academic performance. In this study, our results identify situations for instructors where current curriculum may be improved to help students master some lacking attributes based on CDM-based feedback. As Leighton and Gierl (2007) stated, recent CDM studies have been applied for post-hoc analyses and item analyses rather than constructing the tests (Chapter 7). Although, our study demonstrated that retrofitting of a CDM via the DINA model can be very useful for the TIMSS assessment, it is evident that more benefit can be obtained from CDM-based analyses when tests are designed using CDMs in advance.

REFERENCES

Choi, M. K., Lee, Y-S., & Park, Y. S. (2015). What CDM can tell about what students have learned: An analysis of TIMSS eighth grade mathematics. *Eurasia Journal of Mathematics, Science & Technology Education, 11*(6), 1563–1577.

Chung, Y.L., & Son, D.H. (2000). Effects of cooperative learning strategy on achievement and science learning attitudes in middle school biology. *Journal of the Korean Association for Research in Science Education, 20,* 611–623.

Common Core State Standards Initiative (2010). *The common core state standards for mathematics*. Washington, D.C.: Author.

Common Core State Standards Initiative (2014). Myths vs. facts. Retrieved from http://www.corestandards.org/about-the-standards/myths-vs-facts/

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,* 343–362.

de la Torre, J., & Lee, Y. S. (2008, March). *Relationships between cognitive diagnosis, CTT and IRT indices: An empirical investigation*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34,* 115–130.

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

DiBello, L., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. V. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26, *Psychometrics*) (pp. 979–1027). Amsterdam: Elsevier.

Dogan, E., & Tatsuoka, K. (2008). An international comparison using a diagnostic testing model: Turkish students' profile of mathematical skills on TIMSS-R. *Educational Studies in Mathematics, 68*(3), 263–272.

Doornik, J. A. (2003). *Object-oriented matrix programming using Ox (version 3.1)* [Computer software]. London: Timberlake Consultants Press.

Fisher, D. L., & Kim, H. B. (1999, April). *Constructivist learning environments in science classes in Korea.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Fu, J., & Li, Y. (2007, April). *An integrated review of cognitively diagnostic psychometric models.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301–323.

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. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191–210.

Holliday, W. G., & Holliday, B. W. (2003). Why using international comparative math and science achievement data from TIMSS is not helpful. *The Education Forum, 67*, 250–257.

House, J.D. (2009). Classroom instructional strategies and science career interest for adolescent students in Korea: Results from the TIMSS 2003 assessment. *Journal of Instructional Psychology, 36*, 13–19.

Im, S., & Park, H. J. (2010). A comparison of US and Korean students' mathematics skills using a cognitive diagnostic testing method: linkage to instruction. *Educational Research and Evaluation, 16*(3), 287–301.

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

Kim, K., Kim, S., Kim, N., Park, S., Kim, J., Park, H., & Jung, S. (2008). *Characteristics of achievement trend in Korea's middle and high school students from International Achievement Assessment (TIMSS/PISA) (KICE Research report, RRE-2008-3-1).* Seoul, Korea: Korea Institute of Curriculum and Evaluation.

Lee, Y.-S., Park, Y.S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing, 11*, 144–177.

Lee, Y.-S., Johnson, M., Park, J. Y., Sachdeva, R., Zhang, J., & Waldman, M. (2013, April). *A multidimensional scaling (mds) approach for investigating students' cognitive weakness and strength on the TIMSS 2007 mathematics assessment.* Paper presented at the 2013 Annual Meeting of the American Educational Research Association Conference in San Francisco, CA.

Leighton, J., & Gierl, M. (2007). Why cognitive diagnostic assessment? In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education* (pp. 3–18). Cambridge: Cambridge University Press.

Mullis, I.V.S., Martin, M.O., Foy, P., & Arora, A. (2012). *TIMSS 2011 International Results in Mathematics.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.). (2012). *Cognitively diagnostic assessment.* Hillsdale, NJ: Erlbaum.

Oh, S. (2005). Discursive roles of the teacher during class sessions for students presenting their science investigation. *International Journal of Science Education, 27*, 1825–1851.

Ozturk, D., & Ucar, S. (2010). By using TIMSS data, determination and comparison of the factors that affects science achievement of 8 grade students from Taiwan and Turkey. *Cukurova University Journal of Social Sciences, 19*(3), 241–256.

Park, K.M. (2004, July). *Mathematics teacher education in East Asian countries: From the perspective of pedagogical content knowledge.* Paper presented at the 10th International Congress on Mathematical Education, Copenhagen, Denmark.

Roussos, L., DiBello, L. V., Stout, W., Hartz, S., Henson, R. A., & Templin, J. H. (2007). The fusion model skills diagnosis system. In J. P. Leighton, & Gierl, M. J. (Ed.), *Cognitively diagnostic assessment for education: Theory and practice.* (pp. 275–318). Thousand Oaks, CA: SAGE.

Rupp, A. A., & Templin, J. L. (2008a). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*(4), 219–262.

Rupp, A. A., & Templin, J. (2008b). 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. L., & Henson, R. A. (2010). *Diagnostic assessment: Theory, methods, and applications.* New York: Guilford Press.

Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics, 12*, 55–73.

Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305.

Toker, T., & Green, K. (2012, April). *An application of cognitive diagnostic assessment on TIMMS-2007 8th Grade Mathematics items.* Paper presented at the annual meeting of the American Educational Research Association, Vancouver, British Columbia, Canada.

von Davier, M. (2005). *A general diagnostic model applied to language testing data*. ETS Research Report RR-05-16.

Wang, J. (2001). TIMSS primary and middle school data: Some technical concerns. *Educational Researcher, 30,* 17–21.

GENİŞ ÖZET

Giriş

TIMSS (Trends in International Mathematics and Science Study) sınavı 1995 yılındaki ilk uygulamasından beri 4. ve 8. sınıf fen bilgisi ve matematik derslerinde katılımcı ülkelerin kendi öğrencilerinin performanslarını diğer katılımcı ülke öğrencilerinin performanslarıyla karşılaştırmalarına yardımcı olmuştur. Her ne kadar TIMSS tarzındaki uluslararası büyük ölçekli sınavların bu tür karşılaştırmalar için kullanılması eleştiriliyor olsa da (Holliday ve Holliday, 2003; Wang, 2001), katılımcı ülkeler bu sınavlardan alınan sonuçlara göre kendi öğretim sistemlerinde ve müfredatlarında düzenlemelere gitmişlerdir. Genel olarak bakıldığında bu büyük ölçekli uluslararası sınavlar toplam skora dayalı bir değerlendirme sistemi içermekte ve her ülkenin öğrencilerine toplam puanlar atayarak ülkeler bazında elde edilen ortalama puanlara göre ülkelerin kendi yerleri hakkında karşılaştırma yapmalarına imkan sağlamaktadır. Tek bir puana dayalı değerlendirme yaklaşımları öğrenci performansları açısından çok detaylı bilgi sunmadığı gerekçesiyle eleştirilmiş (Nichols, 2012; Leighton ve Gierl, 2007) ve bunların yerine daha detaylı değerlendirmeye olanak sağlayan bilişsel tanı modelleri geliştirilmiştir (Rupp, Templin, ve Henson, 2010). Bilişsel tanı modellerine ait detaylar Rupp, Templin ve Henson (2010) ve DiBello, Roussos ve Stout (2007) çalışmalarında bulunabilir. Bu bilişsel tanı modellerinden en yaygın olarak kullanılanlardan bir tanesi olan DINA (*deterministic inputs, noisy "and" gate,* Haertel, 1989; de la Torre, 2009; Junker ve Sijtsma, 2001) modeli bu çalışmada kullanılmıştır. Temel olarak DINA modeli bir maddenin doğru cevaplanabilmesi için o madde için gerekli olan özellikler neler ise cevaplayıcının bu özelliklerde yeterlilik kazanmasını şart koşar. Her madde için madde kayması (item slipping) ve madde tahmini (item guessing) olmak üzere iki parametre sonucu elde etmemizi sağlar.

Son zamanlarda çeşitli çalışmalar bilişsel tanı modellerini kullanarak öğrencilerin TIMSS, PISA, ve PIRLS gibi uluslararası büyük ölçekli sınavlardaki başarılarını karşılaştırmışlardır. Bu çalışmalar öğrencilerin performansları ve becerileri, öğretmenlerin öğretim yöntemleri ve öğrencilerin performansları arasındaki ilişki, ve katılımcı ülkelerin eğitim sistemleri ile müfredatları hakkında çok kullanışlı bilgi edinme imkanı sağlamıştır. Örneğin, Dogan ve Tatsuoka (2008) Türk ve Amerikan sekizinci sınıf öğrencilerinin TIMMS-R 1999 sınavındaki matematik performanslarını karşılaştırmışlardır. Bu çalışmaya göre Türk öğrencilerin cebir ve olasılık/istatistik gibi sınavlarda Amerikan öğrencilere göre daha düşük performans sergiledikleri ortaya çıkmıştır. Benzer şekilde, Im ve Park (2010) Güney Kore ve Amerikan sekizinci sınıf öğrencilerinin TIMMS 2003 sınavındaki matematik performanslarını karşılaştırmışlardır. Çalışmanın bulguları Güney Kore ve Amerikan öğrencilerinin performansları arasında çok önemli farklılıklar olduğunu göz önüne çıkarmıştır. Bu farklılıklar özellikle problemlerin yeniden yapılandırılması ve akıl yürütme ile ölçme ve geometri konularında önemli değişiklikler göstermiştir.

Bu çalışmanın asıl amacı TIMSS 2011 matematik sınavındaki konular açısından sekizinci sınıf Türk öğrencilerinin güçlü ve zayıf yanlarını incelemektir. Bu amacı gerçekleştirmek için bu çalışmada Türk ve Güney Kore'li öğrencilerin göreceli performansları karşılaştırılmıştır. Güney Kore'li öğrencilerin düzenli olarak TIMSS matematik sınavında ilk üç sırada yer almaları Kore'yi referans ülke olarak almamızda temel neden olmuştur. Bu doğrultuda bu çalışma tanılayıcı değerlendirme yaklaşımını kullanarak aşağıdaki iki araştırma sorusunu cevaplamaya çalışmaktadır:

1) Türk ve Kore'li sekizinci sınıf öğrencilerinin TIMSS 2011 matematik performansları göreceli olarak nasıl farklılıklar göstermektedir?

2) Kore'li öğrencilerle karşılaştırdığında, Türk öğrencilerinin TIMSS 2011'deki matematik konularında güçlü ve zayıf yanları nelerdir?

Yöntem

Bu çalışmada Türkiye ve Güney Kore ülkelerine ait sekizinci sınıf TIMSS 2011 matematik veri setleri kullanılmıştır. Seçilen örneklemlerde 368 Güney Kore'li ve 488 Türkiye'li öğrenci bulunmaktadır. TIMSS 2011'de uygulanan 14 test kitapçığından 2 numaralı kitapçık bu çalışmadaki analizler için seçilmiştir. Kitapçık 2'de 15 çoktan seçmeli ve 17 kısa yanıtlı madde bulunmaktadır. Seçilen bu 2 numaralı kitapçık bilişsel tanı modellerinden olan DINA model kullanılarak analiz edilmiştir. İki kategorili DINA model kullanıldığı için çoktan seçmeli test maddeleri ve kısa yanıtlı maddeler 0 (yanlış cevap) ve 1 (doğru cevap) şeklinde kodlanmıştır. DINA model analizleri için gerekli olan Q-matris maddeleri doğru cevaplamak için gerekli olan özellikler göz önünde bulundurularak dört matematik eğitimcisi tarafından bağımsız bir şekilde kodlanmıştır. Bu dört eğitimcinin görüşlerine göre toplamda 13 tane özellik Common Core State Standards for Mathematics (CCSSM; Common Core State Standards Initiative, 2010) müfredatı kullanılarak oluşturulmuştur. Q-matris ve öğrenci cevaplarının 1-0 şeklinde kodlanmış olduğu veri setleri kullanılarak DINA model analizleri yapılmıştır. İki ülkeye ait veriler OxEdit programı kullanılarak maksimum olabilirlik yöntemi vasıtasıyla analiz edilmiştir. TIMSS 2011'in uygulayıcı kurum (IEA) tarafından üç parametreli madde tepki kuramı (MTK) ile analiz edilmiş olmasından dolayı üç parametreli MTK modelinden elde edilen iki ülkeye ait sonuçlar da karşılaştırma amaçlı sunulmuştur.

Sonuç ve Tartışma

DINA modeli kullanılarak analiz edilen iki ülke veri setine ait 31 madde için elde ettiğimiz madde parametreleri madde kayma (slipping), madde tahmin (guessing) ve bu iki parametre kullanılarak hesaplanan madde ayırt ediciliği (discrimination) değerleri şeklinde ayrı ayrı rapor edilmiştir. Maddeleri çözmek için gerekli olan özelliklere ait olarak da özellik yaygınlığı (attribute prevalence) yüzdelikler şeklinde sunulmuştur. Türk öğrenciler hem madde parametreleri hem de özellik parametreleri açısından Kore'li öğrencilerden farklılık göstermişlerdir. Genel olarak Kore verisinden elde edilen madde parametreleri yüksek tahmin (guessing) ve düşük kayma (slipping) değerleri içerirken bu durum Türk öğrenciler için tam tersi olarak gözlenmiştir. Koreli öğrenciler testteki maddeleri çözmek için gerekli olan özelliklerin hemen hemen hepsinde yeterlilik kazanmışken Türk öğrenciler çoğu özellikte yeterlilik kazanamamakla beraber en çok Özellik 1, 8 ve 13'te düşük yeterlilik göstermişlerdir.

Dogan ve Tatsuoka (2008) çalışmasına benzer olarak, Türk öğrenciler Cebir, Data Analizi ve Şans konularında Kore'li öğrencilere göre düşük performans sergilemişlerdir. Bu konuların yanında Türk öğrenciler ayrıca Geometri konusunda da Kore'li öğrencilere göre daha az başarılı olmuşlardır. Yine Dogan ve Tatsuoka (2008) çalışmasına benzer olarak Türk öğrenciler açık-uçlu sorularda yeterli başarıyı gösterememişlerdir. Türk öğrencilerin açık-uçlu soruları cevaplamaktaki yetersizliği Türkiye'nin çoktan seçmeli testler üzerine dayalı olan eğitim sisteminin bir neticesi olarak yorumlanabilir. Test sisteminin yanı sıra, Ozturk ve Ucar (2010)'ın da bahsettiği üzere Türk öğrencilerinin düşük performanslarında sosyo-ekonomik nedenler ile, ailelerin eğitim durumları, ve okullardaki öğretimin kalitesi gibi faktörler de etkili olmuş olabilir. Benzer şekilde, Im ve Park (2010) Güney Kore'li öğrencilerin yüksek başarısının sosyal faktörler ve öğretim ile ilgili faktörlere bağlı olduğunu belirtmiştir. Bu çalışma, içerinde sunulan bulguların Türk eğitimcilerine matematik müfredatının nasıl geliştirebileceğine dair bilişsel tanı modeline dayalı geri dönütler vermesi açısından kayda değerdir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2015; 6(2); 254-267

Kayıp Veri Sorununun Çözümünde Kullanılan Farklı Yöntemlerin Ölçeklerin Geçerlik ve Güvenirliği Bağlamında Karşılaştırılması*

Comparison of Various Methods Used in Solving Missing Data Problems in the context of Validity and Reliability of the Scales

Merve ŞAHİN KÜRŞAD ** Zekeriya NARTGÜN **

Öz
Bu araştırmanın amacı kayıp veri sorununun çözümünde kullanılan farklı yöntemlerin etkililiğini ölçeklerin geçerliği ve güvenirliği bağlamında karşılaştırmaktır. Bu amaçla, PISA 2012 Türkiye örneklemi ve "Matematik Çalışma Etiği" ölçeğinden yararlanılmıştır. Analizler için Türkiye örnekleminden rastgele 200 kişilik tam veri seti çekilmiştir. Tam veri setinden, tamamıyla rassal olarak kayıp (TROK) mekanizması altında, farklı oranlarda veri silme ve bu verileri farklı kayıp veri yöntemleriyle yeni tam veri setlerine dönüştürme işlemlerinden sonra geçerlik ve güvenirliğe ilişkin analizler gerçekleştirilmiştir. Kayıp veri içeren setlerin yeni tam veri setlerine dönüştürülmesinde seri ortalaması, yakın noktaların ortalaması, yakın noktaların medyanı, doğrusal değer kestirimi, noktanın doğrusal eğimi, liste bazında silme, beklenti maksimizasyonu, regresyon ataması ve çoklu atama kayıp veri yöntemleri kullanılmıştır. Yeni tam veri setlerinden geçerlik ve güvenirliğe ilişkin elde edilen değerleri karşılaştırarak yorumlamada tam veri setinden elde edilen değerler referans değerler olarak kullanılmıştır. Araştırma sonuçlarına göre liste bazında silme yöntemi için elde edilen değerler, genel olarak tam veri setinden elde edilen değerlere en az benzerlik gösteren değerler olmuştur. Yaklaşık değer atama yöntemleri için elde edilen değerler kayıp veri oranının düşük olduğu durumlarda genel olarak tam veri setinden elde edilen değerlere yakın veya aynı değerleri verirken, tüm kayıp veri oranları için tam veri setinden elde edilen değerlere en yakın değer veren yöntemler çoklu atama, beklenti maksimizasyonu ve regresyon ataması yöntemleri olmuştur.

Anahtar Kelimeler: Kayıp veri, geçerlik, güvenirlik

Abstract
The purpose of this research is to compare the effectiveness of various methods used in solving missing data problems in the context of validity and reliability of the scales. For this purpose PISA 2012 Turkey sample and "Math Work Ethics" scale was used. For the analysis, complete data set of 200 persons were chosen from the Turkey sample at random. After the process of data deletion at different rates from complete data set and transforming them into new complete data sets, under *missing completely* at random (MCAR) mechanism, analysis of validity and reliability were realized. During the phase of transforming missing data set into new complete data sets, series mean, mean of nearby points, median of nearby points, linear interpolation, linear trend at point, listwise deletion, expectation maximization, regression imputation and multiple imputation methods were used. The values obtained from the complete data set were used as reference values in interpreting the values by comparing the values of validity and reliability at the new complete data sets. The research results reveal that the values obtained for the listwise deletion at different rates of missing data are the values with the least similarity to the ones generally obtained from the complete data set. While the values obtained for the approximate value imputation methods resulted in proximal or same values as the ones generally obtained from

*Bu çalışma, birinci yazarın Doç. Dr. Zekeriya Nartgün danışmanlığında tamamlanan yüksek lisans tezinden türetilmiştir.
** Arş. Gör. Abant İzzet Baysal Üniversitesi, Eğitim Fakültesi, Bolu-Türkiye, sahinmerv@gmail.com
*** Doç. Dr. Abant İzzet Baysal Üniversitesi, Eğitim Fakültesi, Bolu-Türkiye, nartgun@yahoo.com

Geliş Tarihi: 13.07.2015
Kabul Tarihi: 09.11.2015

the complete data set in cases where the missing data rate is low multiple imputation, expectation maximization and regression imputation methods resulted in close proximal values at all missing data rates as the ones obtained from the complete data set.

Key Words: Missing data, validity, reliability

GİRİŞ

Yapılan birçok araştırmada, toplanan verilerde eksiklerle karşılaşılmaktadır. Verilerde bulunan bu eksikler kayıp veri veya kayıp değer olarak adlandırılır. Bu durum istatistiksel analizlerde çoğu araştırmacının karşılaştığı önemli bir sorundur (Çokluk ve Kayri, 2011). Örneğin, bilindiği üzere, psikolojik değerlendirmenin temelinde gözlenen verilerden hareketle örtük değişkenler hakkında çıkarımlarda bulunmak vardır. Ancak gözlenen verilerde bulunacak eksikler, örtük değişkenler hakkında çıkarımda bulunulmasını da zorlaştıracaktır (Hohensinn ve Kubinger, 2011). Bundan dolayı araştırmacılardan kayıp veri sorununu ya baştan engellemeleri ya da bu sorunla karşılaştıktan sonra olabildiğince düzeltmeleri beklenir (McKnight, McKnight, Sidani ve Figueredo, 2007).

Kayıp veri sorunuyla en çok sosyal bilimler (Vansteelandt, Carpenter ve Kenward, 2010) ve kişisel davranışların ölçüldüğü alanlarda karşılaşılmaktadır (Ginkel, Sijtsma, Van der Ark ve Vermunt, 2010). Uzun anketlerde soruların atlanması, deneysel çalışmalarda verilerin kaydedilmemesi gibi mekanik hatalar, seksüel davranışlar vb. hassas konularda araştırma yapma (Field, 2009), soruların dikkatsizlik sonucu veya cevabı bilinmediği için boş bırakılması (Finch ve Margraf, 2008) gibi sebepler kayıp veri sorununun başlıca sebepleri arasındadır. Brown ve Kros (2003) bu sebepleri 3 grupta sınıflandırmıştır. Bunlardan birincisi süreçle ilgili nedenlerdir. Süreçle ilgili nedenler veri girişi sırasında karşılaşılan sorunları kapsamaktadır. Bu sorunlar verilerin girilmemesi ve veri gruplaması sırasında karşılaşılan problem durumlarını içerir. Bir diğer neden ise cevaplamayı reddetme ile ilgili nedenlerdir. Böyle durumlar, bazı soruların ağır veya hassas konularla ilgili olmasından kaynaklanmaktadır. Üçüncü neden ise herhangi bir durum, grup veya konuyla ilgisi olmayan soruların sorulmasıdır, yani soruların özel bir gruba sorulması yerine daha geniş bir kitleye sorulmasından kaynaklanan kayıp veri durumudur. Örneğin evli olmayan bir gruba kaç yıllık evlisiniz gibi bir soru sorulması kayıp veriye neden olacaktır.

Araştırmalarda yer alan kayıp veriler, bilgi eksikliğini temsil eder, dolaylısıyla da bilgi kaybına neden olurlar (Bal, 2003). Araştırmalarda, katılımcılar tüm soruları cevaplandırmaları için bilgilendirilse bile çalışmalarda kayıp veri durumu ile karşılaşılmaktadır (Ginkel, Van der Ark, Sjitsma ve Vermunt, 2007). İstatistiksel analizler için gerekli olan paket programlar tam veri setlerine göre düzenlendiği için, toplanan verilerde bulunan kayıplar istatistiksel analizlerde önemli sorunlara neden olmaktadır (Bal, 2003). Kayıp verilerden kaynaklanan problemlerin yaşanmaması için araştırmacıların tam veri setleri ile çalışması gerekmektedir.

McKnight ve diğerlerine (2007) göre kayıp veriler çalışma sonuçlarını şu şekilde etkilemektedir. Eğer kayıp veri miktarı fazlaysa elde edilen sonuçların güvenirliği, genellenebilirliği ve yapılacak istatistiksel çıkarımlar bundan önemli derecede etkilenecektir. Bunun sonucunda da yapılan istatistiksel çıkarımlar yanıltıcı olacaktır. Ayrıca kayıp veriler çalışmanın geçerliğini de olumsuz yönde etkileyecektir.

İlgili alan yazın yukarıda belirtilen hususların yanı sıra kayıp verilerin ölçme sonuçlarının ortalaması, standart sapması, basıklık ve çarpıklık değerleri gibi istatistikleri de etkilediğini göstermektedir (Bal, 2003; Demir, 2013). Alan yazında kayıp verilere herhangi bir işlem uygulamadan yapılan betimsel istatistik, güvenirlik ve geçerlik kestirimlerinin kayıp verinin niteliğine ve miktarına bağlı olarak yanlılık ve kestirim hataları üretebileceği belirtilmektedir (Demir ve Parlak, 2012).

Kayıp veri ile karşılaşılan durumlarda araştırmacılar genellikle kayıp veri içeren durumları analiz dışı bırakmayı tercih etmektedirler. Ancak bunun yapılabilmesi için öncelikle kayıp veriye neyin neden olduğunun anlaşılması gerekmektedir (Demir ve Parlak, 2012). Çünkü kayıp veriler farklı nedenlerle ve farklı örüntülerle ortaya çıkmaktadır. Ayrıca, kayıp verilerin tamamen rassal olarak kayıp, rassal olarak kayıp veya rassal olmayan kayıp veri mekanizmalarından hangisi ile ilişkili olduğunun

belirlenmesi, kayıp veri sorunuyla baş etmede hangi kayıp veri yönteminin kullanılacağını belirlemek açısından da önemlidir (Allison, 2003).

Araştırmalarda karşılaşılan kayıp veri sorunun çözümünde kullanılmak üzere geçmişten günümüze farklı yöntemler geliştirilmiştir. Kayıp veri ile analize devam etme, eksik gözlemleri analiz dışı bırakma, eksik gözlemler yerine veri atama veya çeşitli istatistiksel yöntemlerle eksik verileri tamamlama gibi yöntemler kayıp verilerle karşılaşıldığı durumlarda kullanılan yöntemlerden bazılarıdır. (Bal, 2003; Carpita ve Manisera, 2011; Duncan, Duncan ve Li, 1998; Downey ve King, 1998; Little, 1988). Bu yöntemler içerisinde araştırmacılar tarafından en çok kullanılan yöntemler liste bazında silme ve çiftler bazında silme gibi eksik verileri analiz dışı bırakma yöntemleridir. Ancak yapılan çalışmalar bu yöntemlerin örneklemde kayba, güvenirlikte azalmaya, tahminlerde yanlılığa neden olduğunu (Allison, 2009; Cumming, 2013; Satıcı ve Kadılar, 2009; Oğuzlar, 2001; Van Der Ark ve Vermunt,2010) ve yanlılıktan kaynaklı olarak da örneklemin evreni temsil etme derecesinin düştüğünü göstermektedir (Demir ve Parlak, 2012, Little, 1988). Belirtilen bu sebeplerden dolayı, son yıllarda, bu yöntemler yerine, beklenti maksimizasyonu ve çoklu atama gibi "modern" yöntemler önerilmektedir. Çünkü bu yöntemler, silme yöntemleri gibi geleneksel kayıp veri yöntemlerinin aksine, yanlılığın azaltılması, etkili parametre tahminlerinin yapılması ve daha büyük istatistiksel gücün sağlanması hususunda daha etkili sonuçlar vermektedir (Enders, 2013).

Bu yöntemlerden bu araştırma kapsamında ele alınanlara ilişkin açıklamalar kısaca özetlenerek aşağıda verilmiştir.

a. Silme Yöntemleri:

Liste bazında silme (LBS): Yöntemde bir ya da daha fazla kayıp veri içeren bireyler veya durumlar listeden çıkartılarak sadece tam veri içeren durumlar kullanılır (Cheema, 2012; Yılmaz, 2014)

b. Yaklaşık değer atama yöntemleri:

Seri ortalaması (SO): Tüm deneklerin belirli bir değişkene ilişkin ortalaması atanır.

Yakın noktaların ortalaması (YNO): Kayıp verinin yakınındaki değerlerin ortalaması alınarak kayıp veri yerine atama yapılır.

Yakın noktaların medyanı (YNM): Kayıp verinin yakınındaki tam verilerin medyanı alınarak kayıp veri yerine atama yapılır.

Doğrusal değer kestirimi (DDK): Eksik veriden önceki son tam gözlem ile eksik veriden sonraki ilk tam gözlem değeri eksik veri yerine atanır.

Noktanın doğrusal eğimi (NDE): Kayıp veri yerine mevcut yapının sahip olduğu eğilim ile uyumlu olarak bir değer atanır (Çokluk ve Kayrı, 2011).

c. Beklenti maksimizasyonu (BM): Kayıp verileri, en çok olabilirlik kestirimleri ile dolduran iki aşamalı yöntemdir. İlk aşama olan beklenti aşamasında, kayıp veriler beklenen değerlerle tamamlanır. İkinci aşama olan maksimizasyon aşamasında ise, ilk aşamada tahmin edilen değerler kullanılarak parametre tahmini yapılır (Enders, 2001)

d. Regresyon ataması (RA): Tam veriler kullanılarak, regresyon modeli elde edilir ve kayıp veriler yerine atama yapılır (Yılmaz, 2014).

e. Çoklu atama (ÇA): Kayıp veri yerine m tane atamanın yapıldığı tekniktir. Atama sayısı genelde 3-10 arasında değişmektedir ve yöntem atfetme, analiz etme ve bir araya getirme adımlarından oluşmaktadır (Oğuzlar, 2001).

Kayıp veriler için birçok farklı yöntem ve her yöntemin kayıp veriyi tahmin etme konusunda kendine göre farklı etkileri vardır. Kayıp veriler çalışma sonuçlarını farklı şekillerde etkilediği için, araştırmacıların örneklem büyüklüğü, kayıp veri oranı vb. faktörleri dikkate alarak uygun yöntem seçimi yapması gerekmektedir (Cheema, 2012). Kayıp verilerle çalışmaya devam edildiği veya uygun yöntem seçilmediği takdirde çalışma sonuçları bu durumdan olumsuz yönde etkilenecektir (Ginkel ve diğerleri, 2007).

Araştırmanın Amacı

Bu araştırmanın amacı, kayıp veri sorununun çözümünde kullanılan farklı yöntemlerin etkililiğini ölçeklerin geçerlik (faktör yapıları-yapı geçerliği) ve güvenirliği (cronbach alfa) bağlamında, normal dağılım, tek faktörlü yapı ve farklı büyüklüklerdeki (%5, %10 ve %20) kayıp veri oranları altında inceleyerek karşılaştırmaktır. Bu amaç doğrultusunda, geçerlik ve güvenirliğe ilişkin olarak tam veri setinden elde edilen değerler ile tamamen rassal olarak kayıp (TROK) mekanizması altında veri eksiltilerek oluşturulan eksik veri setlerinin kayıp veri sorununun çözümünde kullanılan farklı yöntemlerle yeni tam veri setlerine dönüştürülmesi neticesinde elde edilen değerler karşılaştırılmıştır.

YÖNTEM

Araştırmanın Modeli

Bu araştırma kayıp veri sorununun çözümünde kullanılan farklı yöntemlerin etkililiğinin, ölçeklerin geçerliği ve güvenirliği bağlamında karşılaştırıldığı bir temel araştırmadır. Temel araştırmalar var olan bilgiye yenilerini eklemek amacıyla gerçekleştirilen teorik veya deneysel nitelikte çalışmalardır (Karasar, 2007).

Veriler ve Verilerin Düzenlenmesi

Araştırmada kullanılan tam veri setini, PISA 2012 sınavı kapsamında yer alan "Matematik Çalışma Etiği" ölçeğindeki tüm maddeleri eksiksiz olarak cevaplayan Türk öğrenciler (n=3127) arasından rastgele seçilen 200 öğrenciye ait veriler oluşturmaktadır. 200 öğrenci seçilmesinin nedeni bu yöntemlerin küçük örneklemlerdeki etkisini incelemektir. Tam veri setinden, araştırmanın amacı doğrultusunda, tamamıyla rassal olarak kayıp mekanizması altında, belirli oranlarda (%5, %10, %20) veri silinerek eksik veri setleri oluşturulmuş daha sonra bu setler kayıp veri sorununun çözümünde kullanılan farklı yöntemler ile yeni tam veri setlerine dönüştürülmüştür. Kayıp veri içeren veri setlerinin yeni tam veri setlerine dönüştürülmesinde seri ortalaması, yakın noktaların ortalaması, yakın noktaların medyanı, doğrusal değer kestirimi, noktanın doğrusal eğimi, liste bazında silme, beklenti maksimizasyonu, regresyon ataması ve çoklu atama kayıp veri yöntemleri kullanılmıştır.

Veri Toplama Aracı

Araştırmada kullanılan veri toplama aracı PISA 2012 öğrenci ölçeklerinden "Matematik Çalışma Etiği" ölçeğidir. Tek faktörlü bir yapıya sahip olan bu ölçek, "Kesinlikle katılıyorum (1)", "Katılıyorum (2)", "Katılmıyorum (3)", "Kesinlikle katılmıyorum (4)" şeklinde 4'lü Likert tipi dereceleme ölçeği formatında olup toplam 9 maddeden oluşmaktadır.

Verilerin Analizi

Verilerin analizi sürecinde öncelikle, tam veri setindeki veriler üzerinden tek boyutluluk ve puan dağılımlarının normalliği test edilmiştir. Temel bileşenler analizi neticesinde ölçekte yer alan tüm maddelerin ilk boyut altında yüksek yük değerleri verdiği ve birinci boyuta ait öz değerin (5,26), ikinci boyuta ait öz değerin (1,14) üç buçuk katından daha fazla olduğu görülmüştür. Bu durum ölçeğin yapısının tek boyutlu olduğunun bir göstergesi sayılmıştır. Birinci boyutun açıkladığı varyans ise toplam varyansın % 58,41'i olarak tespit edilmiştir. Tek boyutlu ölçeklerde açıklanan varyansın minimum % 30 olması durumunun kabul edilebilir olduğu (Büyüköztürk, 2007) dikkate alındığında bu değerin oldukça yüksek olduğu söylenebilir. Tam veri seti için hesaplanan çarpıklık ve basıklık katsayıları ise sırasıyla 0,313 ve 0,214'tür. Bu katsayıların -1 ile +1 değerleri arasında oluşu dağılımların normalliğinin bir göstergesi sayılmıştır (Huck, 2012).

Verilerin analizi sürecinin ikinci aşamasında 200 kişilik tam veri setinden, %5, %10 ve %20 oranlarında veri silinerek eksik veri setleri oluşturulmuştur. Verilerin silinmesinde rastgelelik dikkate alınmış olsa da, rastgeleliğin sağlandığını görmek amacıyla verilere Little'ın TROK testi uygulanmıştır. Bunun için SPSS 20.0 programında bulunan BM algoritmasına bağlı olarak hesaplanan Little'ın TROK testi kullanılmıştır. Verilere BM algoritması uygulanırken 2 varsayımın sağlanması gerekmektedir; bunlardan birincisi verilerin kesikli olmaması, ikincisi de verilerin normal dağılmasıdır (Oğuzlar, 2001). Araştırma verileri bu iki varsayımı da sağladığı için BM algoritması uygulanmıştır.

Little'ın TROK testi neticesinde *p* değerleri %5 kayıp veri oranı için 0,99; %10 kayıp veri oranı için 0,74 ve %20 kayıp veri oranı için ise 0,88 olarak elde edilmiştir. Bu değerlerin 0,05'ten büyük olması verilerin TROK mekanizmasına uyduğunun göstergesidir (IBM,2012). Dağılımların normalliği ve TROK sonuçlarının incelenmesi neticesinde verilerin ilgili analizler için uygun olduğu sonucuna varılmıştır.

Tek boyutluluk, puan dağılımının normalliği ve Little'ın TROK testi sonuçları incelendikten sonra farklı oranlarda eksik veri içeren veri setleri kayıp veri sorununun çözümünde kullanılan farklı kayıp veri yöntemleri ile yeni tam veri setlerine dönüştürülmüştür. Dönüştürmede kullanılan yöntemler silme yöntemlerinden liste bazında silme ve atama yöntemlerinden yaklaşık değer atama (seri ortalaması, yakın noktaların ortalaması, yakın noktaların medyanı, doğrusal değer kestirimi, noktanın doğrusal eğimi) ile beklenti maksimizasyonu, regresyon ataması ve çoklu atama yöntemleridir.

Hem tam veri seti hem de farklı kayıp veri yöntemleri ile oluşturulan yeni tam veri setleri üzerinde geçerliğe ilişkin analizler ve karşılaştırmalar temel bileşenler analizine dayalı açımlayıcı faktör analizi ile elde edilen değerler üzerinden gerçekleştirilmiştir. Güvenirliğe ilişkin yapılan analizler ve karşılaştırmalarda ise Cronbach alfa güvenirlik katsayısı ve Fisher'in z istatistiklerinden yararlanılmıştır. 200 kişilik tam veri setinden geçerlik ve güvenirliğe ilişkin analizler neticesinde elde edilen değerler, karşılaştırmalarda referans değerler olarak kullanılmıştır.

BULGULAR

Geçerliğe ilişkin olarak yapılan analizlerin sonuçları, %5, %10 ve %20 kayıp veri oranları için ayrı ayrı olmak üzere, aşağıda Tablo 1, Tablo 2 ve Tablo 3'te sunulmuştur.

200 kişilik tam veri seti için hesaplanan madde faktör yük değerleri 0,71 ile 0,81 arasında değişmektedir. Bu veri seti için hesaplanan öz değer 5,26, açıklanan varyans değeri ise 58,41'dir. Bu değerler farklı kayıp veri yöntemlerinden elde edilen değerlerle yapılan karşılaştırmalarda referans değerler olarak kullanılmışlardır.

Örneklem büyüklüğü 200 ve kayıp veri oranı %5 olan durum için geçerliğe ilişkin (faktör yükleri, öz değer, açıklanan varyans) elde edilen bulgular

Tablo 1**.** Örneklem Büyüklüğü 200 ve Kayıp Veri Oranı %5 Olan Durum için Geçerlik Analizi Sonuçları

	Kayıp Veri Yöntemleri									
Maddeler	**Tam veri**	**Seriler ortalaması**	**Yakın noktaların ortalaması**	**Yakın noktaların medyanı**	**Doğrusal değer kestirimi**	**Noktanın doğrusal eğimi**	**Liste bazında silme**	**Beklenti maksimizasyonu**	**Regresyon ataması**	**Çoklu atama**
1	,71	,71	,71	,71	,70	,71	,71	,71	,72	,71
2	,78	,79	,79	,79	,79	,79	,79	,79	,79	,79
3	,77	,77	,77	,77	,77	,77	,77	,77	,77	,77

4	,73	,73	,77	,77	,72	,73	,72	,73	,73	,73
5	,80	,80	,80	,80	,80	,80	,80	,80	,80	,80
6	,81	,81	,81	,81	,80	,81	,80	,81	,81	,81
7	,77	,77	,77	,77	,77	,77	,77	,77	,77	,77
8	,75	,75	,76	,75	,75	,75	,76	,76	,76	,76
9	,75	,75	75	,75	,75	,75	,74	,75	,75	,75
Ö	5,26	5,24	5,24	5,23	5,21	5,24	5,25	5,27	5,27	5,26
AV	58,4	58,2	58,17	58,07	57,93	58,22	58,31	58,54	58,60	58,46

Ö = Öz değer; AV = Açıklanan Varyans

Kayıp veri sorununun çözümünde kullanılan farklı yöntemler çerçevesinde elde edilen madde faktör yük değerleri Tablo 1'de incelendiğinde, seriler ortalaması ve noktanın doğrusal eğimi yöntemlerinde ikinci madde hariç, tam veri seti ile aynı faktör yük değerlerinin elde edildiği görülmektedir. Yakın noktaların medyanı (2.ve 4.madde), beklenti maksimizasyonu (2.ve 8.madde) ve çoklu atama (2.ve 8.madde) yöntemlerinde ikişer madde hariç tam veri seti ile aynı faktör yük değerleri elde edilmiştir. Yakın noktaların ortalaması (2. 4. ve 8. madde) ve regresyon ataması (1. 2. ve 8.madde) yöntemlerinde ise üçer madde hariç, tam veri seti ile aynı sonuçlar elde edilmiştir. Doğrusal değer kestirimi yönteminde 4 maddede (1. 2. 4. ve 6. madde), liste bazında silme yönteminde ise 5 maddede (2. 4. 6. 8 ve 9. madde), tam veri setinden elde edilen değerlere göre, farklı faktör yük değerleri elde edilmiştir.

Öz değerler incelendiğinde bu değerlerin 5,21 ile 5,27 arasında değiştiği görülmektedir. Çoklu atama yöntemi için hesaplanan öz değer tam veri seti için hesaplanan öz değerle aynıdır (5,26). En düşük ve tam veri setinden elde edilen öz değere en uzak değer doğrusal değer kestirimi (5.21) yönteminden elde edilirken, en yüksek öz değeri beklenti maksimizasyonu ve regresyon ataması (5,27) yöntemleri vermiştir.

Açıklanan varyans değerleri incelendiğinde ise tam veri setinden elde edilen değere en yakın değeri veren yöntemin çoklu atama (58,46) yöntemi olduğu görülmektedir. En büyük açıklanan varyans değeri regresyon ataması (58,60) yönteminden elde edilirken, en düşük değer doğrusal değer kestirimi (57,93) yönteminden elde edilmiştir.

Örneklem büyüklüğü 200 ve kayıp veri oranı %10 olan durum için geçerliğe ilişkin (faktör yükleri, öz değer, açıklanan varyans) elde edilen bulgular

Tablo 2**.** Örneklem Büyüklüğü 200 ve Kayıp Veri Oranı %10 Olan Durum için Geçerlik Analizi Sonuçları

		Kayıp Veri Yöntemleri								
Maddeler	**Tam veri**	**Seriler ortalaması**	**Yakın noktaların ortalaması**	**Yakın noktaların medyanı**	**Doğrusal değer kestirimi**	**Noktanın doğrusal eğimi**	**Liste bazında silme**	**Beklenti maksimizasyonu**	**Regresyon ataması**	**Çoklu atama**
1	,71	,72	,72	,72	,71	,72	,73	,73	,72	,73
2	,78	,79	,79	,78	,79	,79	,80	,79	,79	,79
3	,77	,77	,77	,77	,77	,77	,77	,77	,78	,77
4	,73	,74	,74	,74	,74	,74	,75	,74	,74	,74
5	,78	,79	,79	,79	,79	,79	,80	,79	,79	,79
6	,81	,81	,80	,80	,80	,81	,82	,81	,81	,81
7	,77	,78	,78	,78	,78	,78	,78	,78	,78	,78
8	,75	,75	,75	,75	,75	,75	,75	,75	,75	,75

9	,75	,74	,74	,74	,74	,74	,75	,75	,74	,74
Ö	5,26	5,27	5,27	5,26	5,25	5,27	5,36	5,31	5,29	5,30
AV	58,41	58,60	58,52	58,45	58,36	58,29	59,57	59,03	58,72	58,97

Ö = Öz değer; AV = Açıklanan Varyans

Kayıp veri sorununun çözümünde kullanılan farklı yöntemler çerçevesinde elde edilen madde faktör yük değerleri Tablo 2'de incelendiğinde, beklenti maksimizasyonu yönteminde 4 madde (3. 6. 8. 9. madde) hariç, tam veri seti ile farklı faktör yük değerleri elde edildiği görülmektedir. Seriler ortalaması (3. 6. ve 8.madde), yakın noktaların medyanı (2. 3. 8. madde), doğrusal değer kestirimi (1. 3. 8. madde), noktanın doğrusal eğimi (3. 6. 8. madde), liste bazında silme (3. 8. 9.madde) ve çoklu atama (3. 6. 8.madde) yöntemlerinde üçer madde tam veri seti ile aynı sonucu vermiştir. Tam veri seti ile en az benzerlik gösteren yöntemler ise yakın noktaların ortalaması ve regresyon ataması yöntemleridir. Bu yöntemler için sadece ikişer madde tam veri seti ile aynı sonucu vermiştir.

Öz değerler incelendiğinde, bu değerlerin 5,25 ile 5,36 arasında değiştiği görülmektedir. Yakın noktaların medyanı yöntemi için hesaplan öz değer (5,26) tam veri seti için hesaplanan öz değerle aynıdır. En düşük öz değer (5.25) doğrusal değer kestirimi yönteminden elde edilirken, en yüksek ve tam veri setine en uzak öz değeri (5,36) liste bazında silme yöntemi vermiştir.

Açıklanan varyans değerleri incelendiğinde ise tam veri setinden elde edilen değere en yakın değeri (58,45) yakın noktaların medyanı yöntemi vermiştir. Farklı kayıp veri yöntemleri için elde edilen en büyük açıklanan varyans değeri (59,57) liste bazında silme yönteminden elde edilirken, en düşük değer ise (58,29) noktanın doğrusal eğimi yönteminden elde edilmiştir.

Örneklem büyüklüğü 200 ve kayıp veri oranı %20 olan durum için geçerliğe ilişkin (faktör yükleri, öz değer, açıklanan varyans) elde edilen bulgular

Tablo 3. Örneklem Büyüklüğü 200 ve Kayıp Veri Oranı %20 Olan Durum için Geçerlik Analizi Sonuçları

	Kayıp Veri Yöntemleri									
Maddeler	**Tam veri**	**Seriler ortalaması**	**Yakın noktaların ortalaması**	**Yakın noktaların medyanı**	**Doğrusal değer kestirimi**	**Noktanın doğrusal eğimi**	**Liste bazında silme**	**Beklenti maksimizasyonu**	**Regresyon ataması**	**Çoklu atama**
1	,71	,72	,72	,72	,72	,72	,73	,72	,73	,72
2	,78	,76	,77	,77	,76	,76	,77	,78	,77	,77
3	,77	,75	,74	,75	,75	,75	,78	,77	,77	,76
4	,73	,72	,72	,71	,71	,72	,71	,73	,73	,73
5	,80	,79	,78	,78	,78	,78	,80	,80	,79	,79
6	,81	,81	,80	,80	,81	,81	,82	,81	,81	,81
7	,77	,78	,77	,78	,77	,78	,79	,77	,77	,77
8	,75	,75	,75	,75	,75	,76	,77	,76	,76	,75
9	,75	,75	,75	,75	,74	,75	,77	,75	,75	,75
Ö	5,26	5,17	5,13	5,12	5,13	5,17	5,33	5,28	5,27	5,24
AV	58,41	57,48	57,00	56,92	57,04	57,47	59,22	58,69	58,52	58,17

Ö = Öz değer; AV = Açıklanan Varyans

Kayıp veri sorununun çözümünde kullanılan farklı yöntemler çerçevesinde elde edilen madde faktör yük değerleri Tablo 3'te incelendiğinde beklenti maksimizasyonu yönteminin tam veri setinden elde

edilen değerlere en yakın sonuçları verdiği görülmektedir. Beklenti maksimizasyonu yönteminde iki madde (1. ve 8. madde) hariç tam veri seti ile aynı madde faktör yük değerleri elde edilmiştir. Regresyon ataması ve çoklu atama yöntemleri beklenti maksimizasyonu yöntemiyle birlikte tam veri setinden elde edilen değerlere en yakın sonuçları veren diğer yöntemlerdir. Regresyon ataması (1. 2. 5. 8. madde) ve çoklu atama (1. 2. 3. 5.madde) yöntemlerinde dörder madde hariç tam veri seti ile aynı faktör yük değerleri elde edilmiştir. Seriler ortalaması (6. 8. 9. madde), yakın noktaların ortalaması (7. 8. 9. madde) ve doğrusal değer kestirimi (6. 7. 8. madde) yöntemlerinde üçer madde tam veri seti ile benzerlik gösterirken; yakın noktaların medyanı (8 ve 9.madde) ve noktanın doğrusal eğimi (6 ve 9.madde) yöntemlerinde ikişer madde tam veri seti ile benzerlik göstermiştir. Liste bazında silme yönteminde ise sadece beşinci maddenin tam veri seti ile aynı sonucu verdiği görülmektedir.

Öz değerler incelendiğinde, bu değerlerin 5,12 ile 5,33 arasında değiştiği görülmektedir. Tam veri setinden elde edilen öz değere en yakın sonucu (5,27) veren yöntem regresyona ataması yöntemi olmuştur. En düşük ve tam veri setinden elde edilen öz değere en uzak değer (5,12) yakın noktaların medyanı yönteminden elde edilirken, en yüksek öz değeri (5,33) liste bazında silme yöntemi vermiştir.

Açıklanan varyans değerleri incelendiğinde ise tam veri setinden elde edilen değere en yakın değeri (58,52) regresyon ataması yönteminin verdiği görülmektedir. Farklı kayıp veri yöntemleri için elde edilen en büyük açıklanan varyans değerini (59,22) liste bazında silme yöntemi verirken en düşük değer (56,92) yakın noktaların medyanı yönteminden elde edilmiştir.

Güvenirliğe ilişkin olarak yapılan analizlerin sonuçları (Cronbach alfa güvenirlik katsayıları ve Fisher'in z testi) %5, %10 ve %20 kayıp veri oranları için ayrı ayrı olmak üzere Tablo 4'te verilmiştir.

200 kişilik tam veri seti için hesaplanan Cronbach alfa güvenirlik katsayısı 0,910'dur. Bu değer farklı kayıp veri yöntemlerinden elde edilen değerlerle yapılan karşılaştırmalarda referans değerler olarak kullanılmışlardır.

Tablo 4'de görüldüğü üzere, %5 oranında kayıp veri içeren veri seti için farklı kayıp veri yöntemleri çerçevesinde elde edilen Cronbach alfa güvenirlik katsayıları 0,908 ile 0,911 arasında değişmektedir. En düşük Cronbach alfa katsayısı (α= 0,908) doğrusal değer kestirimi yönteminden elde edilirken, liste bazında silme, beklenti maksimizasyonu ve çoklu atama yöntemleri tam veri seti ile aynı katsayıyı (α= ,910) vermişlerdir. Seri ortalaması, yakın noktaların ortalaması, yakın noktaların medyanı, noktanın doğrusal eğimi ve regresyon ataması yöntemleri ise tam veri setine yakın katsayılar verirken, en yüksek katsayı (α= 0,911) regresyon ataması yönteminden elde edilmiştir.

%10 oranında kayıp veri içeren veri seti için farklı kayıp veri yöntemleri çerçevesinde elde edilen Cronbach alfa güvenirlik katsayılarının 0,910 ile 0,914 arasında değiştiği görülmektedir. Seri ortalaması yöntemi hariç yaklaşık değer atama yöntemleri tam veri seti ile aynı katsayıyı (α= 0,911) verirken; en yüksek katsayıyı (α=0,914) liste bazında silme yöntemi vermiştir. Seri ortalaması, regresyon ataması ve çoklu atama yöntemleri ise tam veri setine göre yüksek katsayılar (α= 0,911) verse de bu katsayılar tam veri setinden elde edilen katsayıya oldukça yakındır.

Tablo 4. Örneklem Büyüklüğü 200 ve Kayıp Veri Oranlarının %5, %10 ve %20 Olduğu Durumlar için Güvenirlik Analizi Sonuçları

Kayıp Veri Yöntemleri	**%0 kayıp (Tam Veri)**	**%5 kayıp**	**%10 kayıp**	**%20 kayıp**
	Cronbach α	**Cronbach α** ***(Fisher z)***	**Cronbach α** ***(Fisher z)***	**Cronbach α** ***(Fisher z)***
Seri ortalaması	0,910	0,909 *(0,000)*	0,911 *(0,000)*	0,906 *(0,290)*
Yakın noktaların ortalaması		0,909 *(0,000)*	0,910 *(0,000)*	0,905 *(0,290)*
Yakın noktaların medyanı		0,909 *(0,000)*	0,910 *(0,000)*	0,904 *(0,290)*
Doğrusal değer kestirimi		0,908	0,910	0,905

	(0,000)	*(0,000)*	*(0,290)*
Noktanın doğrusal eğimi	0,909 *(0,000)*	0,910 *(0,000)*	0,906 *(0,290)*
Liste bazında silme	0,910 *(0,000)*	0,914 *(-0,290)*	0,913 *(-0,290)*
Beklenti maksimizasyonu	0,910 *(0,000)*	0,912 *(0,000)*	0,911 *(0,000)*
Regresyon ataması	0,911 *(0,000)*	0,911 *(0,000)*	0,910 *(0,000))*
Çoklu atama	0,910 *(0,000)*	0,911 *(0,000)*	0,909 *(0,000)*

%20 oranında kayıp veri içeren veri seti için farklı kayıp veri yöntemleri çerçevesinde elde edilen cronbach alfa güvenirlik katsayılarının ise 0,904 ile 0,913 arasında değiştiği görülmektedir. En düşük katsayıyı (α= 0,904) yakın noktaların medyanı yöntemi verirken, en yüksek katsayı (α= 0,913) liste bazında silme yönteminde elde edilmiştir. Regresyon ataması yöntemi ile elde edilen katsayı (α= 0,910) tam veri setinden elde edilen katsayı ile aynı iken beklenti maksimizasyonu (α= 0,911) ve çoklu atama (α= 0,909) yöntemleri tam veri setinden elde edilen katsayıya en yakın katsayıları vermiştir. Yaklaşık değer atama yöntemlerinden elde edilen katsayıların ise tam veri setinden elde edilene göre düşük olduğu görülmektedir.

Tam veri setinden elde edilen Cronbach alfa güvenirlik katsayısı ile farklı oranlarda kayıp veri içeren veri setlerine uygulanan farklı kayıp veri yöntemleri çerçevesinde elde edilen Cronbach alfa güvenirlik katsayıları arasında manidar bir farklılık bulunup bulunmadığına ilişkin olarak gerçekleştirilen Fisher'in z testi sonuçları incelendiğinde, tüm kayıp veri oranları ve tüm kayıp veri yöntemleri için hesaplanan z değerlerinin sınır değerler olan -1,96 ile + 1,96 aralığında (Akhun, 1994; Kenny, 1987) değerler aldığı görülmektedir. Bu bulgu, tam veri seti ile kayıp veri yöntemlerinin uygulandığı veri setlerinden elde edilen Cronbach alfa güvenirlik katsayıları arasında manidar bir farklılığın bulunmadığını göstermektedir.

SONUÇLAR ve TARTIŞMA

Bu araştırmada, kayıp veri sorununun çözümünde kullanılan farklı yöntemlerin etkililiği ölçeklerin geçerlik (faktör yapıları-yapı geçerliği) ve güvenirliği (cronbach alfa) bağlamında, normal dağılım, tek faktörlü yapı ve farklı büyüklüklerdeki (%5, %10 ve %20) kayıp veri oranları altında incelenerek karşılaştırılmıştır.

Geçerliğe ilişkin olarak açımlayıcı faktör analizi bağlamında yapılan karşılaştırmalar, farklı kayıp veri oranları için tüm kayıp veri yöntemlerinin, tam veri setlerine benzer biçimde, tek faktörlü bir yapıyı gösterdiği görülmüştür. Bu bulgu Çokluk ve Kayri (2011) ile Chen, Wang ve Chen'in (2012) çalışmaları ile tutarlık göstermektedir

Kayıp veri oranı arttıkça madde faktör yük değerlerinin tam veri setlerinden elde edilen değerlere benzerliği küçük bir miktar azalsa da düşük ve yüksek faktör yük değeri veren maddelerin neredeyse tüm koşullar altında benzerlik gösterdiği gözlemlenmiştir. Beklenti maksimizasyonu, regresyon ataması ve çoklu atama yöntemleri farklı kayıp veri oranlarının tamamı için, özellikle de kayıp veri oranının yüksek olduğu durumlar için, genel olarak tam veri setlerinden elde edilen değerlere en yakın madde faktör yük değerlerini vermişlerdir. Yaklaşık değer atama ve silme yöntemlerinde ise kayıp veri setlerinin tamamı için kayıp veri oranı arttıkça, madde faktör yük değerlerinin tam veri seti ile benzerliği küçük düzeylerde olsa da azalmıştır.

Özdeğerler ve açıklanan varyans değerleri bakımından, genel olarak, yaklaşık değer atama yöntemlerinden elde edilen değerler tam veri setlerinden elde edilen değerlerden çok az da olsa düşük bulunmuştur. Kayıp veri yöntemleri içinde özellikle doğrusal değer kestirimi yöntemi tam veri setlerinden elde edilen değerlere göre en düşük değerleri veren yöntem olurken, liste bazında silme yöntemi ise genel olarak en yüksek değerleri veren yöntem olmuştur. Beklenti maksimizasyonu ve

regresyon ataması yöntemleri çoklu atama yönteminden sonra tam veri setlerinden elde edilen değerlere en yakın değerleri veren yöntemler olmuşlardır. Ulaşılan bu sonuçlar konu ile ilgili olarak yapılan çeşitli araştırmaların sonuçları ile büyük benzerlik göstermektedir (Bernaards ve Sijtsma, 2000; Bal, 2003; Bernaards ve Sijtsma, 2000; Cheema, 2012; Enders, 2004; Graham, Hofer ve Piccinin,1994; Musil, Warner, Yobas ve Jones, 2002; Peng, Harwell, Liou ve Ehman, 2006; Streiner, 2002).

İlgili alan yazında kayıp veri yöntemlerinden herhangi birinin her koşul altında benzer ve doğru sonuç verdiğine ilişkin bir bulguya rastlamak zordur. Kayıp veri yöntemlerinin etkililiği karşılaşılan kayıp veri mekanizmasına, kayıp veri miktarına, örneklem büyüklüğüne vb. etkenlere göre değişmektedir. Bu çalışmada ele alınan kayıp veri yöntemleri tamamıyla rassal olarak kayıp (TROK) mekanizması temelinde değerlendirilmiştir. Veri yapısı TROK olduğu zaman elde edilen değerler tam veri setlerinden elde edilen değerlerle birebir aynı çıkmayabilir ancak yapılan kestirimlerin kayıp veri yapısının ihmal edilebilir olmasından dolayı yansız olduğu kabul edilmektedir (Bernaards ve Sijtsma, 2000; Buhi, Goodson, ve Neilands, 2008; Graham, 2009). Bu araştırmada oluşturulan yeni veri setleri ile tam veri setlerinden elde edilen değerler arasında büyük benzerliklerin oluşu, ya da bir başka deyişle, çok küçük farklılıkların bulunuşunun nedeni veri yapısının TROK olmasıyla açıklanabilir.

Sonuç olarak, veri yapısı TROK olduğunda yöntemlerin birbirlerine göre üstünlüğü yok gibi görünse de çoklu atama, beklenti maksimizasyonu ve regresyon ataması yöntemlerinin diğer yöntemlere göre daha yüksek performans gösterdiği söylenebilir.

Güvenirliğe ilişkin olarak ise farklı kayıp veri oranları için farklı kayıp veri yöntemlerinin kullanılması neticesinde elde edilen Cronbach alfa güvenirlik katsayıları ile tam veri setinden elde edilen katsayıların genel olarak birbirlerinden farklılaştığı ancak bu farklılıkların çok küçük değerler düzeyinde olduğu görülmüştür. Kayıp veri oranının düşük olduğu durumda tam veri setinden elde edilen değerlere çok yakın güvenirlik katsayıları elde edilirken, kayıp veri miktarı arttıkça elde edilen katsayılar tam veri setlerinden elde edilen katsayılardan az da olsa farklılaşmıştır. Özellikle kayıp veri oranının yüksek olduğu durumlarda, yaklaşık değer atama yöntemlerinden elde edilen güvenirlik katsayıları tam veri setinden elde edilenlerden daha düşük çıkmıştır. Liste bazında silme yöntemi kayıp veri oranının düşük olduğu durumda genelde tam veri setlerinden elde edilen katsayılara yakın veya aynı katsayı değerini verirken, kayıp veri oranı arttıkça tam veri setlerinden elde edilen katsayılardan daha yüksek değerler almıştır. Bu sonuç Demir'in (2013) çalışmasının sonuçlarıyla tutarlıdır. Yaklaşık değer atama yöntemlerinde, kayıp veri oranının fazla olduğu durumlarda yapılan kestirimlerin tam veri setinden elde edilen değerlere göre daha düşük olduğu görülmektedir. Bu sonuç Çokluk ve Kayri (2011) ile Demir'in (2013) çalışmasında ulaşılan sonuçlarla benzerlik göstermektedir. Beklenti maksimizasyonu, regresyon ataması ve özelikle de çoklu atama yöntemleri tüm farklı kayıp veri oranları için tam veri setlerinden elde edilen katsayılara çok yakın veya aynı değerleri vermiştir. Beklenti maksimizasyonu yönteminin tam verilerden elde edilen güvenirlik katsayılarına benzer kestirimlerde bulunabildiğine ilişkin elde edilen bu bulgu Enders'in (2004) çalışmasını destekler niteliktedir. Bunun yanı sıra, kayıp veri oranının yüksek olduğu durumda, çoklu atama yönteminin diğer kayıp veri yöntemlerine göre daha iyi sonuçlar vermesi Granberg-Rademacker (2007), Leite ve Beretvas (2010) ve Young, Weckman ve Holland'ın (2011) çalışmalarıyla da tutarlık göstermektedir.

Güvenirlik katsayıları bakımından tam veri setlerinden elde edilen değerlerle farklı kayıp veri yöntemlerinin kullanılması ile farklı kayıp veri oranları için elde edilen değerler arasında bir takım farklılıklar gözlemlense de Fisher'in *z* testi bu farklıkların istatistiksel olarak manidar olmadığını göstermiştir. Sonuç olarak, farklı kayıp veri yöntemlerinin, ölçeklerin güvenirliği bağlamında, araştırmada dikkate alınan koşullar altında, tam veri seti için hesaplanan katsayıya benzer sonuçlar verdiği söylenebilir. Ancak, yine de, özellikle kayıp veri oranının yüksek olduğu durumlarda, yaklaşık değer atama yöntemlerinin kullanılması neticesinde elde edilen Cronbach alfa katsayılarının tam veri setlerinden elde edilen katsayılardan daha düşük olmasından dolayı, kayıp verinin fazlalığı durumunda, bu yöntemlerin ölçeklerin iç tutarlılık güvenirliği üzerinde olumsuz etki yaratabileceği düşünülmektedir.

Araştırmada ulaşılan bu sonuçlar çerçevesinde özellikle kayıp veri miktarının fazla olduğu durumlarda, kayıp verilerin çoklu atama, beklenti maksimizasyonu veya regresyon ataması yöntemlerinden biri kullanılarak tamamlanması önerilebilir. Kayıp veri miktarının düşük olduğu durumda ise liste bazında silme yöntemi hariç diğer tüm yöntemler kullanılabilir.

Bu çalışma kapsamında farklı kayıp veri oranına göre farklı kayıp veri yöntemleri karşılaştırılmıştır. Çalışmada gerçek veriler kullanılmıştır. Çalışma kapsamında kullanılan veriler tek faktörlü bir yapıya ve normal dağılıma sahiptir. Kayıp veriler ise TROK mekanizmasındadır. Gelecekteki araştırmalara yönelik olarak farklı kayıp veri mekanizmaları altında çok faktörlü, çarpık dağılıma sahip, farklı örneklem büyüklüğü ve farklı kayıp veri oranlarına göre farklı yöntemler kullanarak yöntemlerin etkililiği karşılaştırılabilir.

KAYNAKÇA

Akhun, İ. (1994). *İstatistiksel formüller ve tablolar*. (4.Baskı). Ankara: Hacettepe Üniversitesi, Eğitim Faültesi

Allison, P.D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology, 112* (4), 545-557, DOI: 10.1037/0021-843X.112.4.545.

Allison, P.D. (2009). *Missing data.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 72-89. London: Sage Publication.

Bal, C. (2003). *Çok gruplu veri setlerinde eksik gözlem sorununun çözümlenmesi ve sağlık alanında bir uygulama*. Yayımlanmamış Doktora Tezi, Osmangazi Üniversitesi, Sağlık Bilimleri Enstitüsü, Eskişehir.

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, DOI: 10.1207/S15327906MBR3503_03.

Brown, M.L. & Kros, J. F. (2003). Data mining and the impact of missing data. *Industrial Management & Data System, 103* (8), 611- 621, DOI: 10.1108/02635570310497657.

Buhi, E.R., Goodson, P. & Neilands, T.B. (2008). Out of sight, not out of mind: Strategies for handling missing data. American Journal of Health Behavior, 32 (1), 83-92.

Büyüköztürk, Ş. (2007). Sosyal bilim için veri analizi el kitabı (7.Baskı). Ankara: Pegem Akademi.

Carpita, M. & Manisera, M. (2011). On the imputation of missing data in surveys with likert-type scales. *Journal of Classical, 28*, 93-112, DOI: 10.1007/s00357-011-9074-z

Cheema, J. (2012). *Handling missing data in educational research using SPSS.* Unpublished doctoral dissertation, George Mason University, USA.

Chen, S.F., Wang, S. & Chen, Y.C. (2012). A simulation study using EFA and CFA programs based the impact of missing data on test dimensionality. *Expert Systems with Applications, 39*, 4026–4031.

Cumming, P. (2013). Missing data and multiple imputation. Clinical Review & Education, 167 (7), 656-661.

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.

Demir, E. ve Parlak, B. (2012). Türkiye'de eğitim araştırmalarında kayıp veri sorunu. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 3*(1), 230-241.

Demir, E. (2013). Kayıp verilerin varlığında çoktan seçmeli testlerde madde ve test parametrelerinin kestirilmesi: SBS örneği. *Eğitim Bilimleri Araştırmaları Dergisi, 3*(2), 47-68.

Downey, R.G. & King, C.V. (1998). Missing Data in Likert Ratings: A Comparison of Replacement Methods. *The Journal of General Psychology, 125* (2), 175-191, DOI: 10.1080/00221309809595542.

Duncan, T.E., Duncan, S.C. ve Li, F. (1998). A comparison of model- and multiple imputation-based approaches to longitudinal analyses with partial missingness. Structural Equation Modeling: *A Multidisciplinary Journal, 5* (1), 1-21, DOI: 10.1080/10705519809540086.

Enders, C.K. (2001). A primer on maximum likelihood algorithms available for use with missng data. *Structural Equation Modeling, 8*(1), 128-141, DOI: 10.1207/S15328007SEM0801_7.

Enders, C.K. (2004). The impact of missing data on sample reliability estimates: Implications for reliability reporting practices. *Educational and Psychological Measurement, 64*(3), 419-436, DOI: 10.1177/0013164403261050.

Enders, C.K. (2013). Dealing with missing data in developmental research. *Child Development Perspectives, 7* (1), 27- 31.

Field, A. (2009). *Discovering statistics using SPSS* (3. Edition). London: Sage Publication.

Finch, H., & Margraf, M. (2008). *Imputation of categorical missing data: A Comparison of multivariate normal and multinomial methods.* 20.11.2014 tarihinde http://www.mwsug.org/proceedings/2008/stats/MWSUG-2008-S05.pdf adresinden alınmıştır.

Ginkel, J.R.V., Van der Ark, L.A., Sijtma, K. & Vermunt, J.K. (2007). Two-way imputation: A Bayesian method for estimating missing scores in tests and questionnaires, and an accurate approximation. *Computational Statistics & Data Analysis, 51,* 4013 – 4027, DOI:10.1016/j.csda.2006.12.022.

Ginkel, J.R.V., Sijtma, K., Van der Ark, L.A. & Vermunt, J.K. (2010). Incidence of missing item scores in personality measurement, and simple item-score imputation. *Methodology, 6* (1), 17-30, DOI: 10.1027/1614-2241/a000003

Graham, J.W., Hofer, S.M. ve Piccinin, A.M.(1994). Analysis with missing data in drug prevention research. Collins, L.M. ve Seitz, L.A. (eds.). *Advances in data analysis for prevention intervention research* (ss. 13- 64) içinde. National Institutes of Health.

Graham, J.W. (2009). Missing data analysis: Making it work in the real World. *Annual Review of Psychology, 60,* 549-576.

Granberg- Rademacker, J.S. (2007). A comparison of three approaches to handling incomplete state level data. *State Politics and Policy Quarterly, 7*(3), 325-338.

Hohensinn, C. ve Kubinger, K.D. (2011). On the impact of missing values on the item fit and the model validness of the Rasch model. *Psychological Test and Assessment Modeling, 53* (3), 380-393.

Huck, S.W. (2012). *Reading statistics and research* (6.Edition). USA: Pearson.

IBM (2012). 28.11.2014 tarihinde http://www-01.ibm.com/support/knowledgecenter/SSLVMB_20.0.0/com.ibm.spss.statistics.cs/mva_describe_rerun_mcartest.htm sitesinden alınmıştır.

Karasar, N. (2007). *Bilimsel araştırma yöntemi: kavramlar, ilkeler, teknikler*. (17. Baskı). Ankara: Nobel Yayın Dağıtım.

Kenny, D.A. (1987). *Statistics for the social and behavioral science*. USA.

Leite, W. ve Beretvas, S.N. (2010). The performance of multiple imputation for likert-type items with missing data. *Journal of Modern Applied Statistical Methods, (9)*1, 64-74.

Little, R.J.A. (1988). Missing data adjustments in large surveys. *Journal of Business & Economic Statistics, 6* (3), 287-296.

McKnight, P.E., McKnight, K.M., Sidani, S. & Figueredo, A.J (2007). *Missing data: A gentle introduction.* United States of America: The Guilford Press.

Musil,C.M., Warner, C.B, Yobas, P.K. & Jones, S.L. (2002). A comparison of imputation techniques for handling missing data. *Western Journal of Nursing Research, 24*(7),815-829,DOI: 10.1177/019394502237390.

Oğuzlar, A. (2001, Eylül). *Alan araştirmalarinda kayıp değer problemi ve çözüm önerileri.* V. Ulusal Ekonometri ve İstatistik Sempozyumu'nda sunulan bildiri. Çukurova Üniversitesi, Adana.

Peng, C.-Y. J., Harwell, M., Liou, S.-M & Ehman L. H. (2006). *Advances in missing data methods and implications for educational research*. In S. Sawilowsky (ed.), *Real data analysis* (ss.31-78) içinde. Greenwich, CT: Information Age Publishing Inc.

Satıcı, E. ve Kadılar, C. (2009). Kayıp gözlem olduğunda kitle ortalamasının tahmini. *Anadolu Üniversitesi Bilim ve Teknoloji Dergisi, 10(2), 549*-556.

Streiner, D.V. (2002). The case of the missing data: Methods of dealing with dropouts and other research vagaries. *Research methods in Psychiatry, 47,* 68-75.

Van der Ark, L. A. & Vermunt, J. K. (2010). New developments in missing data analysis. *Methodology, 6*(1), 1-2, DOI: 10.1027/1614-2241/a000001

Vansteelandt, S., Carpenter, J. & Kenward, M.G. (2010). Analysis of incomplete data using inverse probability weighting and doubly robust estimators. *Methodology, 6*(1), 37-48. DOI: 10.1027/1614-2241/a000005.

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.

Young, W., Weckman, G. & Holland, W. (2011) A survey of methodologies for the treatment of missing values within datasets: limitations and benefits, *Theoretical Issues in Ergonomics Science, 12* (1), 15-43, DOI: 10.1080/14639220903470205

EXTENDED ABSTRACT

Introduction

One of the major problems in researches in social and behavioral sciencies is missing data (Vansteelandt, Carpenter, Kenward, 2010; Ginkel, Sijtsma, Van der Ark, Vermunt, 2010). Mechanical

mistakes such as missing questions in long questionnaires, not recording the data in experimental works, and making research on delicate topics such as sexual behaviour (Field, 2009), leaving questions blank because of inattention or not knowing the answer (Finch and Margraf, 2008) are among the main reasons of missing data. Missing data in researches means lack of knowledge and thus causes infromation loss (Bal, 2003). According to McKnight et.al. (2007), missing data affect the results as follows: If the missing data amount is high, the realiability, generalizability and statistical inferences of the obtained data are affected dramatically, which will result in misleading statistical inferences. Besides, missing data will adversely affect the validity of the study.

In cases of missing data, reserachers generally prefer to exclude the missing data cases from the analysis. However, to do this, it is required to understand what causes missing data (Demir and Parlak, 2012) because missing data is caused by various reasons and patterns. Moreover, defining whether missing data is related to complete missing at random or missing at random orelse missing at nonrandom mechanisms is important in terms of determining which missing data method is going to be used in solving the missing data problem (Allison,2003).

Different methods have been developed to be used in solving the missing data problem in researches. Methods such as continuing the analysis with missing data, excluding the missing observations from the analysis, data allocation for the missing data or completing the missing data with various statistical methods are some of the methods used in missing data situations (Bal, 2003; Carpita ve Manisera, 2011; Duncan, Duncan ve Li, 1998; Downey ve King, 1998; Little, 1988).

In this research, the effects of different methods used in solving the missing data problem on the validity (factor structures- construct validity) and reliability (Cronbach alpha) of the scales were compared in the context of normal distribution and single-factor structure, at missing data rates of different sizes (5%, 10%, and 20%)

Method

"Math Work Ethics" scale in PISA 2012 examination was used as data collection tool. The tool is in the format of Likert type rating scale and consists of nine items. The complete data set used in the resarch consists of data obtained from 200 students randomly selected among the Turkish students (n=3127) who completely answered all the items at this scale. Data were deleted at certain rates (5%, 10% , 20%) from the complete data set and missing data sets were obtained. Afterwards, these sets were transformed into new complete data sets with different methods used in solving missing data problem. In transforming the data sets with missing data into complete data sets series mean, mean of nearby points, median of nearby points, linear interpolation, linear trend at point, listwise deletion, expectation maximization, regression imputation and multiple imputation methods were used.

During the data analysis, the single dimensionality and normality of score distributions of data in complete data set were tested. The consistency of data with *missing completely* at random (MCAR) mechanism was also controlled.

Analysis and comparisons on validity of the new complete data sets composed with both complete data set and different missing data methods were made with the values obtained with the exploratory factor analysis based on principal component analysis. On the other hand, the analyses and comparisons on reliability were made with Cronbach alpha reliability coefficient and Fisher's z statistics. The values obtained as a result of the analyses on validity and reliability of complete data set of 200 persons were used as reference values in comparisons.

Results and Discussion

The comparisons made in exploratory factor analysis of validity revealed that all missing data methods showed a single-factor structure, similar to complete data sets, for different missing data rates.

As the missing data rate increased, the similarity of item-factor loadings to the values obtained from complete data sets decreased with a low level but the items with low and high factor loadings were similar almost at all conditions.

The values obtained from approximate value imputation methods generally were found to be lower, despite slightly, than the ones obtained from complete data sets with respect to eigenvalues and explained variances. Expectation maximization and regression imputation methods were found to be the ones giving the values most proximal to the values obtained from complete data sets following the multiple imputation method.

For reliability, it has been found that the Cronbach alpha reliability coefficients obtained as a result of different missing data methods used for different missing data rates generally differentiated from the coefficients obtained from complete data set but the differences are at low levels and statistically not significant.

However, especially in cases of high rate of missing data, as the Cronbach alpha coefficients obtained as a result of using approximate value imputation methods are lower than the ones obtained from the complete data sets, these methods may cause adverse effect on the internal consistency reliability of the scales in cases of high level of missing data.

In light of the research results, especially in cases when missing data level is high, missing data is recommended to be completed by using multiple imputation, expectation maximization or regression imputation. When the missing data level is low, all methods except for listwise deletion can be used in completing missing data.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2015; 6(2); 268-278

5. Sınıf Seçmeli Ders Tercihlerinin Sıralama Yargıları Kanunuyla Ölçeklenmesi

Scaling 5th Grade Elective Course Preferences with Rank-Order Judgments

Selda ÖRS ÖZDİL * Esra KINAY **

Öz
Bu araştırmada, 4. sınıf öğrencilerinin, 5. sınıflar için MEB tarafından belirlenen 15 seçmeli dersi hangi sırayla tercih ettiklerini, tercih sıralamasının cinsiyet ve okul türü değişkenlerine göre farklılaşıp farklılaşmadığını sıralama yargıları kanunuyla ölçeklenmesi ile belirlemek amaçlanmıştır. Araştırma, Ankara ilinde özel okul ve devlet okullarında öğrenim gören toplam 316 öğrenci üzerinde yürütülmüştür. Verilerin toplanmasında araştırmacılar tarafından geliştirilen "5. Sınıf Seçmeli Dersler Tercih Sıralaması Formu" kullanılmıştır. Araştırma sonucunda, kız öğrencilerin daha çok "Yabancı Dil" ve "Görsel Sanatlar" derslerini, erkek öğrencilerin "Spor ve Fiziki Etkinlikler" ile "Zekâ Oyunları" derslerini; devlet okulu öğrencilerinin daha çok "Kur'an-ı Kerim" ve "Hz. Muhammed'in Hayatı" derslerini, özel okul öğrencilerinin ise "Spor ve Fiziki Etkinlikler" ile "Yabancı Dil" derslerini ilk sıralarda tercih ettikleri görülmüştür. Bu durum, cinsiyet ve okul türü değişkenlerinin seçmeli dersler tercih sırasında farklılaşmaya neden olduğunu göstermektedir.

Anahtar Kelimeler: 5. sınıf seçmeli dersler, ölçekleme, sıralama yargıları kanunuyla ölçekleme

Abstract
In this research it was aimed to determine 4th grade students' selection order and whether this order differentiate by means of gender and school type variables of 15 elective courses which are specified by the Ministry of Education for 5th grade students via using rank order judgment scaling method. This research was conducted with a sample of 316 students who were attending private and public schools in Ankara. Data were collected by using "5th Grade Elective Courses Preferences Ordering Form" which was developed by the researchers. According to the results in terms of gender, female students preferred "Foreign Language" and "Visual art" courses while male students preferred "Sports and Physical Activities" and "Mind Games", besides in terms of school type, public school students preferred "Quran" and "The life of the Prophet Mohammed" while private school students preferred "Sports and Physical Activities" and "Foreign Language" uppermost. This shows that gender and school type variables are effective on the selection order of the elective courses.

Key Words: 5th grade elective course, scaling, rank-order judgments scaling

GİRİŞ

Günümüz dünyasının ve bireylerinin, ekonomide, bilişim teknolojilerinde, sosyal ve kültürel hayatta yaşanan değişikliklere bağlı olarak istek ve beklentileri sürekli değişmektedir. Bu durum, okullardan beklenen işlevlerin niteliğini etkilemekte ve bireylerin ihtiyaçlarının karşılanmasını zorlaştırmaktadır. Birey, bir taraftan kendi toplumu ve dünya toplumunun bir üyesi olarak uyumlu bir yaşam için gerekli bilgi, beceri ve duygusal özellikleri bir bütünlük içinde kazanma; diğer taraftan da kendi ilgi ve yeteneklerini tanıma, geliştirme, hangi işleri daha iyi yapabileceğini yordayabilme ve gelecekteki eğitimini planlayabilmek için gerekli davranışları kazanma

*Doktora öğrencisi, Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara-Türkiye, seldaors85@gmail.com
**Doktora öğrencisi, Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara-Türkiye, esrakinay@gmail.com

Geliş Tarihi: 15.07.2015
Kabul Tarihi: 09.11.2015

ihtiyacındadır (Ülgen, 1992). Eğitim kurumları da bireylerin bu ihtiyaçlarına yönelik olarak, bireylerin farklılıklarını göz önünde bulunduran çeşitlendirilmiş eğitim programlarını geliştirmek ve uygulamakla yükümlüdür.

30.03.2012 tarihinde kabul edilen temel eğitim kanunu ile 8 yıllık zorunu kesintisiz eğitimin yerine 12 yıllık zorunlu kademeli eğitim sistemine geçiş yapılması kabul edilmiştir. Eylül 2012 tarihi itibariyle bu eğitim sistemi uygulamaya koyulmuştur. Kamuoyunda 4+4+4 olarak bilinen 12 yıllık zorunlu kademeli eğitim ilkokul, ortaokul ve lise kademelerinden oluşmaktadır. Eğitim sisteminde yapılan değişiklikler arasında, 5. sınıf öğretim kademesinin ortaokula dâhil edilmesi ile seçmeli derslerin ders saatinin ve çeşidinin arttırılması yer almaktadır.

Seçmeli derslerin öğrencileri hayata hazırlaması, ilgi ve yeteneklerini ortaya çıkarmada faydalı olması, okul programlarının ayrılmaz bir parçası olarak öğrencilerin gelişimlerine destek olması, ayrıca bilişsel (bilgi, beceri), duyuşsal (ilgi, tutum) ve sosyal gelişimlerine katkı sağlaması beklenmektedir. Hızla değişen dünyada öğrencilerin bu değişime ayak uydurabilmeleri için, yaşam becerilerinin de geliştirilmesi gerekmektedir (Eğitimi Araştırma Geliştirme Dairesi Başkanlığı [EARGED], 2008). 2012-2013 eğitim öğretim yılında uygulamaya koyulan seçmeli dersler ile farklı ilgi, ihtiyaç ve yeteneklere sahip öğrencilere farklı ders seçenekleri sunulduğu belirtilmektedir. Milli Eğitim Bakanlığı (MEB) tarafından yayımlanan seçmeli dersler genelgesinde, 5. sınıf öğrencileri için toplam 15 seçmeli ders belirlenmiş ve haftalık ders saati sayısına göre toplam 8 saat olacak şekilde ders seçebilecekleri ifade edilmiştir (MEB, 2012; Talim Terbiye Kurulu Başkanlığı [TTKB], 2012).

MEB tarafından belirlenen seçmeli dersler ile öğrencilerin akademik başarılarının yanında ilgi ve yeteneklerini keşfetmeleri ve geliştirmelerinin amaçlandığı; bu nedenle seçmeli derslerin öğrencilerin ilgi, yetenek ve istekleri doğrultusunda velisinin de rehberliği ile öğrenci tarafından seçilmesi gerektiği belirtilmiştir (MEB, 2012). Ancak bazı okullarda öğrenci talebi olmadığı için ya da öğretmen ve donanım eksikliği gerekçesiyle MEB tarafından belirlenen seçmeli derslerin tamamı açılamamış, ayrıca öğrenciler çoğunlukla veli ve yönetici isteklerine göre ders seçmek durumunda kalmıştır (Eğitim Reformu Girişimi [ERG], 2014; Çelik, Boz, Gümüş ve Taştan, 2013; Kaya, 2013; Memduhoğlu ve Mazlum 2013; Tanrıverdi ve Kardaş, 2013; Yayla ve Kozikoğlu, 2013). Ayrıca öğrenci ve velilerin seçmeli ders içerikleri ile ilgili bilgilendirilmediği belirtilmiştir (ERG, 2014).

2012-2013 öğretim yılında uygulamaya koyulan ortaokullardaki seçmeli dersler ile ilgili yapılan araştırmalar incelendiğinde (Yayla ve Kozikoğlu, 2013; Kaya, 2013; Memduhoğlu ve Mazlum, 2013; Karagözoğlu, 2015; Yayla ve Tat, 2013; Uçar, İpek ve Uçar, 2013; Tanrıverdi ve Kardaş, 2013; Çelik ve diğ., 2013) öğretmen, veli ya da yöneticilerin seçmeli ders seçim ve uygulama süreçleri ile ilgili görüşlerinin alındığı belirlenmiştir. Bunun yanında öğrenci tercihlerinin de incelendiği araştırmaların (ERG, 2014; Karagözoğlu, 2015; Tanrıverdi ve Kardaş, 2013; MEB, 2015) sınırlı sayıda olduğu ve bu araştırmalarda herhangi bir değişkene göre öğrencilerin seçmeli ders tercihlerinin incelenmediği görülmüştür. Bu nedenle, öğrencilerin MEB tarafından belirlenen seçmeli dersleri hangi sırada tercih ettiklerinin belirlenmesi önemlidir. Bu amaç doğrultusunda kullanılabilecek tekniklerden biri de ölçeklemedir.

Ölçekleme, uyarıcıların fiziksel büyüklükleri ile algılanan büyüklükleri arasındaki bağıntıyı bulmaya çalışan psikofizik bilim dalında ortaya çıkmıştır. Ölçekleme, ölçme sonucunda elde edilen ölçümlerin, belirli nitelikler kazandırmak amacıyla işlemlere tabii tutulması olarak tanımlanmaktadır. Ölçeklemede kullanılan deneysel yöntemler "Tepki Yaklaşımı" ve "Yargı Yaklaşımı" olmak üzere ikiye ayrılmaktadır (Turgut ve Baykul, 1992).

Tepki Yaklaşımı Yönteminde, K tane uyarıcı N kişilik bir denek grubuna uygulanarak onların tepkileri toplanır. Bu yaklaşımda tepkiyi veren kişiler tarafsız bilirkişi olarak değil, kendi tepkilerini belirten deneklerdir. Bu yöntemde denekler, her uyarıcının ölçekleme boyutundaki yerini aynı boyuttaki kendi yerlerine göre belirlemektedirler. Tepki yönteminin en bilinen örneği Likert tipi tutum ölçeği geliştirme çalışmalarıdır. Yargı Yaklaşımı Yönteminde ise, eldeki uyarıcılar, gözlemci veya bilirkişi yargılarına dayanarak belirlenmiş bir boyutta ölçeklenmektedir. N tane gözlemcinin her birinden K tane uyarıcının her birinin uyarıcılık derecesini belli bir yöntemle belirtmesi istenir. Gözlemcinin görevi, her uyarıcının ölçekleme boyutundaki büyüklüğünü diğer uyarıcılara göre

belirtmektir. Herhangi bir uyarıcı için gözlemci yargılarının ortalama değeri, onun ölçek değeri olarak kabul edilir. Bu yöntemde gözlemcilerin, kendi öznel yargılarını değil her bir uyarıcının diğer uyarıcılara göre bağıl durumunu olabildiğince tarafsız olarak belirlemeleri istenir. Bu yaklaşımın ise tipik örnekleri Thurstone yöntemiyle ölçeklemelerde görülmektedir. Yargı Yaklaşımı Yönteminde, her bir uyarıcının uyarıcılık derecesi ikili karşılaştırma, sınıflama, mutlak ve sıralama yargılarıyla ölçekleme yöntemleriyle belirlenmektedir (Turgut ve Baykul, 1992; Crocker ve Algina, 1986).

İkili karşılaştırmalar yöntemiyle ölçeklemede, N tane gözlemciye U_j ve U_k uyarıcılarından hangisinin uyarıcılık değerinin daha büyük olduğu sorulur. Örneğin gözlemciler belirtilen iki uyarıcıdan hangisinin "daha büyük", "daha iyi", "daha verimli" olduğuna karar verirler. Sınıflama yöntemiyle ölçeklemede, gözlemcilere K uyarıcının tümü verilir ve her uyarıcının önceden tanımlanmış sıralı sınıflardan hangisine düştüğünü belirtmeleri istenir. Mutlak yargılarla ölçeklemede, gözlemci veya bilirkişilere birbirinden oldukça farklı K tane uyarıcı verilerek, bir başlangıç noktası ve bir birim tanımlayabilmeleri için ipuçları verilir. Gözlemcilerin grafik veya sayısal bir dereceleme aracı üzerinde derece belirtmeleri, bir ucu sıfır noktası ve eşit aralıklı olarak belirlenmiş sınıflar üzerinde sınıflama işlemi yapmaları istenir. Mutlak yargıları toplama yöntemine bir tür sınıflama yöntemi olarak da bakılabilir. Sıralama yöntemiyle ölçeklemede ise, N tane gözlemciden K tane uyarıcının tümünü belirli bir nitelikte büyükten küçüğe ya da küçükten büyüğe doğru sıralaması ve her birine bir sıra sayısı vermesi istenir. Yapılan bir dizi istatistiksel işlem sonucunda ölçek değerleri elde edilir (Turgut ve Baykul, 1992). Yapılan araştırmada, öğrencilerin seçmeli ders tercihlerini en çok istedikleri dersten en az istedikleri derse doğru sıralamaları istendiğinden sıralama yargıları kanunuyla ölçekleme kullanılmıştır.

Yukarıda yapılan tartışmalar ışığında, 2012-2013 öğretim yılında yeniden düzenlenerek uygulamaya koyulan seçmeli dersler ile ilgili öğrenci görüşlerinin alınarak bu görüşlerin farklı değişkenlere (cinsiyet, okul türü, anne-baba eğitim düzeyi vb.) göre incelendiği bir araştırmaya rastlanmadığından, öğrencilerin seçmeli ders tercihlerinin incelenmesi önemli görülmüştür. Ayrıca araştırmanın, sıralama yargıları kanunu ile yapılan bir ölçekleme araştırması olması, Türkiye'de bu yöntemle yapılan araştırmaların (Şahin, Boztunç Öztürk ve Taşdelen Teker, 2015; Yalçın ve Avşar, 2014; Bal, 2011; Kan, 2008) az sayıda olması nedeniyle de alanyazına katkıda bulanacağı düşünülmektedir.

Araştırmanın Amacı

Bu araştırmanın amacı, 4. sınıfta öğrenim gören öğrencilerin bir sonraki öğretim yılında almak istedikleri seçmeli derslerin tercih sırasını belirlemektir. Bu genel amaç doğrultusunda aşağıdaki sorulara yanıt aranmıştır.

Sıralama yargılarına dayalı ölçekleme yönteminden;

1. tüm grup için elde edilen ölçek değerleri nasıldır?

2. cinsiyet değişkenine göre elde edilen ölçek değerleri nasıldır?

3. okul türü değişkenine göre elde edilen ölçek değerleri nasıldır?

YÖNTEM

Bu araştırmada, var olan bir durum var olduğu biçimiyle betimlendiğinden, tarama modelinde betimsel bir araştırmadır (Karasar, 2008).

Çalışma Grubu

Araştırma, 2012-2013 öğretim yılında Ankara ilinde, devlet okulunda ve özel okulda 4. sınıfta öğrenim görmekte olan toplam 316 öğrenci üzerinde yürütülmüştür. Çalışma grubunun cinsiyet ve okul türü değişkenlerine göre dağılımı Tablo 1'de verilmiştir.

Tablo 1. Çalışma Grubunun Cinsiyet ve Okul Türü Değişkenlerine Göre Dağılımı

		Okul Türü					
		Özel Okul		Devlet Okulu		Toplam	
		f	%	f	%	f	%
Cinsiyet	Kız	99	62,3	60	37,7	159	100
	Erkek	96	61,1	61	38,9	157	100
	Toplam	195	61,7	121	38,3	316	100

Tablo 1'e göre, araştırmaya dâhil olan kız ve erkek öğrencilerin sayılarının birbirine yakın olduğu ve özel okulda uygulama yapılan öğrenci sayısının devlet okuluna göre daha fazla olduğu görülmektedir.

Veri Toplama Aracı

Bu araştırmada, araştırmacılar tarafından geliştirilen "5. Sınıf Seçmeli Dersler Tercih Sıralaması Formu" kullanılmıştır. Form, MEB tarafından 31.08.2012 tarihinde yayımlanan seçmeli dersler konulu genelgede (MEB, 2012) yer alan 15 seçmeli ders dikkate alınarak geliştirilmiştir. Genelgede yer alan seçmeli dersler aşağıda verilmiştir.

1. Kur'an-ı Kerim
2. Hz. Muhammed'in Hayatı
3. Temel Dini Bilgiler
4. Okuma Becerileri
5. Yazarlık ve Yazma Becerileri
6. Yaşayan Diller ve Lehçeler
7. Yabancı Dil
8. Bilim Uygulamaları
9. Matematik Uygulamaları
10. Bilişim Teknolojileri ve Yazılım
11. Görsel Sanatlar (Resim, Geleneksel Sanatlar, Plastik Sanatlar vb.)
12. Müzik
13. Spor ve Fiziki Etkinlikler
14. Drama
15. Zekâ Oyunları

Hazırlanan formda, belirtilen seçmeli derslerin yanı sıra, öğrenci özelliklerini belirlemek amacıyla cinsiyet ve okul türü değişkenleri de yer almış ve bu değişkenler verilerin çözümlenmesinde bağımsız değişken olarak kullanılmıştır.

İşlem

5. Sınıf Seçmeli Dersler Tercih Sıralaması Formu, 4. sınıfta öğrenim görmekte olan 316 öğrenciye uygulanmış ve öğrencilerden formda yer alan derslerden en çok seçmek istediklerinin yanına "1", en az seçmek istediklerinin yanına "15" yazacak şekilde tüm dersleri tercih sırasına koymaları istenmiştir. Öğrenciler, farklı derslere aynı sıra numarasını vermemeleri konusunda uyarılmıştır. Formlar, belirlenen bir ders süresi içerisinde okul ortamında uygulanmış ve ardından toplanmıştır. Bu süreçte öğrencilerin, ders seçimlerine yönelik herhangi bir yönlendirmede bulunulmamış, öğrencilerin kendi istekleri doğrultusunda ders seçim sıralaması yapmaları istenmiştir. Öğrencilerin yaptıkları tercih sıralamaları sadece araştırma kapsamında kullanılmıştır.

Verilerin Analizi

Verilerin analizinde, sıralama yargılarıyla ölçekleme yöntemi kullanılmıştır. Öncelikle uyarıcılara ait sıra frekansları matrisi oluşturulmuş ve bu matris üzerinden frekanslar matrisi, frekanslar matrisine

bağlı olarak oranlar matrisi oluşturulmuştur. Oranlar matrisi üzerinden birim standart normal sapmalar matrisi (z_{ik}) elde edilmiştir. V. hal denklemleriyle ölçekleme yapılarak ölçek değerleri (S_i) hesaplanmıştır. Eksenin başlangıcı (O noktası) en küçük S_i değerine kaydırılarak ölçek değerleri (S_c) bulunmuştur. Her bir dersin ölçek değerleri, bağımsız değişken kategorilerinin her biri için ayrı ayrı hesaplanmıştır.

Ölçek değerlerinin elde edilmesinde kullanılan yöntemin varsayımlarının sağlanıp sağlanmadığının ve yargıların oluşturulmasında öğrencilerin dikkatli davranıp davranmadıklarının kontrol edilmesi için ölçek değerlerinin iç tutarlılık anlamında güvenirlik düzeyine bakılmıştır. Ölçek değerlerinin iç tutarlılığı, gözlenen frekanslardan elde edilen oranların (p_{ik}), ölçek değerlerinden elde edilen (teorik) oranlarla (p'_{ik}) ne dereceye kadar bağdaştığının belirlenmesi yoluyla yapılmaktadır (Turgut ve Baykul,1992).

Gözlenen ve teorik oranlar matrisleri yardımıyla ortalama hata değeri 0,008 olarak hesaplanmıştır. Hesaplanan ortalama hata, gözlenen değerlerle ampirik değerler arasındaki uyumun ortalama değerini, yani uyumun bir ölçüsünü vermektedir. Hesaplanan küçük ortalama hata değerleri, gözlemci yargılarının güvenilir olduğunu; büyük ortalama hata değerleri ise, gözlemci yargılarının güvenilir olmadığını ya da modeldeki varsayımların sağlanmadığını ifade etmektedir. Ancak bu değer, uyum derecesinin anlamlı olup olmadığı hakkında bilgi vermediğinden uyum derecesinin anlamlılığının test edilmesi için ki-kare istatistiği kullanılmıştır (Öğretmen, 2008; Turgut ve Baykul, 1992). Hesaplanan ki-kare değerinin 91 serbestlik derecesi ve 0,05 anlamlılık düzeyinde tablo ki-kare değerini aşmadığı görülmüştür [χ^2=16,836 < $\chi^2_{\text{Tablo (91; 0,05)}}$ = 69,126]. Bu durum, gözlemci yargılarının tutarlılığını ya da yöntemin varsayımlarının sağlandığını göstermektedir. Dolayısıyla araştırmada kullanılan verilere uygulanan ölçekleme yönteminin uygun ve ölçek değerlerinin iç tutarlılığa sahip olduğunu göstermektedir.

Bağımsız değişkenlerin her bir kategorisine ait frekans tablolarının oluşturulmasında "SPSS Statistics 15.0" programından; ölçek değerlerinin hesaplanmasında ve ölçek değerlerinin iç tutarlılığının belirlenmesinde "Microsoft Office Excel 2007" programından yararlanılmıştır.

BULGULAR

Araştırmanın bu bölümünde, ilk olarak öğrencilerin tümü üzerinden seçmeli ders tercihlerinin ölçeklenmesine ait bulgular sunulmuştur. Daha sonra öğrencilerin cinsiyet ve okul türüne göre ölçekleme işlemine ait bulgular sunularak elde edilen sonuçlar karşılaştırılmıştır.

Tüm Grup İçin Elde Edilen Ölçek Değerleri

Araştırma kapsamında yer alan tüm öğrencilerin, 5. sınıf seçmeli ders tercih sıralaması formuna verdikleri yanıtlardan elde edilen verilere ilişkin ölçek değerleri Tablo 2'de verilmiştir.

Tablo 2. Tüm Öğrencilerin 5. Sınıf Seçmeli Ders Tercihleri İçin Elde Edilen Ölçek Değerleri

Dersler	Sıra Numarası	Ölçek Değerleri
Spor ve Fiziki Etkinlikler	1	1,021
Yabancı Dil	2	0,997
Zekâ Oyunları	3	0,950
Görsel Sanatlar (Resim, Geleneksel Sanatlar, Plastik Sanatlar vb.)	4	0,939
Hz. Muhammed'in Hayatı	5	0,806
Matematik Uygulamaları	6	0,799
Kur'an-ı Kerim	7	0,775
Drama	8	0,696
Bilişim Teknolojileri ve Yazılım	9	0,643
Bilim Uygulamaları	10	0,608
Müzik	11	0,546
Temel Dini Bilgiler	12	0,506
Okuma Becerileri	13	0,386
Yazarlık ve Yazma Becerileri	14	0,349
Yaşayan Diller ve Lehçeler	15	0,000

Tablo 2'ye göre 4. sınıf öğrencilerinin, 5. sınıfta en çok seçmek istedikleri dersin "Spor ve Fiziki Etkinlikler" dersi; en az seçmek istedikleri dersin ise "Yaşayan Diller ve Lehçeler" dersi olduğu görülmektedir. "Spor ve Fiziki Etkinlikler" dersinin yanı sıra öğrencilerin, "Yabancı Dil" ve "Zekâ Oyunları" derslerini ilk sıralarda, "Okuma Becerileri" ile "Yazarlık ve Yazma Becerileri" derslerini ise son sıralarda seçmek istedikleri görülmektedir.

Cinsiyet Değişkenine Göre Elde Edilen Ölçek Değerleri

Araştırma kapsamında yer alan kız öğrencilerin, 5. sınıf seçmeli ders tercih sıralaması formuna verdikleri yanıtlardan elde edilen verilere ilişkin ölçek değerleri Tablo 3'te verilmiştir.

Tablo 3. Kız Öğrencilerin 5. Sınıf Seçmeli Ders Tercihleri İçin Elde Edilen Ölçek Değerleri

Dersler	Sıra Numarası	Ölçek Değerleri
Yabancı Dil	1	1,169
Görsel Sanatlar (Resim, Geleneksel Sanatlar, Plastik Sanatlar vb.)	2	1,109
Zekâ Oyunları	3	0,927
Spor ve Fiziki Etkinlikler	4	0,892
Müzik	5	0,862
Kur'an-ı Kerim	6	0,832
Drama	7	0,817
Matematik Uygulamaları	8	0,808
Hz. Muhammed'in Hayatı	9	0,805
Bilişim Teknolojileri ve Yazılım	10	0,666
Bilim Uygulamaları	11	0,640
Temel Dini Bilgiler	12	0,541
Yazarlık ve Yazma Becerileri	13	0,496
Okuma Becerileri	14	0,462
Yaşayan Diller ve Lehçeler	15	0,000

Tablo 3'e göre 4. sınıf kız öğrencilerinin, 5. sınıfta en çok seçmek istedikleri dersin "Yabancı Dil" dersi; en az seçmek istedikleri dersin ise "Yaşayan Diller ve Lehçeler" dersi olduğu görülmektedir. "Yabancı Dil" dersinin yanı sıra kız öğrencilerin, "Görsel Sanatlar (Resim, Geleneksel Sanatlar, Plastik Sanatlar vb.)" ve "Zekâ Oyunları" derslerini ilk sıralarda, "Yazarlık ve Yazma Becerileri" ile "Okuma Becerileri" derslerini ise son sıralarda seçmek istedikleri görülmektedir.

Araştırma kapsamında yer alan erkek öğrencilerin, 5. sınıf seçmeli ders tercih sıralaması formuna verdikleri yanıtlardan elde edilen verilere ilişkin ölçek değerleri Tablo 4'te verilmiştir.

Tablo 4. Erkek Öğrencilerin 5. Sınıf Seçmeli Ders Tercihleri İçin Elde Edilen Ölçek Değerleri

Dersler	Sıra Numarası	Ölçek Değerleri
Spor ve Fiziki Etkinlikler	1	1,180
Zekâ Oyunları	2	0,987
Yabancı Dil	3	0,838
Hz. Muhammed'in Hayatı	4	0,818
Matematik Uygulamaları	5	0,800
Görsel Sanatlar (Resim, Geleneksel Sanatlar, Plastik Sanatlar vb.)	6	0,779
Kur'an-ı Kerim	7	0,726
Bilişim Teknolojileri ve Yazılım	8	0,627
Bilim Uygulamaları	9	0,584
Drama	10	0,582
Temel Dini Bilgiler	11	0,477
Okuma Becerileri	12	0,313
Müzik	13	0,224
Yazarlık ve Yazma Becerileri	14	0,201
Yaşayan Diller ve Lehçeler	15	0,000

Tablo 4'e göre 4. sınıf erkek öğrencilerinin, 5. sınıfta en çok seçmek istedikleri dersin "Spor ve Fiziki Etkinlikler" dersi; en az seçmek istedikleri dersin ise "Yaşayan Diller ve Lehçeler" dersi olduğu görülmektedir. "Spor ve Fiziki Etkinlikler" dersinin yanı sıra erkek öğrencilerin, "Zekâ Oyunları" ve "Yabancı Dil" derslerini ilk sıralarda, "Müzik" ile "Yazarlık ve Yazma Becerileri" derslerini ise son sıralarda seçmek istedikleri görülmektedir.

Cinsiyet değişkeninden elde edilen bulgulara göre, kız öğrencilerin daha çok "Yabancı Dil" ve "Görsel Sanatlar" derslerini ilk sıralarda tercih ettikleri; erkek öğrencilerin ise "Spor ve Fiziki Etkinlikler" ile "Zekâ Oyunları" derslerini ilk sıralarda tercih ettikleri görülmektedir. Bu durum, cinsiyet değişkeninin seçmeli dersler tercih sırasında farklılaşmaya neden olduğunu göstermektedir.

Okul Türü Değişkenine Göre Elde Edilen Ölçek Değerleri

Araştırma kapsamında yer alan devlet okulu öğrencilerinin, 5. sınıf seçmeli ders tercih sıralaması formuna verdikleri yanıtlardan elde edilen verilere ilişkin ölçek değerleri Tablo 5'te verilmiştir.

Tablo 5. Devlet Okulu Öğrencilerinin 5. Sınıf Seçmeli Ders Tercihleri İçin Elde Edilen Ölçek Değerleri

Dersler	Sıra Numarası	Ölçek Değerleri
Kur'an-ı Kerim	1	1,325
Hz. Muhammed'in Hayatı	2	1,168
Yabancı Dil	3	0,950
Görsel Sanatlar (Resim, Geleneksel Sanatlar, Plastik Sanatlar vb.)	4	0,887
Zekâ Oyunları	5	0,879
Matematik Uygulamaları	6	0,866
Spor ve Fiziki Etkinlikler	7	0,834
Temel Dini Bilgiler	8	0,685
Müzik	9	0,575
Bilim Uygulamaları	10	0,484
Bilişim Teknolojileri ve Yazılım	11	0,463
Okuma Becerileri	12	0,459
Drama	13	0,442
Yazarlık ve Yazma Becerileri	14	0,377
Yaşayan Diller ve Lehçeler	15	0,000

Tablo 5'e göre 4. sınıf devlet okulu öğrencilerinin, 5. sınıfta en çok seçmek istedikleri dersin "Kur'an-ı Kerim" dersi; en az seçmek istedikleri dersin ise "Yaşayan Diller ve Lehçeler" dersi olduğu görülmektedir. "Kur'an-ı Kerim" dersinin yanı sıra devlet okulu öğrencilerinin "Hz. Muhammed'in Hayatı" ve "Yabancı Dil" derslerini ilk sıralarda, "Drama" ile "Yazarlık ve Yazma Becerileri" derslerini ise son sıralarda seçmek istedikleri görülmektedir.

Tablo 6. Özel Okul Öğrencilerinin 5. Sınıf Seçmeli Ders Tercihleri İçin Elde Edilen Ölçek Değerleri

Dersler	Sıra Numarası	Ölçek Değerleri
Spor ve Fiziki Etkinlikler	1	1,160
Yabancı Dil	2	1,043
Zekâ Oyunları	3	1,008
Görsel Sanatlar (Resim, Geleneksel Sanatlar, Plastik Sanatlar vb.)	4	0,985
Drama	5	0,863
Matematik Uygulamaları	6	0,769
Bilişim Teknolojileri ve Yazılım	7	0,760
Bilim Uygulamaları	8	0,692
Hz. Muhammed'in Hayatı	9	0,604
Müzik	10	0,534
Kur'an-ı Kerim	11	0,468
Temel Dini Bilgiler	12	0,398
Okuma Becerileri	13	0,342
Yazarlık ve Yazma Becerileri	14	0,333
Yaşayan Diller ve Lehçeler	15	0,000

Tablo 6'ya göre 4. sınıf özel okul öğrencilerinin, 5. sınıfta en çok seçmek istedikleri dersin "Spor ve Fiziki Etkinlikler" dersi; en az seçmek istedikleri dersin ise "Yaşayan Diller ve Lehçeler" dersi olduğu görülmektedir. "Spor ve Fiziki Etkinlikler" dersinin yanı sıra özel okul öğrencilerinin, "Yabancı Dil" ve "Zekâ Oyunları" derslerini ilk sıralarda, "Okuma Becerileri" ile "Yazarlık ve Yazma Becerileri" derslerini ise son sıralarda seçmek istedikleri görülmektedir.

Okul türü değişkeninden elde edilen bulgulara göre, devlet okulu öğrencilerinin daha çok "Kur'an-ı Kerim" ve "Hz. Muhammed'in Hayatı" derslerini ilk sıralarda tercih ettikleri; özel okul öğrencilerinin ise "Spor ve Fiziki Etkinlikler" ile "Yabancı Dil" derslerini ilk sıralarda tercih ettikleri görülmektedir. Bu durum, okul türü değişkeninin de seçmeli dersler tercih sırasında farklılaşmaya neden olduğunu göstermektedir.

SONUÇLAR ve TARTIŞMA

Bu araştırmada, 4. sınıf öğrencilerinin 5. sınıfta seçecekleri MEB tarafından belirlenen 15 seçmeli dersi tercih etme önceliğini belirlemek amaçlanmıştır. Öğrencilerin seçmeli dersleri tercih etme önceliği cinsiyet ve okul türü değişkenlerine göre ölçeklenmiştir.

Tüm grup üzerinde yapılan ölçekleme çalışması sonucunda, öğrencilerin ilk sıralarda sırasıyla "Spor ve Fiziki Etkinlikler", "Yabancı Dil" ve "Zekâ Oyunları" derslerini seçtikleri belirlenmiştir. Karagözoğlu (2013) ve ERG (2014)'nin yaptıkları araştırmalarda, "Yabancı Dil" ve "Spor ve Fiziki Etkinlikler" derslerinin ilk sıralarda tercih edilmesi bu araştırmanın bulgularıyla benzerlik göstermektedir. Tanrıverdi ve Kardaş (2013)'ın yaptığı araştırmada "Spor ve Fiziki Etkinlikler"; MEB (2015)'de yayımlanan istatistiklere göre ise "Yabancı Dil" dersinin ilk sıralarda seçilmesi araştırma bulgularını desteklemektedir. Araştırma bulgularından farklı olarak, Karagözoğlu (2013), ERG (2014), MEB (2015), Tanrıverdi ve Kardaş (2013), "Matematik Uygulamaları" ve "Kur'an-ı Kerim" derslerinin de öğrenciler tarafından ilk sıralarda tercih edildiğini ifade etmişlerdir.

Tüm grup üzerinde yapılan ölçekleme çalışması sonucunda, öğrencilerin son sıralarda "Yaşayan Diller ve Lehçeler" ve "Yazarlık ve Yazma Becerileri" derslerini tercih ettikleri belirlenmiştir. Bu bulgu, ERG (2014)'nin yaptığı araştırma sonuçlarıyla benzerlik göstermektedir. Karagözoğlu (2013)'nun yaptığı araştırmada ise "Okuma Becerileri" ve "Bilim Uygulamaları" derslerinin son sıralarda tercih edildiği belirtilmiştir. "Yaşayan Diller ve Lehçeler" dersinin adı itibariyle çalışma grubundaki öğrencilerin ilgisini çekmediği; ders içeriği incelendiğinde ise bu dersin çalışma grubundaki öğrencilere hitap etmediği düşünülmektedir.

Devlet okulu öğrencilerinden elde edilen bulguların dışında, diğer gruplarda "Spor ve Fiziki Etkinlikler", "Yabancı Dil", "Zekâ Oyunları" ve "Görsel Sanatlar (Resim, Geleneksel Sanatlar, Plastik Sanatlar vb.)" derslerinin öncelikle seçilmek istendiği, devlet okullarında ise "Kur'an-ı Kerim" ve "Hz. Muhammed'in Hayatı" derslerinin ilk iki sırada seçilmek istendiği görülmüştür. Bazı okullarda, öğretmen ve donanım yeterliliklerinin, yönetici görüş ve isteklerinin, velilerin sosyo-ekonomik düzeylerinin, görüş ve inançlarının öğrencilerin seçmeli ders tercih sıralamasındaki farklılaşmada etkili olduğu düşünülmektedir.

Cinsiyet ve okul türü değişkenlerinin seçmeli dersler tercih sıralamasında farklılaşmaya neden oldukları belirlenmiştir. Bunların yanı sıra hemen hemen tüm gruplarda "Yaşayan Diller ve Lehçeler", "Yazarlık ve Yazma Becerileri" ve "Okuma Becerileri" derslerinin en sonlarda tercih edildiği sonucuna ulaşılmıştır. Okuma ve yazma becerilerinin değerlendirildiği uluslararası araştırmalarda da Türkiye'nin okuma becerileri yeterlilik düzeylerin düşük olduğu ortaya konmaktadır. Araştırmanın sınırlılıkları da dikkate alınarak, okuma ve yazma becerilerine gerekli önemin verilmediği düşünülmektedir.

Araştırma sürecindeki gözlemlere dayanarak, okul yöneticilerinin, ders seçme sürecinde öğrenci ve velileri, ders içerikleri hakkında bilgilendirmeleri ancak belirli derslerin seçilmesi için öznel görüş ve inançlarından bağımsız olarak yönlendirmeleri önerilmektedir. Ayrıca, çağın gerektirdiklerine bağlı olarak öğrencilerin bireysel farklılıklarını ortaya çıkaracak şekilde seçmeli derslerin çeşitlendirilmesi önerilmektedir. Bu araştırmadaki sonuçlar doğrultusunda, benzer bir araştırma temsil gücü yüksek

bir örneklem üzerinde, cinsiyet ve okul türü değişkenlerinin yanı sıra farklı değişkenlere (anne-baba eğitim durumları, sosyoekonomik düzey vb.) göre gerçekleştirilebilir. Seçmeli ders tercihlerinin okul türü ve cinsiyet değişkenine göre farklılaşmasının nedenleri nitel bir araştırma ile incelenebilir.

KAYNAKÇA

Bal, Ö. (2011). Seviye belirleme sınavı (SBS) başarısında etkili olduğu düşünülen faktörlerin sıralama yargıları kanunuyla ölçeklenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 2(2), 200-209.

Crocker, L. & Algina J. (1986). *Introduction to classical and modern test theory*. Orlando: Harcourt Brace Jovanovich Inc.

Çelik, Z., Boz, N., Gümüş, S. ve Taştan, F. (2013). *4+4+4 Eğitim reformunu izleme raporu*. Eğitimciler Birliği Sendikası. http://www.egitimbirsen.org.tr/ebs_files/files/yayinlarimiz/267-egitimbirsen.org.tr-267.pdf adresinden edinilmiştir.

EARGED. (2008). *Seçmeli derslerin seçim kriterlerinin değerlendirilmesi araştırması*. Ankara: MEB. http://yegitek.meb.gov.tr/tamamlanan/secmeli_dersler_arastirmasi.pdf adresinden edinilmiştir.

ERG. (2014). *Temel eğitimin kademelendirilmesi sürecinin izlenmesi*. Eğitim Reformu Girişimi. http://erg.sabanciuniv.edu/sites/erg.sabanciuniv.edu/files/444.ArastirmaRaporu.04.03.14.WEB__0.pdf adresinden edinilmiştir.

Kan, A. (2008). Yargıcı kararlarına dayalı ölçekleme yöntemlerinin karşılaştırılması üzerine ampirik bir çalışma. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 35, 186-194.

Karagözoğlu, N. (2015). Ortaokul 5. sınıflarda tercih edilen seçmeli dersler ve tercih nedenlerinin öğrenci ve veli görüşlerine göre değerlendirilmesi. *Pegem Eğitim ve Öğretim Dergisi*, 5(1), 69-94.

Karasar, N. (2008). *Bilimsel Araştırma Yöntemi*. (18. Baskı). Ankara: Nobel Yayın Dağıtım.

Kaya, K. (2013). Okul idarecilerinin gözüyle seçmeli ders uygulaması. *Kesintili Oniki Yıllık Zorunlu Eğitim Modelinde Seçmeli Dersler Sempozyumu*. 24-25 Haziran 2013. Van.

Memduhoğlu, H. B. ve Mazlum, M. M. (2013). Seçmeli ders uygulamasının sosyal ve pedagojik temelleri ve yansımaları. *Kesintili Oniki Yıllık Zorunlu Eğitim Modelinde Seçmeli Dersler Sempozyumu.* 24-25 Haziran 2013. Van.

MEB. (2012). Seçmeli Dersler [Genelge]. http://tegm.meb.gov.tr/meb_iys_dosyalar/2012_08/31022530_semeliders.pdf adresinden edinilmiştir.

MEB. (2015). Öğrencilerin gözde dersleri belli oldu [Haber metni]. http://www.meb.gov.tr/ogrencilerin-gozde-dersleri-belli-oldu/haber/8377/tr adresinden edinilmiştir.

Öğretmen, T. (2008). Alan tercih envanteri: ölçeklenmesi, geçerliği ve güvenirliği. *Türk Eğitim Bilimleri Dergisi*, 6(3), 507-522.

Şahin, G. M., Öztürk Boztunç, N. ve Teker Taşdelen, G. (2015). Öğretmen adaylarının başarılarının değerlendirilmesinde tercih ettikleri ölçme araçlarının belirlenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 6(1), 95-106.

TTKB. (2012). Milli eğitim bakanlığı ilköğretim kurumları haftalık ders çizelgesi. http://nigde.meb.gov.tr/meb_iys_dosyalar/2012_06/27112918_ttk_69_25062012.pdf.pdf adresinden edinilmiştir.

Tanrıverdi, S. ve Kardaş, F. (2013). Öğretmenlerin, idarecilerin ve okul psikolojik danışmanlarının ortaokullarda seçmeli ders sürecine ilişkin görüşlerinin incelenmesi: Van ili örneği. *Kesintili Oniki Yıllık Zorunlu Eğitim Modelinde Seçmeli Dersler Sempozyumu*. 24-25 Haziran 2013. Van.

Turgut, M. F. ve Baykul, Y. (1992). *Ölçekleme teknikleri*. Ankara: ÖSYM Yayınları.

Uçar, R., İpek, Y. ve Uçar, İ. H. (2013). Ortaokul müdürlerinin seçmeli derslere yönelik tutumlarının incelenmesi. *Kesintili Oniki Yıllık Zorunlu Eğitim Modelinde Seçmeli Dersler Sempozyumu*. 24-25 Haziran 2013. Van.

Ülgen, G. (1992). İlköğretim okullarının 6, 7, 8., sınıflarında seçmeli dersler. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 1992, S. 8, s. 107-114.

Yalçın, S. ve Şengül Avşar, A. (2014). Eğitim fakültesi meslek bilgisi derslerinin sıralama yargıları kanunuyla ölçeklenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 5(2), 79-90.

Yayla, A. ve Kozikoğlu, İ. (2013). Seçmeli derslerin işlevselliği ve öğretmen görüşleri. *Kesintili Oniki Yıllık Zorunlu Eğitim Modelinde Seçmeli Dersler Sempozyumu*. 24-25 Haziran 2013. Van.

Yayla, A. ve Tat, O. (2013). Öğretmen perspektifinden seçmeli ders uygulaması: Problemler ve çözüm önerileri. *Kesintili Oniki Yıllık Zorunlu Eğitim Modelinde Seçmeli Dersler Sempozyumu*. 24-25 Haziran 2013. Van.

EXTENDED ABSTRACT

Introduction

Due to the basic education law legislated on March 30, 2012, 8-year-compulsory and uninterrupted education system was replaced with 12-year-compulsory and progressive education system. In this system, known as the 4+4+4 system, school levels are elementary, middle and high schools. Fifth grade was included in middle school and elective courses were offered in addition to the compulsory courses. In 2012-2013 academic year, The Ministry of National Education (MoNE) defined 15 elective courses for the fifth grade students. Elective courses are to be chosen by the students in accordance with their interests and aptitudes under the guidance of their guardians. Students' ranking the elective courses will be informative in terms of their interests, wishes and needs. Therefore, the elective course preferences of prospective fifth-grade students were needed to be explored. The purpose of this study is to identify the elective course preferences of the fourth grade students in 2012-2013 academic year, for the following year.

Method

This study is a descriptive research in survey model since examining the fifth grade elective course preferences of the fourth grade students is targeted. The study was carried out in 2012-2013 academic year with 316 fourth-grade-students. A data collection instrument, in which students are asked to rank 15 elective courses (the Quran, the life of the Prophet Mohammed, Basic Religious Knowledge, Reading Skills, Writing Skills, Living Languages and Dialects, Foreign Language, Science Practices, Mathematics Practices, Information Technologies and Software, Visual Arts, Music, Sports and Physical Activities, Drama, Mind Games) according to their preferences, was developed. The instrument also included gender and school type variables. The students were asked to rank the courses from 1 to 15, from the most desired to the least desired, respectively.

Rank order judgment scaling method was used for data analysis. Firstly, mean error was calculated and its significance was tested via chi-square statistics, in order to examine the degree to which the assumptions are met and the internal consistency. The mean error was 0,008, the assumptions were met according to the chi-square statistics and the scale values had internal consistency. Scale values were calculated primarily for the whole sample, and then according to gender and school type variables.

Results and Discussion

The scaling of the whole sample showed that the students chose "Sports and Physical Activities" to be their first preference and "Living Languages and Dialects" to be the last. Besides, it was observed that "Foreign Language" and "Mind Games" were at the top of the list, whereas "Reading Skills" and "Writing Skills" were towards the end. Similarly according to the Research conducted by ERG (2014) it was determined that "Foreign Languages" and "Sports and Physical Activities" courses are being selected in the first order on the other hand "Living Languages and Dialects" and "Writing Skills" courses are being selected lastly and these results are supporting the finding of this study. Moreover according to the research conducted by Karagözoğlu (2013), "Foreign Language" and "Sports and Physical Activities" courses; according to the research conducted by Tanrıverdi ve Kardaş (2013), Sports and Physical Activities" course also according to the statistics published by Ministry of Education "Foreign Language" course were selected uppermost. These results also show similarity with this research's findings.

According to the results of scaling in terms of gender, female students preferred "Foreign Language" at the top and " "Living Languages and Dialects" at the bottom. In addition, it was observed that "Visual Arts" and "Mind Games" were at the top of the list, whereas "Writing Skills" and "Reading Skills" were towards the end. Male students named "Sports and Physical Activities" to be their first preference and "Living Languages and Dialects" to be the last. Moreover, it was observed that "Mind

Games" and "Foreign Language" were at the top of the list, whereas "Music" and "Writing Skills" were towards the end. This result shows that elective course preferences differ by gender.

According to the results of scaling in terms of school type variable, state school students preferred "the Quran" at the top and " "Living Languages and Dialects" at the bottom. In addition, it was observed that "the life of the Prophet Mohammed" and "Foreign Language" were at the top of the list, whereas "Drama" and "Writing Skills" were towards the end. Private school students preferred "Sports and Physical Activities" at the top and "Living Languages and Dialects" at the bottom. Furthermore, it was observed that "Foreign Language" and "Mind Games" were at the top of the list, whereas "Reading Skills" and "Writing Skills" were towards the end. This result shows that elective course preferences differ by school type.

In addition, "Living Languages and Dialects", "Writing Skills" and "Reading Skills" were found to be at the end of the list for all groups, regardless of gender and school type variables.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2015; 6(2);194-206

Öğretim Yöntem ve Tekniklerinin Öğrenci Görüşlerine Göre İkili Karşılaştırma Yöntemiyle Ölçeklenmesi

Scaling of Teaching Methods and Techniques which is Used in Education Environment According to The Students Opinions By Pair-Wise Comparasion Method

Gökhan AKSU * Nuri DOĞAN **

Öz

Bu araştırma üniversite öğrencilerinin matematik dersinde hangi öğretim yöntem ve tekniklerin akademik başarıları bakımından yararlı olduğunu düşündüklerini belirlenmesi amacıyla gerçekleştirilmiştir. Bu amaç kapsamında uzman görüşleri doğrultusunda belirlenen 8 farklı yöntem ve tekniğin ikili karşılaştırma yöntemiyle ölçeklenmesi amaçlanmıştır. Bu kapsamda hazırlanan ölçme aracı 8x8 kare matris formatında hazırlanarak 2013 – 2014 Eğitim Yılı Bahar Yarıyılında Ege Bölgesinde bulunan bir devlet üniversitesinin merkez kampüsünde yer alan bir meslek yüksekokulunda öğrenim gören 310 öğrenciye uygulanmıştır. Yapılan ölçekleme işleminde önce V. Hal denklemi sonrasında III. Hal denklemi yardımıyla ölçekleme yapılmıştır. Yapılan ölçekleme işlemi sonucunda öğrenci görüşlerine göre matematik dersinde en faydalı bulunan öğretim yöntem ve tekniklerinin sırasıyla gösterip yaptırma, soru-cevap, düz anlatım, problem çözme, beyin fırtınası, tartışma, grup çalışması ve rol oynama-drama yöntemi olduğu belirlenmiştir. Elde edilen sonuçlara göre ders vermekle yükümlü öğretmen ve öğretim elemanlarının hangi yöntem ve teknikleri kullanmaları gerektiği konusunda önerilerde bulunulmuştur.

Anahtar Kelimeler: Öğretim yöntemleri, öğretim teknikleri, ölçekleme, ikili karşılaştırma

Abstract

The aim of this study is to determine the which teaching techniques and methods is more useful for academic success in math class according to university students' opinions. Prepared in accordance with expert opinion 8 different techniques and methods is aimed to scaling by using pair-wise comparision. The measurement tool prepared 8x8 square matrix format conducted to 310 students from a state university located in the Aegean region studying in a vocational school located in the central campus in 2013-2014 Academic Year. First the scaling process is performed by V. Case Equation then III. Case Equation. Afters caling process, according to students' opinions the most useful teaching methods and techniques in math class are demonstration, question and answer, lecture, problem solving, brain storming, discussion, group work and role-plays-drama method, respectively. According to the obtained results, suggestions were made for teachers and lecturers about the need to use which methods and techniques

Key Words: Teaching methods, teaching techniques, scaling, pairwise comparison method

*Öğr. Gör., Adnan Menderes Üniversitesi, Aydın MYO, Aydın-Türkiye, e-posta:gokhanaksu1983@hotmail.com
** Doç. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, ndogan@gmail.com

Geliş Tarihi: 02.02.2015
Kabul Tarihi: 01.08.2015

GİRİŞ

Sürekli değişen ve gelişen dünyada bugünün ve yarının gereksinimlerine yanıt vermesi gereken 21. yüzyılın öğretmenlerinin, öğrencilere yalnızca ders veren ve onları yılsonunda yaptıkları sınavlarla değerlendiren bireyler olmaları beklenmemektedir. Günümüzde öğretmenlerden, öğretme-öğrenme süreçlerini örgütleyebilen, iyi bir yönetici, iyi bir gözlemci ve nitelikli bir rehber olmaları beklenmektedir. Bu nedenle öğretmenlik mesleği artık daha fazla nitelik ve yeterlilik gerektiren bir meslek gurubu olarak görülmektedir (Gökçe, 1994). Klasik anlayışın aksine öğretmenlerden hem öğretim yöntem ve tekniklerini en iyi şekilde kullanmaları hem de çağdaş eğitimin gereksinimi olan daha modern sınıf yönetimi ve bunları öğrenme ortamları ile bütünleştirebilmeleri beklenmektedir (Kahyaoğlu ve Yangın, 2007). Sınıf ortamında düzenlenen eğitim-öğretim faaliyetleri planlı bir şekilde ilerleyen bir süreç olduğundan çağdaş eğitim anlayışına göre bu sürecin şekillenmesi ve kontrolü, sınıf içi faaliyetlerin düzenlenmesinde rehber olarak görülen öğretmenlere düşmektedir (Önen, Mertoğlu, Saka ve Gürdal, 2009). Günümüzde öğretmenlerden farklı öğretim yöntem ve teknikleri kullanmaları beklenmekte ancak öğretmenler daha çok kendilerinin merkezde olduğu, kendilerinin dersi ve öğrencileri yönlendirdiği, değerlendirmeyi kendilerinin yaptığı ve özellikle öğrencinin pasif alıcı durumunda olduğu yöntem ve teknikleri kullanma eğiliminde oldukları görülmektedir (Marbach, Seal ve Sokolove 2001, Junst, Licklider ve Wiersema 2003, Covill 2011).

Sınıf ortamında öğretmenleri etkin öğrenme yaşantısı oluşturma, onu sürdürüp geliştirmede ve sonuçlandırmadaki başarısı, çağdaş eğitim görüş ve ilkelerine, yöntemlerine dayandırılırsa gerçekleşebilmektedir (Kayabaşı, 2012). Yapılan araştırmalara göre (Önen, Saka, Erdem, Uzal ve Gürdal, 2008) farklı branşlardaki öğretmenlerin derslerinde öğretim yöntem-teknik ve stratejilere ilişkin yeterli bilgi ve beceriye sahip olmadıklarından, genellikle kendilerinin merkezde olduğu yöntem-teknik ve stratejileri tercih ettikleri gözlenmiştir. Araştırmalar, öğretmeni dinlerken ve okurken, zamanla öğrencilerin dikkati dağılarak, sıkılmaya başladıklarını göstermektedir (Larson at al., 1991; Lammers ve Murphy, 1999). Dolayısıyla sınıf içerisinde öğretme-öğrenme sürecinin etkili olabilmesi uygun öğretim yöntem ve tekniklerinin seçimiyle doğru orantılıdır (Hesapçıoğlu 2011; Demirel 1999). Bu nedenle eğitim öğretim ortamında en fazla temele alınması gereken yöntem aktif öğretim yöntemleri olmalıdır (Yeşilyurt, 2013). Bu yöntemlerden hangisinin iyi yöntem olduğu konusunda Vural (2006), öğrencilerde öğrenme isteği uyandıran, öğrencileri beden, zihin etkinliğine ve onların düşünmeye yönelten yöntemin iyi yöntem olduğunu belirtmektedir.

Diğer yandan bir öğretmenin yöntem ve teknikleri seçimini etkileyecek pek çok faktör vardır. Bunların en belli başlıları şöyle sıralanabilir; öğretmenin yöntem ve tekniğe olan yatkınlığı, yöntem ve tekniğin kullanılması için gerekli zaman ve fiziksel olanaklar, yöntem ve tekniğin maliyeti, öğrenci grubunun büyüklüğü, konunun içeriği ve özelliği, öğretim sonucunda öğrencide geliştirilmek istenen nitelikler ve öğrenci özellikleri vb. (Küçükahmet, 1995).

Öğretmenin, konunun içeriğine göre uygulayacağı yöntem veya teknik ile ders daha akıcı hale gelebilmekte, zamandan tasarruf edilip, ders daha eğlenceli hale gelebilmektedir (Ergani, 2010). Öğretmenlerin bilgiyi doğrudan aktarmak yerine, öğrencileri bilgiye yönlendirecek şekilde derslerini düzenlemesinin ve bu süreçte farklı yöntem-teknik ve yaklaşımları kullanmasının oldukça önemli olduğu söylenebilir. Ancak yapılan araştırmalar, öğretmenlerin öğrencilerinin derse aktif katılımını sağlayacak farklı yöntem ve teknikler hakkında yeterli bilgiye sahip olmadığını göstermektedir (Gönen ve Kocakaya, 2006). Araştırmalar, öğretmeni dinlerken ve okurken, zamanla öğrenci sıkıntısının çoğaldığını göstermektedir (Larson at al., 1991; Lammers ve Murphy, 1999). Böyle durumlarda öğretmen, uyguladığı yöntemi değiştirerek, örneğin öğrencilerin tartışmasını sağlamalı, konuyla ilgili değişik yaklaşımlara yönelmelidir (Cangelosi, 1988: 19) ve öğrencileri düşündürmeye yöneltecek ve onlarda eleştirel düşünme, probleme dayalı öğretim gibi bir takım düşünme becerilerini geliştirecek stratejiler kullanmalıdır (Ishiyama ve John 1999).

Artık çağımızda "Bilen öğretir" sloganı geçerliliğini kaybetmektedir. Bilenin bildiğini organize bir biçimde nasıl öğreteceğini de kesinlikle bilmesi gerekmektedir. Bu ise, öğretmenlerin meslek bilgisi konusundaki yetkinlikleriyle doğru orantılıdır denilebilir (Küçükahmet, 1983). Öğretim yöntemlerini

seçme ve kullanmada en önemli sorumluluk öğretmene düşmektedir. Yöntem kavramı, bir amaca ulaşmak için izlenen, tutulan yol, usul, sistem, prosedür, politika (TDK, 2014), düşünülmüş ve planlanmış bir hareket biçimi (Hesapcıoğlu, 2011) olarak tanımlanmaktadır. Demirel (2011), yöntemi hedefe ulaşmak için önceden belirlenmiş ya da izlenecek en kısa yol olarak tanımlamakta ve sınıf içi öğrenme-öğretme surecinin etkili olabilmesinin kullanılacak yöntemlerin seçimiyle doğru orantılı olduğunu belirtmektedir.

Eğitim öğretim ortamında çok sayıda yöntem ve teknik kullanılmaktadır. Kullanılan yöntemlerden bazıları anlatma, tartışma, örnek olay, gösterip yaptırma, problem çözme, yaratıcı drama ve bireysel çalışmadır (Demirel, 1999). Eğitim ortamında en çok kullanılan teknikler ise grupla öğretim teknikleri (Beyin fırtınası, soru-yanıt, gösteri, benzetim, ikili ve grup çalışmaları, mikro öğretim, eğitsel oyunlar, altı şapkalı düşünme tekniği), bireysel öğretim teknikleri (bireyselleştirilmiş öğretim, programlı öğretim, bilgisayar destekli öğretim) ve sınıf dışı öğretim teknikleri (görüşme, gözlem, sergi, proje, ödev) olarak sınıflanmaktadır. Aşağıdaki şekilde öğretim yöntem ve teknikleri verilmektedir

Şekil 1. Öğretim yöntem ve teknikleri

Kaynak: Büyükkaragöz ve Çivi (1996); Özden (2000); Saban (2004)

Çalışma kapsamında ele alınan bazı yöntem ve teknikler aşağıda kısaca açıklanmıştır.

1. Anlatım Yöntemi: Öğretmenin sahip olduğu bilgilerini sınıf ortamında pasif durumda olan öğrencilere aktarması esastır (Saban, 2004). Anlatma yöntemi, öğretmen merkezli bir öğretme yöntemi olup, daha çok öğretmenin bilgiyi öğrenenlere aktarması sürecini içermektedir. Anlatma yöntemi sözlü anlatıma ağırlık verdiği için anlatmayı gerektiren her türlü derste kullanılır, özellikle de sosyal bilgiler derslerinde yaygın olarak kullanılmaktadır. Bu yöntemin sınırlılıkları: Sürenin uzaması halinde ders sıkıcı hale gelir, iletişim az ya da yoktur, Dinleyiciler pasiftir, Öğrenme düzeyinin ne düzeyde olduğunu ölçmek zordur, daha çok bilişsel alana yöneliktir (Taşpınar, 2005).

2. Beyin Fırtınası Tekniği: Beyin fırtınası bir konuya çözüm getirmek, karar vermek ve hayal yoluyla düşünce ve fikir üretmek için kullanılan yaratıcı bir tekniktir. Bu teknikte doğru ve yanlış diye bir şey yoktur; önemli olan konu üzerinde çok miktarda fikir üretmektir. Önerilen her çözüm teklifi diğer grup üyelerini yeni fikirler üretmeye yönlendirir. Yöntemin amacı, belirli bir yargıya ya da sonuca ulaşmak değil, öğrencilerin yaratıcı düşünme becerilerini geliştirmek ve sorun ya da konu hakkında istenen alternatiflerin sayısını arttırmaktır (Saban, 2004). Yöntemde bilgi kadar görüş ve düşüncelerin aktarılması önemli olduğundan belirtilen tüm görüşler eleştirilmeden not alınmalıdır (Ergani,2010)

3. Soru-Cevap Tekniği: Sınıf içi uygulamalarda en yaygın kullanılan tekniklerden biridir. Bu teknik, öğrencilere düşünme ve konuşma alışkanlıklarını kazandırma bakımından oldukça önemlidir ve her dersin öğretiminde kullanılır. Konu hakkında önceden hazırlanmış olan soruların öğrencilere yöneltilmesi (Karaağaçlı, 2005) olarak da tanımlanan yöntemde soruları öğretmen sorabileceği gibi öğrencilerin öğretmene ya da öğrencilerin birbirlerine soru sormalarına olanak sağlanmalıdır. Bu yöntemle öğrenci düşünmeye özendirilir, güdülenir. Etkili konuşma alışkanlığı kazanır. Konuşma becerisi geliştirilir. Soru cevap yöntemi, öğrencilere cevap almak için soru sormak ve alınan cevapları eleştirerek öğretimin sağlanması yöntemi olarak da tanımlanabilir (Büyükkaragöz ve Çivi, 1996).

4. Tartışma Yöntemi: Tartışma bir konu üzerinde öğrencileri düşünmeye yöneltmek, iyi anlaşılamayan noktaları açıklamak ve verilen bilgileri pekiştirmek amacıyla kullanılan bir yöntemdir. Bu yöntem daha çok bir konunun kavranması aşamasında karşılıklı görüşler ortaya konulurken, bir problemin çözüm yollarını ararken ve değerlendirme çalışmaları yaparken kullanılır. Bu yöntem öğrencilerin konu üzerinde kendi düşüncelerini belirtmelerine ve konuya ilişkin yorum yapmalarına yardımcı olur. Karşılıklı saygı çerçevesinde demokratik duyguların da gelişmesine katkı sağlayan yöntem, öğrenmenin kalıcılığı açısından da önem arz etmektedir (Taşpınar, 2005)

5. Rol Oynama ya da Drama Tekniği: Rol yapma öğrencinin kendi duygu ve düşüncelerini başka bir kişiliğe girerek ifade etmesini sağlayan bir öğretme tekniğidir. Rol yapma sosyodrama olarak da adlandırılır. Diğer bir tanımla öğrencilere, insan ilişkileri konusunda daha çok bilgi, beceri ve anlayış kazandırmayı öngören; oyunlardan (drama) yararlanma temeline dayalı deneysel bir eğitim tekniğidir. Drama tekniği ile öğrenciler hangi durumlarda nasıl davranmaları gerektiğini öğrenirler. Problem çözme ve iletişim kurma yeteneğini geliştirir. Yöntemin bütün aşamalarının kontrollü bir şekilde gerçekleştirilmesi ile yöntemde elde edilecek faydalardan en önemlisi (Özden, 2000); anlatımında güçlük çekilen veya görece soyut konuların anlatımının kolay hale gelmesi ve öğrencilerin farklı rollere girmesi yoluyla gelişimlerini ve kendilerine olan güven duygusunu geliştirmesidir.

6. Gösterip Yaptırma Yöntemi: Bu yöntem bir işlemin uygulanmasını, bir araç gerecin çalıştırılmasını önce gösterip açıklama sonra da öğrenciye alıştırma ve uygulama yaptırarak öğretme yoludur. Bu yöntem, bir konuya ilişkin bilgilerin açıklanması ve bu bilgilerin beceriye dönüştürülmesi için gerekli uygulamaların yapılması aşamasında kullanılır. Gösteri öğretmen merkezli, yapma işlemi ise öğrenci merkezlidir. Öğrenciler kazandırılacak becerileri yaparak, yaşayarak öğrenirler. Yöntemin amacına ulaşabilmesi için her öğrenci tarafından gerçekleştirilmesi sağlanmalıdır. Bu yüzden ders sırasında öğrencilere eşit fırsatlarla gösteri yapabilmesi sağlanmalıdır. Yöntem hem görsel hem de işitsel alana hitap ettiği için öğrenme daha rahat ve kalıcı olur, Gösteride öğrencilerden de yardım alınması öğrenmeyi eğlenceli ve dikkat çekici hale getirmesi faydalarındandır. Ancak bunun yanında yöntemin uygulanmasında sınırlılıklar vardır. Gösterinin hazırlanması zaman alabilir, gösterilen bir etkinlik tekrar edilmediği zaman unutulabilir, kalabalık sınıflarda uygulamada sorunlar yaşanabilir ve öğretmenin öğrenci dönütlerine dikkat etmemesi öğrenimin sadece taklit edilmesi ile sonuçlanabilir (Büyükkaragöz ve Çivi, 1996).

7. Grup Çalışması (İşbirlikli Öğrenme) Tekniği: İşbirlikli öğrenme; sınıflarda ilerleme ve motivasyonu arttırmak için kullanılır. Öğrenciler ortak bir amaç için ortaklaşa çalışırlar ve bunu küçük gruplar kurarak gerçekleştirirler. Sınıflarda öğrenci sayısına göre en az iki ve en çok sekiz ya da on olacak şekilde gruplar oluşturulur. Ancak her grup çalışması işbirlikli öğrenme olmayabilir. İşbirlikli öğrenmede: öğrencinin kendisinin ve arkadaşlarının en üst seviyede çaba göstermeleri ve öğrenme gayreti içinde olmaları gerekir. Grup çalışması yönteminde, konu ya da sorun ile ilgili anlaşılmayan noktaları açıklığa kavuşturmak için o konu ya da sorun ile ilgili grubun üyelerini çözüm için düşünmeye yöneltmek esastır (Tan, Kayabaşı ve Erdoğan, 2002). Bu yöntemde bazı öğrenciler alışkanlık kazanmaları için gruba dahil edilirken bazı öğrenciler de lokomotif görevi ile gruba dahil edilirler (İşman ve Eskicumalı, 1999).

8. Problem Çözme Yöntemi: Problem çözme, istenilen hedefe varabilmek için etkili ve yararlı olan araç ve davranışları türlü olanaklar arasından seçme ve kullanmadır. Daha çok araştırma yoluyla öğretme yaklaşımında, bilişsel alanın uygulama düzeyindeki davranışların kazandırılmasında ve duyuşsal alanın analiz ve sentez özelliklerini geliştirmede kullanılır. Aşamaları: Problemin farkına varma, problemi tanıma ve sınırlama, problemin çözümü için hipotezler kurma, veri toplama ve toplanan verileri analiz etme, sonuçları yorumlama, denenceleri kabul ya da ret etme, çözümü uygulama ve elde edilen sonuçlara göre önerilerde bulunmadır. Yöntemin sağlayacağı faydalardan bazıları (Küçükahmet, 1983). Öğrenci bu yöntemde sürece aktif olarak katılım sağlanması, öğrenci düzenli ve planlı bir şekilde çalışmış ve bu şekilde çalışmaya alışmış olur, Öğrenciler başkaları ile çalışma, bilgilerinden faydalanma gibi olumlu alışkanlıklar kazanma, sorumluluk duygularında gelişme gibi değerlere sahip olurlar.

Alan yazında öğretmenler tarafından hangi yöntemin tercih edildiğine ilişkin farklı çalışmalar bulunmaktadır. Kılıç (2010) tarafından yapılan çalışmada devlet okullarında görev yapan öğretmenlerin düz anlatım yöntemini kullanmalarına karşın dershanede görev yapan öğretmenlerinin soru cevap yöntemini daha çok kullandıkları belirtilmiştir. Yıldız (2008) tarafından yapılan çalışmada anlatım yönteminin öğretmene zaman kazandırması bakımından avantaj sağladığı için tercih edildiği açıklanmıştır. Ergani (2010) çalışmasında derslerde öğretmenlerin en çok soru cevap yöntemini kullandıklarını belirtmektedir. Uysal (2010) tarafından yapılan çalışmada öğretmenlerin sınıf ortamında en fazla soru cevap ve anlatım yöntemini kullandıkları belirtilmiştir. Tohumcu (2004) tarafından yapılan çalışmada matematik dersinde başarılı olan sınıflarda öğretmenlerin en fazla kullandıkları yöntem ve tekniklerin sırasıyla soru cevap, gösterip yaptırma, düz anlatım, tanımlar yardımıyla öğretim ve problem çözme yöntemi olduğu belirtilmiştir. Ancak ilgili literatür incelendiğinde öğrencilerin bu tekniklere verdiği önem sırasına ilişkin görgül bir araştırma bulgusuna rastlanamamıştır. Araştırmanın problem cümlesi matematik dersinde kullanılan öğretim yöntem ve tekniklerin öğrenci gözüyle önem sırasının nasıl olduğunun belirlenmesidir.

Araştırmanın Amacı

Bu çalışmanın amacı öğrenci görüşlerine göre hangi yöntemlerin birbirine benzer düzeyde rağbet gördüğü; hangi tekniklerin daha çok veya az rağbet gördüğünün belirlenmesidir.

YÖNTEM

Çalışma ilgili fakülte ve yüksekokullarda ders vermekle yükümlü öğretim elemanları tarafından belirlenmiş olan öğretim yöntem ve tekniklerinden en çok kullanılan sekiz farklı yöntem ve tekniği öğrenci yargılarına göre kendi içinde ikili karşılaştırma esasına dayanmaktadır. Thurstone (1927) tarafından geliştirilen karşılaştırmalı yargılar kanunu temel alınarak yapılan analizlerde öğrenciler her bir yöntemi ikili karşılaştırarak kendileri ve bilişsel gelişimleri için hangi yöntemi daha yararlı buluyorlarsa ilgili matriste üstün olan yönteme 1, diğerine 0 puan vererek kodlama yapmışlardır. Bu bilgiler kullanılarak ikili karşılaştırma yapılan tekniklerin ölçek değerleri elde edilmiştir. Bu araştırmanın amacı var olan durumun ne olduğunu ortaya çıkarmak olduğundan betimsel bir çalışmadır.

Örneklem/Çalışma Grubu/Katılımcılar

Araştırmanın çalışma grubunu Ege Bölgesinde yer alan bir üniversitenin farklı fakülte ve yüksekokullarında öğrenim gören öğrenciler oluşturmaktadır. Uygulama 2013–2014 eğitim-öğretim yılı bahar yarıyılında yapılmıştır. Çalışma gönüllülük esasına dayalı olarak, uygulamaya katılmak isteyen öğrenciler ile gerçekleştirilmiştir. Kolayda örnekleme yöntemiyle belirlenen 320 öğrenciye uygulama formu dağıtılmıştır. Dağıtılan formlardan 10 tanesi eksik doldurulması sebebiyle değerlendirilmeye alınmamıştır.

Veri Toplama Araçları

Öğrencilerin belirlenen 8 farklı yöntem ve teknikten ikili karşılaştırma yaparak hangisini diğerine kıyasla faydalı bulduklarını belirlemek amacıyla 8x8 türünde kare matris oluşturulmuştur. Matrisin alt ve üst köşegen matrisinde benzer karşılaştırmaları iki defa yapmamak amacıyla alt köşegen matris kullanılarak karşılaştırma yapmaları istenmiştir (Ek-1). Geliştirilen matriste yer alacak yöntem ve tekniklerin neler olacağını belirlemek amacıyla sayısal ve sözel ağırlıklı bölümlerde görev yapan öğretim elemanları ve öğretim üyelerinden oluşan 20 kişilik bir ekipten uzman görüşü alınmıştır. Uzman görüşleri doğrultusunda frekans değeri en yüksek 8 farklı öğretim yöntem ve tekniği matris üzerine yerleştirilerek uygulama formuna son şekli verilmiştir. Ayrıca uygulama formunun arka yüzüne her bir yöntem ve tekniğin açıklaması yazılmıştır.

İşlem

Bu çalışmada araştırmanın ana problemine bağlı kalınarak gerçekleştirilen işlemler aşağıdaki belirtildiği sırada yapılmıştır.

1. İkili karşılaştırma matrisinde yer alacak yöntem ve tekniklerin belirlenmesi amacıyla uzman görüşüne başvurulmuştur. Matematik dersi için belirlenen yöntem ve tekniklerin neler olduğu konusunda 24 uzmanın görüşü alınmıştır.
2. Uzman görüşleri doğrultusunda son şekli verilen ölçme aracını uygulamadan önce uygulama yapılacak tüm öğrencilere her bir yöntem ve teknik örnekler verilerek 2 ders saati süresince açıklanmıştır.
3. Uygulamada öğrencilere matematik dersinde öğretim elemanları tarafından en çok kullandıklarını belirtikleri ilk sekiz yöntem ve teknik konusunda hazırlanmış Powerpoint sunusu ve basılı materyaller ile 2 saatlik bir eğitim verilmiştir
4. Açıklamaların ardından araştırmacının bizzat kendisi tarafından uygulama yönergesi açıklanarak örnek bir uygulama sınıfla beraber yapılmıştır.
5. Uygulamanın ardından öğrencilere 30 dakika süre verilerek yöntem ve teknikleri ikili karşılaştırma yapmaları istenmiştir.

Matrisin 8x8 türünde kare matris ve alt üçgen üzerinde işaretleme yapılması ve toplam gözenek sayısının 32 olması sebebiyle her hücreye en az 5 veri girmesi (32*5 = 160) amacıyla katılımcı sayısı yüksek tutulmuştur.

Verilerin Analizi

Verilerin analizinde psikolojik özelliklerin ölçeklenmesinde yaygın olarak kullanılan tekniklerinden ikili karşılaştırma yöntemiyle ölçekleme tekniğinden yararlanılmıştır (Thurstone, 1927; Turgut ve Baykul, 1992). Thurstone (1927) tarafından geliştirilen karşılaştırmalı yargılar kanunu temel alınarak yapılan analizlerde öğrenciler her bir yöntemi ikili karşılaştırarak kendileri ve bilişsel gelişimleri için hangi yöntemi daha yararlı buluyorlarsa ilgili matriste üstün olan yönteme 1, diğerine 0 puan vererek kodlama yapmışlardır. Bu bilgiler kullanılarak ikili karşılaştırma yapılan tekniklerin ölçek değerleri elde edilmiştir. Araştırmada elde edilen verilerin analizine önce V. Hal yöntemiyle ölçekleme yapılmış, hesaplanan hata ve χ^2 değerinin .05 anlamlılık düzeyinde beklenen değerden büyük çıkması nedeniyle III. Hal denklemiyle ölçekleme yapılmıştır. Ölçeklemede ikili karşılaştırmada kullanılan temel formüllerden yararlanılmıştır (Turgut ve Baykul, 1992). Öğrencilerin vermiş oldukları yanıtlarda çelişkili üçlüler olup olmadığı ki kare değerleri ile kontrol edilerek yanıtlarında çelişkili üçlü bulunan öğrencilere ait anket sonuçları analiz kapsamına alınmamıştır.

BULGULAR

Çalışmada öğrencilerden matematik derslerinde kendileri için daha faydalı olduğunu düşündükleri yöntem ve teknikleri ikili karşılaştırma yaparak belirtmeleri istenmiştir. Öğrencilerin bu ikili karşılaştırma sonucu verdikleri tepkilere ilişkin frekans değerleri belirlenmiş ve sonuçlar Tablo 1'de verilmiştir.

Tablo 1. Öğretim Yöntem ve Tekniklerine İlişkin Frekans Matrisi (F)

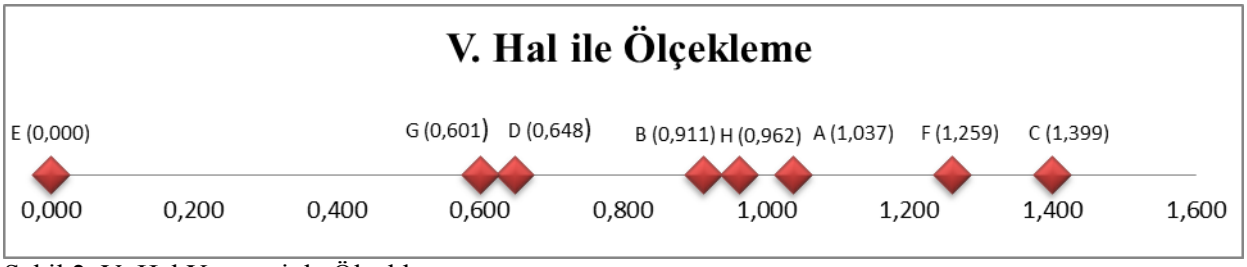
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
A		90	189	116	51	194	117	161
B	220		180	121	44	191	113	172
C	121	130		58	38	150	78	128
D	194	189	252		49	203	120	195
E	259	266	272	261		253	237	257
F	116	119	260	44	57		57	94
G	193	197	232	190	73	253		177
H	149	138	182	115	53	216	133	

Frekans matrisi öğrencilerin tercih ettikleri öğretim yöntem ve tekniklerinin tercih edilme sıklıklarına göre oluşturulmuştur. Tabloda görüldüğü üzere B yöntemi ile A yöntemi karşılaştırıldığında çalışma grubunda bulunan 310 öğrencinin 220'si B yöntemini kendileri için daha faydalı bulmalarına rağmen 90'ı A yöntemini daha faydalı bulduklarını belirtmişlerdir. Frekans Matrisi tüm tekniklerin ikili karşılaştırılmasına dayanan sonuçlardan oluşturulmuştur. Frekans matrisinde esas köşegen üzerindeki hücrelerde her yöntem kendisiyle karşılaştırma yapıldığı anlamına geldiğinden bu hücreler boş bırakılmıştır. Frekans matrisi oluşturulduktan sonra her bir hücrede yer alan değerlerin toplam öğrenci sayısı olan 310'a bölünmesiyle oranlar matrisi elde edilmiştir. Oranlar matrisine ilişkin sonuçlar Tablo 2'de verilmiştir.

Tablo 2. Öğretim Yöntem ve Tekniklerine İlişkin Oranlar Matrisi (P)

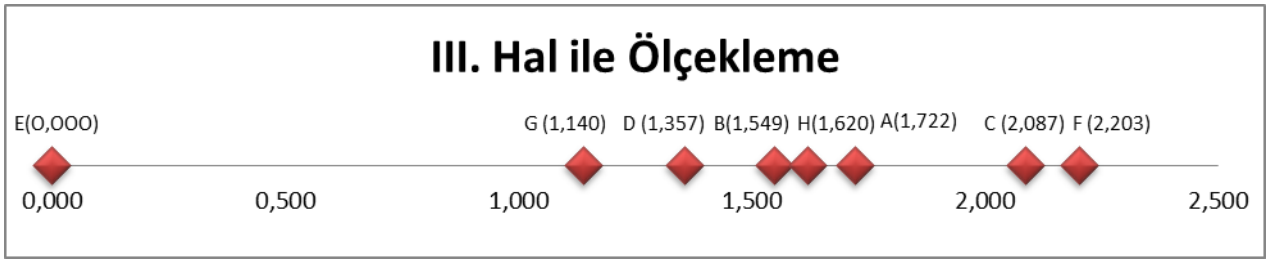
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
A		,29	,61	,37	,16	,63	,38	,52
B	,71		,58	,39	,14	,62	,36	,55
C	,39	,42		,19	,12	,48	,25	,41
D	,63	,61	,81		,16	,65	,39	,63
E	,84	,86	,88	,84		,82	,76	,83
F	,37	,38	,84	,14	,18		,18	,30
G	,62	,64	,75	,61	,24	,82		,57
H	,48	,45	,59	,37	,17	,70	,43	

Oluşturulan tabloda esas köşegene simetrik olan elemanların toplamı 1 olduğu görülmektedir. Daha sonra oranlar matrisindeki hücrelerde yer alan değerlere karşılık gelen birim normal değerler (z değerleri) elde edilmiştir. Z Matrisinde sütun toplamları bulunduktan sonra bu değerler sekiz farklı yöntem ve teknik olduğundan 8'e bölünerek S_j değerleri elde edilmiştir. Hesaplanan S_j değerlerinden en küçük olanı (-.495) sıfır yapmak amacıyla her bir S_j değerine bu değerin mutlak değeri olan .495 eklenerek her bir yöntem ve tekniğin ölçek değeri belirlenmiştir. Her bir yöntem ve tekniğe ilişkin ölçek değerleri Şekil 2'de gösterilmiştir.



Şekil 2. V. Hal Yöntemiyle Ölçekleme

Çalışmada önce V. Hal Yöntemiyle ölçekleme yapılmış ancak hesaplanan ortalama hata değeri 0,386 ve hesaplanan χ^2değerinin 21 serbestlik derecesinde %95 güven derecesinde beklenen değerden büyük olması sebebiyle verilerin uyum ölçütlerini karşılamaması nedeniyle verilere III. Hal denklemi yardımıyla ölçekleme yapılmasına karar verilmiştir. İki yöntem arasındaki fark ise ölçekleme yapılırken V. Hal denkleminde ayırt etme yargılarının varyansı eşit kabul edilirken III. Hal denklemi ayırt etme yargıları varyansını eşit kabul etmemekte ve her bir yargı için tek tek varyansları hesaplaması sebebiyle kullanılmaktadır (Turgut ve Baykul, 1992). Hem V. Hal hem de III. Hal yöntemi yardımıyla yapılan ölçekleme çalışmaları sonucunda uygulanan öğretim yöntem ve tekniklerin sıralarında büyük bir değişiklik olmadığı görülmüştür.



Şekil 3. III. Hal Yöntemiyle Ölçekleme

Çalışmada V. Hal denklemi yardımıyla elde edilen ve III. Hal denklemi yardımıyla elde edilen ölçek değerleri karşılaştırıldığında sıralama anlamında en iyi iki yöntem olarak kabul edilen Soru cevap (C) ile Gösterip yaptırma yöntemi sıralarının yerleri değişmiştir. Bunun dışında diğer yöntem ve tekniklerin sıralarında bir değişme olmamıştır. Ölçek değeri aralıklarına bakıldığında III. Hal denklemi yardımıyla elde edilen ölçek değeri aralıklarının V. Hal denklemine göre biraz daha genişlediği görülmüştür. Bunun yanında bazı yöntem ve tekniklerin birbirinden daha iyi ayrılabildiği belirlenmiştir. Bu sonuçlara göre öğrenci görüşlerine göre en faydalı olduğu düşünülenden en az faydalı olduğu düşünülen yönteme doğru sıralama değerleri Tablo 3'de hem V. Hal hem de III. Hal Yöntemiyle elde edilen ölçek değerleri için gösterilmiştir.

Tablo 3. V. ve III. Hal Yöntemiyle Elde Edilen Değerlerin Karşılaştırılması

	V. HAL YÖNTEMİ		III. HAL YÖNTEMİ	
Sıra No	Yöntemin Adı	Ölçek Değeri	Yöntemin Adı	Ölçek Değeri
1	Soru – Cevap (C)	1,399	Gösterip Yaptırma (F)	2,203
2	Gösterip Yaptırma (F)	1,259	Soru – Cevap (C)	2,087
3	Anlatım (A)	1,037	Anlatım (A)	1,722
4	Problem Çözme (H)	0,962	Problem Çözme (H)	1,620
5	Beyin Fırtınası (B)	0,911	Beyin Fırtınası (B)	1,549

6	Tartışma (D)	0,648	Tartışma (D)	1,357
7	Grup Çalışması (G)	0,601	Grup Çalışması (G)	1,140
8	Rol Oynama (E)	0,000	Rol Oynama (E)	0,000

Tablo incelendiğinde soru cevap ve gösterip yaptırma yönteminin öğrenciler için matematik derslerinde en faydalı yöntemler olarak görüldüğü ve ardından sırasıyla anlatım, problem çözme, beyin fırtınası, tartışma, grup çalışması ve rol oynama tekniklerinin yer aldığı görülmektedir.

SONUÇLAR ve TARTIŞMA

Bu çalışmada üniversite öğrencilerinin matematik dersinde akademik başarıları bakımından en faydalı olacağını düşündükleri 8 farklı öğretim yöntem ve tekniği ikili karşılaştırmalar yaparak ölçeklendirmeleri amaçlanmıştır. Yargıcı kararlarına dayalı ölçekleme tekniklerinden biri olan ikili karşılaştırmalar yöntemiyle öğrencilerin matematik dersinde kendileri için en faydalı olduğunu düşündükleri yöntemleri karşılaştırmaları istenmiştir (Kan, 2008). Elde edilen sonuçlara göre öğrenciler soru cevap ve gösterip yaptırma yöntemini matematik derslerinde başarılı olabilmeleri için en önemli iki yöntem olarak görmektedir. Bu yöntemleri sırasıyla anlatım, problem çözme, beyin fırtınası, tartışma ve rol oynama teknikleri takip etmektedir. Bu sonuç Acar Güvendir ve Özkan (2013) tarafından yapılan çalışmanın sonuçlarıyla benzerlik göstermektedir. Araştırmacılar eğitim fakültesi öğrencilerinden ölçme ve değerlendirme dersini tekrar almaları mümkün olduğunda hangi yöntem ve teknikle dersin işlenmesini istediklerini belirtmeleri istenmiş ve araştırma sonucunda elde edilen bulgulara göre sırasıyla örnek olay, gösterip yaptırma, problem çözme, tartışma, bireysel çalışma ve anlatma tekniğini yer almaktadır.

Öğrenciler tarafından genel olarak zor bir ders olarak görülen matematik dersinde (Peker ve Mirasyedioğlu, 2003) öğrenciler sırasıyla gösterip yaptırma, soru cevap, anlatım, problem çözme, beyin fırtınası, tartışma, grup çalışması ve rol oynama öğretim yöntem ve tekniklerini kendileri için faydalı bulmaktadırlar. Çalışmada grup çalışması yönteminin son sıralarda yer alması öğrencilerin derslerinde daha çok bireysel çalışmalara yer verildiğinin bir göstergesi olarak yorumlanabilir. Acar Güvendir ve Özkan (2013) tarafından yapılan çalışmada bireysel çalışma yöntemi en az tercih edilen yöntem olduğu göz önüne alındığında öğrencilerin özellikle matematik dersinde gösterip yaptırma yönteminin en çok tercih edilen yöntem olması öğrencilerin dersi vermekle yükümlü öğretim elemandan büyük bir beklenti içerisinde olduklarını göstermektedir. Öğrencilerle uygulama sonrası yapılan görüşmelerde öğrencilerin önce öğretmen veya öğretim elemanın konuyu anlatıp, konuyla ilgili farklı örnekleri çözmeleri sonrasında kendilerinin konuyla ilgili örnekleri çözmek istedikleri belirlenmiştir. Bunun yanında öğrencilerin rol oynama (drama) tekniğini matematik dersi için kendileri için en az faydaya sahip olduğu bir yöntem olarak gördükleri belirlenmiştir. Nitekim matematik derslerinde bu yönteme yer verilmemesi bu sonucun doğal bir nedeni olarak düşünülmektedir. İlgili alan yazında öğretim yöntem ve tekniklerinin öğrenci görüşlerine göre karşılaştırılmasına ilişkin görgül bir araştırma bulgusuna rastlanamamıştır. Çalışma bu yönüyle özellikle matematik dersini vermekle yükümlü öğretmen ve öğretim elemanlarına önemli sonuçlar vermektedir. Özellikle derslerde soru-cevap, gösterip yaptırma, anlatım ve problem çözme tekniklerine başvurulması öğrenciler tarafından istenilen bir yaklaşımdır. Bu nedenle ders vermekle yükümlü öğretmen ve öğretim elemanlarının bu yöntemleri kullanmaları derse ilişkin ilgi ve dikkati artırmada etkili olacağı düşünülmektedir. Öğrencilerin ilgi ve dikkatlerinin artırılması sayesinde ders başarılarının da dolaylı olarak artacağı düşünülmektedir.

Ölçekleme yöntemleri ile ilgili çalışma yapacak araştırmacılara ikili karşılaştırmalar yöntemiyle elde ettikleri sonuçların özellikle sıralama yargıları kanunuyla ölçekleme yöntemleriyle elde ettikleri sonuçları ile tutarlılığını araştırmaları ve sonuçları raporlamaları önerilmektedir.

KAYNAKÇA

Acar Güvendir, M. ve Özer Özkan, Y. (2013). İki Ölçekleme Yönteminin Karşılaştırılması: İkili Karşılaştırma ve Sıralama Yargıları. *Eğitim Bilimleri Araştırmaları Dergisi*, 3 (1), 105-119

Baykul, Y. ve Turgut, M. F. (1992) *Ölçekleme Teknikleri,* Ankara: Meteksan Anonim Şirketi

Büyükkaragöz, S. ve Çivi, C. (1996), *Genel Öğretim Metotları,* İstanbul: Öz Eğitim Yayınları

Covill, A. (2011) College Students Perceptions of the Traditional Lecture Method. *College Student Journal*, 45 (1), 2-15

Demirel, Ö. (1999)*Türk Eğitim Sisteminde Öğretim Programlarının Geliştirilmesinde Bilimsel Yaklaşım ve 2000'li Yıllar İçin Öneriler*, Eğitimde Yansımalar: 21. Yüzyılın Eşiğinde Türk Eğitim Sistemi Ulusal Sempozyumu, 25-27 Kasım. 328 –335

Demirel, Ö. (2011). *Öğretme Sanatı.* Ankara: Pegem Akademi.

Ergani, K. (2010) *İlköğretim 4. ve 5. Sınıf Sosyal Bilgiler Dersi Öğretim Yöntem ve Teknikleri İle Materyal Kullanımına İlişkin Öğretmen Görüşleri*, Yayınlanmış Yüksek Lisans Tezi, Dumlupınar Üniversitesi Sosyal Bilimler Enstitüsü, İlköğretim Anabilim Dalı, Kütahya.

Gökçe, F. (1994). Eğitimde Denetimin Amaç ve İlkeleri, *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 10, 73-78.

Gönen, S., ve Kocakaya, S. (2006), Fizik Öğretmenlerinin Hizmet İçi Eğitimler Üzerine Görüşlerinin Değerlendirilmesi, *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, 19, 37-44,

Hesapcıoğlu, M. (2011). *Öğretim İlke ve Yöntemleri,* Ankara: Nobel Yayınları.

İşman, A. ve Eskicumalı, A. (1999). *Eğitimde Planlama ve Değerlendirme.* Adapazarı: Değişim Yayınları.

Jungst, S.,Licklider, B. Ve Wiersema, J. (2003). Providing support for faculty who wish to shift to a learning-centered paradigm in their higher education classrooms, *The Journal of Scholarship of Teaching and Learning*, 3, 69-81.

Ishiyama, J.T., McClure, M., H. ve Amico, J. (1999) Critical thinking disposition and locus of control as predictors of evaluations of teaching strategies, *College Student Journal*, 33 (2), 10-44.

Kan, A. (2008) Psikolojik Değişkenleri Ölçmek İçin Kullanılan Ölçekleme Yaklaşımları Üzerine Bir Karşılaştırma, *Eğitimde Kuram ve Uygulama,* 4 (1), 2-18.

Kayabaşı, Ş. ve Erdoğan, Y. A. (2002) *Öğretimi Planlama ve Değerlendirme*, Ankara: Anı Yayınevi.

Kahyaoğlu, M ve Yangın, S. (2007) İlköğretim Öğretmen Adaylarının Mesleki Öz Yeterliklerine İlişkin Görüşleri, *Kastamonu Eğitim Dergisi*, 15, 83-105.

Karaağaçlı, M. (2005), *Öğretimde Yöntemler ve Yaklaşımlar*, Ankara: Pelikan Yayıncılık.

Kılıç, C. (2010) *İlköğretim Okullarında (Devlet-Özel) ve Dershanelerde Görev Yapan Fen ve Teknoloji Öğretmenlerinin Kullandıkları Öğretim Yöntem ve Teknikleriyle İlgili Öğrenci Görüşleri,* Yüksek Lisans Tezi, Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, İlköğretim Bölümü Fen Bilgisi Öğretmenliği Bilim Dalı, Ankara.

Küçükahmet, L. (1983) *Öğretim İlke ve Yöntemleri*, Ankara Üniversitesi Eğitim Bilimleri Fakültesi Yayınları No: 124.

Küçükahmet, L. (1995) *Öğretim İlke ve Yöntemleri*, Ankara: Gazi Büro Kitabevi.

Larson R. W. ve Richards, M. H. (1991) Daily companionship in late childhood and early adolescence: Changing developmental contexts, *Child Development*, 62, 284–300

Marbach-Ad, G., Seal, O. ve Sokolove, P. (2001). Student Attitudes and Recommendations On Active learning, *Journal of College Science Teaching*, JO, 434-438.

Murphy, B. C., Eisenberg, N., Fabes, R. A., Shepard, S. ve Guthrie, I. K. (1999) Consistency and change in children's emotionality and regulation: A longitudinal study. Merrill-Palmer Quarterly: *Journal of Developmental Psychology*. 45, 413–444.

Önen, F., Mertoğlu, H., Saka, M. ve Gürdal, A. (2009) Hizmet içi Eğitimin Öğretmenlerin Öğretim Yöntem ve Tekniklerine ilişkin Bilgilerine Etkisi: ÖPYEP Örneği, *Ahi Evran Üniversitesi Eğitim Fakültesi Dergisi,* 10 (3), 9-23.

Önen, F., Saka, M., Erdem, A., Uzal, G. ve Gürdal, A., (2008) Hizmet İçi Eğitime Katılan Fen Bilgisi Öğretmenlerinin Öğretim Tekniklerine İlişkin Bilgilerindeki Değişimin Tespiti: Tekirdağ Örneği, *KEFAD*, 9 (1), 45-57

Özden, Y. (2000) *Öğrenme ve Öğretmen*, Pegem A Yayıncılık, Ankara.

Peker, M. ve Mirasyedioğlu, Ş. (2003) Lise 2. Sınıf Öğrencilerinin Matematik Dersine Yönelik Tutumları ve Başarıları Arasındaki İlişki, *Pamukkale Eğitim Fakültesi Dergisi*, 2 (14), 157-166.

Saban, A. (2004) *Öğrenme Öğretmen Süreci Yeni Teori ve Yaklaşımlar*, Nobel Yayınevi, Ankara.

Tan, Ş., Kayabaşı, Y ve Erdoğan, A. (2002) *Öğretimi Planlama ve Değerlendirme (3. Baskı),* Ankara: Anı Yayıncılık.

Taşpınar, M. (2005) *Kuramdan Uygulamaya Öğretim Yöntemleri*, Ankara: Nobel Yayınevi.

TDK (2014). Türk Dil Kurumu Genel Türkçe Sözlük. Erişim Tarihi: 25.05.2014.

Thurstone, L. L. (1927) The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology*, 21, 384-400.

Tohumcu, T. (2004) *Adıyaman Merkez İlköğretim Okulları 5. Sınıf Öğrencilerinin Matematik Dersindeki Başarıları ile Bu Öğrencilerin Sınıf Öğretmenlerinin Öğretim Yöntemleri Arasındaki İlişkinin İncelenmesi,* Yayınlanmış Yüksek Lisans Tezi, İnönü Üniversitesi Sosyal Bilimler Enstitüsü, Malatya.

Turgut, M. F. ve Baykul, Y. (1992) *Ölçekleme Teknikleri,* Ankara: ÖSYM Yayınları.

Uysal, A. (2010) *Sınıf Öğretmenlerinin 2009 Hayat Bilgisi Öğretim Programında Belirtilen Strateji, Yöntem ve Teknikleri Uygulamadaki Yeterlilik Düzeylerinin Belirlenmesi,* Yayınlanmış Yüksek Lisans Tezi, Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Vural, B. (2006). *Eğitim-öğretimde planlama-ölçme ve stratejiler,* İstanbul: Hayat Yayıncılık.

Yeşilyurt, E. (2013) Öğretmenlerin Öğretim Yöntemlerini Kullanma Amaçları ve Karşılaştıkları Sorunlar, *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi,* 17 (1), 163-188.

Yıldız, S. (2008) *Özel Eğitim Sınıflarında Çalışan Sınıf Öğretmenlerinin Matematik Öğretiminde Kullanılan Öğretim Yöntemlerine İlişkin Görüşlerin Değerlendirilmesi,* Yayınlanmış Yüksek Lisans Tezi, Selçuk Üniversitesi Sosyal Bilimler Enstitüsü, Konya

EXTENDED ABSTRACT

Introduction

Today's teachers are expected to be good administrators, observers and qualified guides who can organize teaching-learning processes. Therefore, teachers have now been considered as a professional group requiring more qualifications and competence (Gökçe, 2000). Contrary to classical understanding, the teachers are both expected to use teaching methods and techniques in the best manner and to integrate them with learning environments and more modern classroom management, which is required by contemporary education. Today teachers are expected to use different teaching methods and techniques, however the teachers were observed to tend to use the methods and techniques in which they have a central role, direct the course and the students, make the evaluation themselves and especially the students have the role of passive receivers (Marbach, Seal & Sokolove 2001, Junst, Licklider & Wiersema 2003, Covill 2011). According to previous research, (Önen, Saka, Erdem, Uzal and Gürdal, 2008) teachers from different branches generally prefer teacher-centered techniques and methods in their lessons as they do not have adequate knowledge and skills about teaching methods and techniques. Research has shown that the students begin to get distracted and bored in time while listening to the teacher and while reading (Larson and others, 1991: 431 and Lammersand Murphy 2002). The lesson can be more fast-moving, more time-saving and more fun through the methods and techniques used by the teacher according to the content of the subject (Ergani, 2010). It can be stated that it is of great importance to organize the lesson in such a way to lead the students into knowledge and to use different methods-techniques in this process, instead of directly transferring knowledge. A wide range of methods and techniques have been used in teaching environment. Some of these methods are lecture, discussion, example case, demonstration and practice, problem-solving, creative drama and individual study (Demirel, 2009). The most commonly used techniques in teaching environment are classified as group teaching techniques (brain storming, question-answer, demonstration, simulation, pair and group work, micro teaching, educational games, six thinking hats), individual teaching techniques (individualized teaching, programmed teaching, computer assisted teaching) and extra-classroom teaching techniques (interview, observation, exhibition, project, homework).

Method

The study is based on the principle of comparison of eight different most commonly used teaching methods and techniques determined by the lecturers lecturing at the related faculties and high schools according to the judgments of the students using pairwise comparison method. Analyses were performed using law of comparative judgements developed by Thurstone (1927). The students were requested to conduct pairwise comparison for each method and to code them by giving 1 point to the superior method and 0 to the other method in the related matrix according to what method they

found more beneficial for themselves and their cognitive development. Scale values were obtained for the techniques that underwent pairwise comparison using that information. This is a descriptive study as it aims to determine an existing situation. The study was carried out in spring semester of 2013-2014 academic year with the students who volunteered to participate in the study. A total of 320 students determined by convenience sampling were distributed the application form. Of the distributed forms, 10 were excluded from evaluation as they were incomplete. 8x8 square matrix was formed to determine which method the students find more beneficial when compared to the others. Views of a team of 20 experts consisting of the teaching staff and faculty members in numerical and verbal departments were taken to determine the methods and techniques that will be included in the developed matrix. Based on the expert views, the application form was designed by placing 8 different teaching methods and techniques with the highest frequency value on the matrix. Furthermore, explanations for each method and technique were written at the back page of the application form.

Results and Discussions

The purpose of this study was to have university students scale 8 different teaching methods and techniques that they believe will be the most beneficial in terms of academic success in mathematics using pairwise comparison technique. The students were requested to compare the methods which they believe are the most beneficial for themselves in mathematics course using pairwise comparison method, which is a scaling technique based on judgment decisions (Kan, 2008; 4). According to the results, the students considered question-answer and demonstration and practice methods as the most important two methods to be successful in mathematic lessons, followed by lecture, problem solving, brain storming, discussion and role play techniques. This result is consistent with the results of Acar, Güvendir and Özkan (2013). The researchers requested the students to state which method they would like the course to be instructed if it was possible for them to retake measurement and evaluation course. It was found that the students preferred example case, demonstration and practice, problem-solving, discussion, individual study and lecture techniques respectively. In mathematics course, which is generally regarded as a difficult lesson by the students (Peker and Mirasyedioğlu, 2003), demonstration and practice, question-answer, lecture, problem solving, brain storming, discussion, group study and role-play teaching method and techniques were found to be beneficial. The fact that group study ranked the last can suggest that the lessons mostly include individual studies. In a study carried out by Güvendir and Özkan (2013) the fact demonstration and practice method was the most preferred method while individual study method was the least preferred method especially in mathematics course indicates that the students have great expectations from the lecturer. Interviews with the students after the application revealed that the students preferred first the teacher or the lecturer to give a lecture, solve relevant examples and then to solve relevant examples by themselves.

EK-1: Veri Toplama Aracı

	1. Anlatım Tekniği	2. Beyin fırtınası	3. Soru-cevap	4.Tartışma	5. Rol oynama ya da Drama	6. Gösterip yaptırma	7. Grup çalışması	8. Problem çözme
1. Anlatım Tekniği								
2. Beyin fırtınası								
3. Soru-cevap								
4.Tartışma								
5. Rol oynama ya da Drama								
6. Gösterip yaptırma		:)						
7. Grup çalışması								
8. Problem çözme								

NOT: Yukarıdaki matriste yatayda (satır) yer alan öğretim yöntem ve teknikleri ile dikeyde (sütun) aynı yöntem ve teknikler bulunmaktadır. Sizden istenilen her bir yöntem ve tekniği ikili karşılaştırma yapmanızdır. Eğitim öğretim ortamında kullanılan yöntemlerden hangisinin diğerine orana daha faydalı (yararlı) olduğunu düşünüyorsanız üstün olduğunu düşündüğünüz tekniği kesişim noktasına yazmanız gerekir. Örneğin 6. Satırdaki gösterip yaptırma ile 2. sütundaki beyin fırtınası karşılaştıran bir öğrenci iki yöntemin kesiştiği noktaya (tabloda gülücük olan hücre) eğer satırdaki teknik (gösterip yaptırma) daha yararlı ise "**gösteri**"; eğer sütundaki teknik (beyin fırtınası) daha yararlı ise "**beyin**" yazması gerekmektedir. Benzer şekilde boş olan tüm hücreler için aynı işlemi yapmanız gerekmektedir. Ayrıca siyah alanlara işaretleme yapmayınız. Özetle satırlarda yer alan yöntemleri daha faydalı bulunlar satırdaki yöntemin adını; ilgili kesişim noktasındaki sütunu faydalı bulanlar sütunda yer alan yöntemin adını yazacaklardır.

İlginiz, sabrınız ve katkılarınız için şimdiden teşekkür ederim. Gökhan AKSU

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2015; 6(2); 293-312

Test Boyutluluğunun Analizinde Kullanılan Programların İncelenmesi

An Overview of Software for Assessing Test Dimensionality

Güler YAVUZ * Nuri DOĞAN**

Öz
Son yıllarda çok boyutlu veri setlerine ve çok boyutlu veri setleriyle yapılan çeşitli test uygulamalarına (test eşitleme, alt test puanlama, değişen madde fonksiyonu gibi) olan ilgi önemli düzeydedir. Bu çalışmada test boyutluluğunun analizinde ve çok boyutlu veri setleriyle çeşitli test uygulamalarında yaygın kullanılan bazı popüler paket programları (Mplus, NOHARM, TESTFACT, IRTPRO, flexMIRT, BMIRT, DIMTEST, DETECT, CCPROX/HAC, Velicer' in MAP testi ve Paralel Analiz) incelenmiştir. Bu programların ulaşılabilirliği, hangi tür veri setleri, modeller, kestirim teknikleri ile analiz yapabildikleri belirlenmiştir. Programların boyutluluk analizini hangi tür yaklaşımla (açımlayıcı, doğrulayıcı) yaptıkları ve program çıktılarında bulunan indekslerin nasıl yorumlandıkları incelenmiştir. Ayrıca programların çeşitli özellikleri açısından karşılaştırılmalarına yönelik yapılmış çalışmalara yer verilmiştir.

Anahtar Kelimeler: boyutluluk, Mplus, IRTPRO, flexMIRT, BMIRT

Abstract
In recent years, there is substantial demand for multidimensional data sets and their various test applications (such as equating test scores, subtest score reporting, differential item functioning). In this study, some popular software packages (Mplus, NOHARM, TESTFACT, IRTPRO, flexMIRT, BMIRT, DIMTEST, DETECT, CCPROX/HAC, Velicer's MAP test and Parallel Analysis) which are used for analyzing test dimensionality and used for various test applications with multidimensional data sets, was investigated. The software packages was investigated in terms of their availability, their capabilities of data sets, models, estimation techniques. The software packages' dimensionality analyses approach (confirmatory, exploratory) was investigated. Also assessing of the indices which can be gained from software packages' outputs was investigated. Finally the simulation studies in which the packages are compared in terms of their various properties was reviewed.

Key Words: dimensionality, Mplus, IRTPRO, flexMIRT, BMIRT

GİRİŞ

Testlerin veya testlerde bulunan maddelerin sadece tek bir örtük özelliği ölçtüğünü varsaymak ve bu varsayım altında ölçmeler yaparak, bireyler hakkında kararlar vermek 1980' lerden sonra daha da tartışılır hale gelmiştir (Ackerman, 1989; Ansley ve Forsyth, 1985; Drasgow ve Parsons, 1983; Harrison, 1986; Way, Ansley ve Forsyth, 1986). Özellikle 1990'lardan itibaren çok sayıda araştırmacı (Sireci, Thissen, ve Wainer, 1991; Yen, 1993; Bradlow, Wainer, ve Wang, 1999; Chen ve Thissen, 1997; De Champlain, 1996; Shealy ve Stout, 1993; Walker ve Beretvas, 2003) çok boyutlu veri setlerini tek boyutlu model uygulamalarıyla incelemiştir. Bu çalışmaların büyük bir çoğunluğundan çıkan sonuçlara dayanarak, araştırmacılar çok boyutlu madde tepki kuramı model, program ve uygulamalarının geliştirilmesi gerektiğini belirtmişlerdir (Ackerman, 1996; Adams, Wilson ve Wang, 1997; Embretson, 1997; McDonald, 1997; Reckase, 1997; Beguin & Glas, 2001; Bolt ve Lall, 2003; Ackerman, Gierl ve Walker, 2003; Walker ve Beretvas, 2003; Yao ve Boughton, 2009; Reckase, 2009).

* Yrd. Doç.Dr., Adıyaman Üniversitesi, Eğitim Fakültesi, Adıyaman-Türkiye, e-posta: gyavuz2010@gmail.com
**Doç.Dr.,Hacettepe Üniversitesi, Eğitim Bilimleri Bölümü, Ankara-Türkiye, e-posta: nurid@hacettepe.edu.tr

Geliş Tarihi: 09.09.2015
Kabul Tarihi: 11.20.2015

Özellikle son yıllarda, test uzmanlarının ve test merkezlerinin odağında başarı ve yetenek testlerine ait çok boyutlu veri setleri ile dikey ölçekleme, test eşitleme, alt test puanlama gibi temel uygulamalara ilişkin daha doğru analizler yapma çabası bulunmaktadır. Test uygulamalarıyla elde edilen sonuçların doğruluğu kullanılan programlarla yakından ilişkilidir. Araştırmacılar ve uygulayıcılar, program seçimini test koşullarına uygun yapmak durumundadır. Farklı modellerle ve test koşulları ile analiz yapabilen programların sahip oldukları kestirim teknikleri sonuçların doğruluğu üzerinde özellikle belirleyicidir.

Bu durum, verilerin boyutluluğunu belirleyebilen, çok boyutlu veri setleriyle çeşitli test uygulamaları yapma olanağı tanıyan, madde ve yetenek parametrelerine yönelik farklı kestirim algoritmalarına ve yollarına sahip bilgisayar programlarının geliştirilmesine yönelik de önemli bir araştırma alanı ortaya çıkarmaktadır.

Günümüzde hızla gelişen bilgisayar teknolojisinden dolayı ve her gün daha iyi ve yeni bir program ve kestirim yöntemi elde etmek mümkün olabildiğinden kalıcı bir program listesi yapmak kolay değildir. Bilgisayar programlarına ilişkin listenin sürekli güncellenmesi gerekmektedir. Bu sebeple çeşitli amaçlarla çok boyutlu modellerle ve veri setleriyle analiz yapmak isteyen araştırmacıların iyi bir alanyazın araştırması yapması gerekmektedir. Boyutluluk analizinde kullanılabilen istatistik programları arasında önemli farklılıklar bulunmaktadır. Herhangi bir araştırmacı bu programlardan biriyle çalışmadan önce programların kullandığı kestirim tekniği, kestirim süresi, yazılım dili, maliyet, ne tür veri setleriyle analiz yapılabildiği, hangi modellerin kullanılabildiği vb. gibi çok sayıda avantaj ve dezavantajları bilmek durumundadır.

Araştırmanın Amacı

Bu çalışmanın amacı, çeşitli disiplinlerle çalışan araştırmacılara ve uygulayıcılara boyutluluğun değerlendirilmesinde kullanılabilen mevcut paket programların ve yöntemlerin tanıtılmasıdır. Bu araştırmada ilk olarak programa erişilebilme yolları, programın kullanımına yönelik dökümanların var olup olmadığı, varsa neler olduğu ele alınmış, ardından programların teknik özellikleri incelenmiştir. Çok boyutlu veri setlerinin boyutluluk analizini sağlayan ve bu veri setleriyle çeşitli test uygulamaları yapma olanağı tanıyan programların hangi tür verilerle (ikili, çok kategorili), modellerle, kestirim teknikleri ve algoritmalarıyla ne tür analizler yapabildikleri incelenmiştir. Bunun yanı sıra programların çıktılarında yer alan indekslerin hangileri oldukları ve bu indekslerin yorumlanmasına ilişkin ölçütler açıklanmıştır. Son olarak programların çeşitli uygulamalar açısından avantaj ve dezavantajlarını inceleyen, programları çeşitli özellikleri açısından karşılaştıran simülasyon çalışmaları verilmiştir.

Araştırmada sırasıyla Mplus, NOHARM, TESTFACT, IRTPRO, flexMIRT, BMIRT, DIMPACK (DETECT, DIMTEST, CCPROX/HCA), Velicer'in MAP Testi, Paralel Analiz olmak üzere boyutluluk analizinde yaygın kullanılan programlar ve yaklaşımlar incelenmiştir. Ayrıca programlara ilişkin özelliklerin karşılaştırıldığı bir tablo Ek 1'de verilmiştir.

PROGRAMLAR

Mplus 7.31:

Mplus programı Muthen ve Muthen (1998-2012) tarafından geliştirilmiştir. Programın 1998-2014 yılları arasında farklı sürümleri geliştirilmiş olup, 7.31 sürümü 2014'de geliştirilmiş ve 7.4 sürümünün 2015 yılının sonlarında kullanıcılara sunulması hedeflenmektedir. Programın kullanımına yönelik olarak ayrıntılı yazılmış bir pdf dosyası bulunmaktadır. Mplus ücretli olup, hem nasıl edinilebileceğine ilişkin ayrıntılar hem de kullanımına ilişkin pdf dosyası geliştiriciler tarafından hazırlanan bir web sitesinde bulunmaktadır.

(https://www.statmodel.com/ugexcerpts.shtml).

Mplus, araştırmacılara çok sayıda model seçeneği, kestirim algoritması, grafik seçenekleri ve verileri analiz etme olanağı sağlayan bir istatistiksel modelleme programıdır. Program tek ve çok düzeyli verilerle, kayıp verilerle, sürekli, sıralı, sınıflama ve oran ölçeğiyle elde edilmiş veri setleriyle ve tüm bu veri türlerinin kombinasyonlarıyla analiz yapma olanağı vermektedir. Ayrıca Mplus programı ile veriyi Monte Carlo simülasyonları kullanarak üretmek de mümkündür (Muthen ve Muthen, 1998-2012). Mplus programı sürekli örtük değişkenler kullanarak regresyon analizi, path analizi, açımlayıcı ve doğrulayıcı faktör analizi, madde tepki kuramı modellemeleri, yapısal eşitlik modellemeleri, örtük büyüme modellemeleri, kesikli zaman serisi ve sürekli zaman serisi analizleri yapmaya olanak vermektedir. Aynı zamanda kategorik örtük değişkenler kullanarak regresyon analizi, path analizi, örtük sınıf analizi, doğrulayıcı örtük sınıf analizi, loglinear modellemesi, çoklu grup modelleme, CACE (Complier Average Casual Effect) modelleme, kesikli ve sürekli zaman serileri analizleri yapmaya da olanak vermektedir. Dolayısıyla hem sürekli hem de kategorik örtük değişkenleri birlikte kullanarak rastgele etkilerle örtük sınıf analizleri, faktör karışımı modellemesi, yapısal eşitlik karışım modellemesi, büyüme karışım modellemesi, kesikli ve sürekli zaman serisi karışım modellemesi yapılabilmektedir (Muthen ve Muthen 1998-2012).

Mplus programı hem Bayesyen hem de olabilirlik (likelihood) yaklaşımlarını kullanarak kestirimde bulunmaktadır. Bayesyen yaklaşımda MCMC (Monte Carlo Markov Chain) algoritmasını, olabilirlik yaklaşımda ise maksimum olabilirlik ve ağırlıklandırılmış en küçük kareler algoritmalarından seçileni kullanmaktadır. Ayrıca Quasi-Newton, Fisher scoring, Newton-Raphson, Expectation Maximization (EM) gibi optimizasyon algoritmaları da seçilebilmektedir (Gelman, Meng ve Stern, 1996; Svetina ve Levy, 2012)

Mplus programı açımlayıcı ve doğrulayıcı yaklaşımlarla boyutluluk analizi yapabilmektedir. Açımlayıcı faktör analiziyle, dik ve eğik döndürme teknikleri uygulanabilmekte, program çıktılarında veri yapısına uygun (polikorik, tetrakorik vb.) korelasyon matrisleri, artık korelasyon matrisleri için özdeğerler verilmektedir. Ayrıca boyutluluğun değerlendirilmesi için RMSR (root mean squared residual), χ^2 istatistiği ve RMSEA (root mean square error of approximation) değerleri programın çıktılarında verilmektedir (Muthen ve Muthen, 1998-2012).

Boyutluluk analizlerinde kullanılan uyum iyiliği ve hata (uyumsuzluk) indekslerinin yorumlanmasına ilişkin farklı görüşler bulunmakla birlikte, çoğunlukla iyiliği indekslerinin değeri 1'e yaklaştıkça uyumun arttığı, diğer bir deyişle modelin veriye uyumunun arttığı, hata indekslerinin ise sıfıra yaklaştıkça model uyumunun arttığı ifade edilmektedir. χ^2 (Ki-kare) uyum istatistiği örneklem büyüklüğüne duyarlılığı nedeniyle (Bentler ve Bonnet, 1980; Jöreskog ve Sörbom, 1993; Kenny ve McCoach, 2003) tartışılsa da yaygın kullanılan uyum istatistiklerinden biridir. Örneklem büyüklüğüne duyarlılığı azaltmak için Wheaton Muthen, Alwin ve Summers (1977) tarafından χ^2 / df oranının kullanılması gerektiği önerilmiştir ve bu oran 2' den küçük değerlere sahipse modelin iyi uyum gösterdiği, 2 ile 5 arasında ise modelin kabul edilebilir olduğu ifade edilmektedir (Tabachnick ve Fidell, 2007).

RMSEA uyum istatistiği Steiger (1990) tarafından geliştirilmiş ve bu indeks de kestirilen parametre sayısına duyarlı olmasına rağmen son yıllarda model uyumunu değerlendirme de yaygın kullanılan indekslerden biridir (Diamantopoulos and Siguaw, 2000). RMSEA değerinin 0.10'un üstünde olması kötü uyumu, 0.08 ile 0.10 arasında olması kabul edilebilir uyumu (MacCallum, Browne ve Sugawara, 1996) göstermektedir. Ancak son yıllarda RMSEA değerinin 0.00 ile 0.06 veya 0.07 arasında olması gerektiği ifade edilmektedir (Hu ve Bentler, 1999; Steiger, 2007). RMR ve SRMR değerlerinin 0 olması mükemmel uyumu gösterirken 0.05- 0.08 aralığında ki değerler kabul edilebilirdir (Hu ve Bentler, 1999 Byrne, 1998; Diamantopoulos ve Siguaw, 2000). GFI, AGFI, CFI gibi indekslerin ise 0.90 üstü olması kabul edilebilir bir uyum iyiliği değerini, 0.95 üstü olması ise iyi bir uyum iyiliğinin göstergesi olarak kabul edilir (Tabachnick ve Fidell, 2007). AIC (Akaike Bilgi Kriteri) ve BIC (Bayesyen Bilgi Kriteri) indekslerinin ise daha küçük değerler alması iyi uyumu göstermektedir.

NOHARM(Normal Ogive Harmonic Analysis Robust Method)

NOHARM (Normal Ogive Harmonic Analysis Robust Method) bir model olarak McDonald (1967) tarafından geliştirilmiş olup, Fraser (1998) tarafından bilgisayar programı (http://noharm.software.informer.com/download/) olarak geliştirilmiştir. Modele yönelik ayrıntılı bilgilere McDonald (1962; 1967; 1981; 1997; 2000) tarafından yapılan çalışmalarda erişebilmek mümkündür. Program iki kategorili verilerle tek ve çok boyutlu normal ogive modeller için analiz yapabilmektedir. Ancak çok kategorili ve kayıp verilerle analiz yapamamaktadır. Program en küçük kareler kestirimi algoritasını kullanmaktadır Program ücretsizdir. Programın temel dezavantajlarından biri şans parametresi kestirimi yapamamasıdır ancak kullanıcıların şans parametresi değerlerini programa manuel olarak girmesine imkan tanımaktadır. (McDonald ve Fraser, 1988; Reckase, 2009; Svetina ve Levy, 2012).

Program faktör çözümleri için dik ve eğik döndürme olanağı tanıyarak açımlayıcı ve doğrulayıcı faktör analizi yapma olanağı tanımaktadır. NOHARM tetrakorik korelasyonları kullanmamakta, normal ogive modellere polinomiyal bir yaklaşım geliştirmektedir ve ağırlıklandırılmış en küçük kareler yaklaşımıyla kestirim yapmaktadır (Çok boyutlu modellere polinomiyal yaklaşım için Reckase (2009) incelenebilir). Program çıktılarında beklenen ve gözlenen oranlar arasındaki farka ilişkin artık matrislerini, RMSR (root mean square residual), Tanaka'nın uyum iyiliği indisi (GFI), döndürülmüş faktör yükleri verilmektedir. Ayrıca NOHARM programının çıktılarındaki değerler kullanarak $X^2_{G/D}$ istatistiği ile ALR (approximate likelihood ratio) istatistiği de hesaplanmaktadır. NOHARM çıktılarında yedi bölüm bulunmaktadır. İlk bölümde madde sayısı, analiz başlığı korelasyon matrisleri, kullanıldıysa sabitlenmiş şans parametreleri, bulunmaktadır. İkinci bölümde örtük yetenek ve genel faktör parametreleri bulunmaktadır. Bu parametreler; güçlük ve ayırt edicilik parametreleri ile RMSR ve Tanaka'nın ağırlıklandırılmamış en küçük kareler uyum iyiliği indisidir.

NOHARM programı k boyutlu normal ogive modeli aşağıda verilen eşitlikteki gibi kullanır.

$$P\{U_i = 1 | \Theta = (\theta_1, ...\theta_k)\} = N\{\beta_{i0} + \beta_{i1}\theta_1 + ... + \beta_{ik}\theta_k\}$$

Verilen eşitlikte N normal ogive fonksiyonunu, θ örtük yetenek vektörünü, β_{i0} madde güçlük parametresini, β_{ik} madde ayırt edicilik parametresini göstermektedir.

NOHARM programı ile doğrulayıcı yaklaşımla boyutluluk analizi yapılmışsa model uyumunu değerlendirmek için RMSR (root mean square residual) ve Tanaka' nın uyum iyiliği incelenebilir. RMSR değerinin sıfır olması mükemmel uyumu gösterirken, değer büyüdükçe model uyumu kötü hale gelmektedir (Kline, 2005). Tanaka' nın indeksi 0 ile 1 arasında değerler almaktadır. Bu değer 1'e yaklaştıkça modelin uyumu artmaktadır. Boyut sayısına açımlayıcı yaklaşımla karar verme sürecinde uyum indeksinde hızlı artış gözlenirse ve RMSR değerindeki düşüşler %10 veya daha fazla ise en uygun modelin belirlendiği görüşü bulunmaktadır (Tate, 2003). Finch ve Habing (2005), 1 ve M model arasında bir seri modelin uyumunu belirleyerek, farklı sayıda boyut (örtük değişken) öneren modeller için $X^2_{G/D}$ farkının belirlenmesini ve yorumlanmasını önermişlerdir.

TESTFACT 4.0

TESTFACT modeli ilk olarak "full information item factor analysis" çalışmalarıyla Bock, Gibbons ve Muraki (1998) tarafından oluşturulmuştur. TESTFACT 4.0 sürümü ise Bock, Gibons, Schiling, Muraki, Wilson ve Wood tarafından (2003) geliştirilmiştir. Program ücretlidir ve araştırmacılar tarafından edinilebilmesi için TESTFACT programının yanı sıra çok sayıda farklı istatistik programı da geliştiren Uluslararası Bilimsel Paket programı (Scientific Software International, http://www.ssicentral.com/irt/index.html) grubunun internet sitesinin incelenmesi gerekmektedir. Programın kullanımına yönelik ayrıntılara Mathide Du Toit (2003)'in editörlüğünü yaptığı "IRT

from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT" isimli kitapta Bock ve Schilling (2003) tarafından yazılan "IRT based item factor analysis" isimli bölümde yer verilmiştir.

TESTFACT programı iki ve çok kategorili veriler, kayıp veriye sahip veri setleri tetrakorik korelasyon matrisleri için analiz yapabilmektedir. Program kestirim algoritması olarak Bock ve Aitkin (1981) tarafından geliştirilen marjinal maksimum olabilirlik kestirimini ve ayrıca yetenek parametreleri için Bayesyen algoritması kullanmaktadır. TESTFACT programının sınırlılıklarından biri doğrulayıcı yaklaşımla boyutluluk analizinde sadece iki faktörlü model için çözüm üretmesi, diğeri ise NOHARM programı gibi şans parametresi kestirimi yapamamasıdır. Ancak TESTFACT programına da şans parametresi için sabit değerlerin kullanıcı tarafından girilebilmesi mümkündür. Program klasik madde analizi, test puanı hesaplama ve maddeler arası tetrakorik korelasyonu kullanarak faktör analizi yapabilmektedir Ayrıca program ile madde tepki kuramına dayanan modern faktör analizi yapmak mümkündür (Reckase, 2009; du Toit, 2003; Svetina ve Levy, 2012).

TESTFACT programı açımlayıcı faktör analiziyle dik ve eğik döndürme yapmaya olanak tanımaktadır. Boyutluluğun analizinde modele bir faktör daha eklendiğinde Ki-kare (χ^2) uyum istatistiğinde meydana gelen değişiklik temel alınmaktadır. Analizlere tek faktörlü modelle başlanmakta ve modele bir faktör daha eklendiğinde model uyumundaki düzelmeler incelenmektedir. Modele bir faktör daha eklendiğinde model uyumunda önemli bir değişiklik olmaz ise yapının tek boyutlu olduğu ifade edilmektedir. Modele bir faktör daha eklendiğinde model uyumunda anlamlı değişiklik olursa testin çok boyutlu olduğuna karar verilmekte ve faktör eklenmeye devam edilmektedir. Testin boyut sayısı eklenen faktörün anlamlı düzeyde değişiklik meydana getirmediği durum olarak kabul edilmektedir.

IRTPRO 2.1

IRTPRO programı Cai, du Toit ve Thissen (2011) tarafından geliştirilmiştir. Programın kullanımına ilişkin ayrıntılı bir pdf dosyası yazarlar tarafından hazırlanmıştır. Program ücretlidir ve araştırmacılar tarafından programın edinilebilmesi ve pdf dosyasının indirebilmesi için Uluslararası Bilimsel Paket programı (Scientific Software International, http://www.ssicentral.com/irt/index.html) grubunun internet sitesinin incelenmesi gerekmektedir. Programın kullanımına yönelik ayrıntılı bir pdf dosyası ise internet sitesindeki (http://coeweb.gsu.edu/coshima/EPRS8410/IRTPRO%20Guide.pdf) linkinde bulunmaktadır. Program iki ve çok kategorili verilerle, erişilmemiş veya kayıp veriye sahip veri setleriyle, tek ve çok boyutlu madde tepki kuramı modellerinin birçoğuyla analizler yapmaya olanak tanımaktadır.

IRTPRO programına çeşitli istatistik paket programlarından veri aktarılabildiği gibi fixed, csv, txt ve xls uzantılı formattaki verilerde aktarılabilmekte ve IRTPRO çıktısı dosyalar .ssig* uzantılı olarak kaydedilmektedir. IRTPRO açımlayıcı ve doğrulayıcı analizleri tek ve çok boyutlu modellerin birçoğuyla (bir, iki ve üç parametreli lojistik modeller, aşamalı tepki modeli, genelleştirilmiş kısmi puan modeli, sınıflamalı model) yapabilmektedir. IRTPRO programı parametre kestiriminde genellikle maksimum olabilirlik tekniğini kullanmaktadır ancak madde parametreleri için önsel dağılımların önceden belirlenmesi durumunda maksimum sonsal (Maximum a posteriori) kestirim tekniği kullanılabilmektedir. Ayrıca her biri farklı test yapılarıyla veya modellerle daha iyi sonuçlar veren çeşitli kestirim teknikleri (Bock Aitkin EM (BA-EM), Bifactor EM, Generalized Dimension Reduction EM, Metropolis Hasting Robbins-Monro (MH-RM)) ile IRTPRO programında madde parametresi kestirimi olanaklıdır.

Madde sayısı, cevap kategorileri ve cevaplayıcılara göre değişmek suretiyle çeşitli uyum iyiliği istatistikleri madde kalibrasyonundan sonra hesaplanmaktadır. 2loglikelihood, Akaike Bilgi Kriteri (Akaike Information Criterion (AIC)) ve Bayesyen bilgi kriteri (BIC) rapor edilen başlıca istatistiklerdir. Bazı modeller için ayrıca M_2 istatistiği hesaplanmaktadır. Bilişsel istatistikler ise LD (yerel bağımsızlık istatistiği) ve SS-X^2 madde uyum istatistikleridir.

flexMIRT 2.0

flexMIRT 2.0 programı Cai (2013) tarafından geliştirilmiştir. Program çok zengin psikometrik ve istatistik özelliklere sahiptir ve ücretlidir. Programın özellikleri ve kullanımına yönelik ayrıntılı bir pdf dosyası Houts ve Cai (2013) tarafından yazılmıştır. Programın tüm özelliklerine yönelik ayrıntılı bilginin olduğu ve programın satın alınabileceği bir web sitesi bulunmaktadır (https://www.vpgcentral.com/irt-software/). Programın kullanımına hızla teknik destek sunan bir internet sitesi de bulunmaktadır (http://www.vpgcentral.com/irt-software/support/).

Program Windows temellidir, grafik ara yüzü sayesinde kullanıcı dostudur. Oldukça hızlıdır ve boşluk, virgül ve sekmeyle ayrılmış veriyi yüklemeye olanak sağlamakta, iki ve çok kategorili verilerle analiz yapabilmektedir. Monte Carlo simülasyonuyla çok sayıda modele ilişkin veri üretmek de mümkündür. Ayrıca program kayıp ve erişilmemiş verisi olan veri setleriyle de analiz yapabilmektedir. flexMIRT programı doğrulayıcı yaklaşımla çok sayıda tek ve çok boyutlu MTK modeli için madde parametresi kestirimi yapabilmektedir (Örneğin, tek boyutlu iki kategorili modellerden Rasch, bir, iki ve üç parametreli lojistik model, Samejima' nın aşamalı tepki modeli, sınıflama modeli, kısmi ve genelleştirilmiş kısmi puan modeli ile dereceli ölçek modeli, çok boyutlu bir, iki, üç parametreli lojistik model, çok boyutlu aşamalı tepki modeli gibi). Ayrıca tek ve çok gruplu, tek ve çok düzeyli veri setlerine ilişkin parametre kestirimi de yapılabilmektedir. Madde parametresi kestirimi için BA-EM ve MHRM tekniklerini kullanılarak flexMIRT programı ile parametre kestirimi yapılabilmektedir. Yetenek parametresi kestirimi için maksimum olabilirlik (ML), maksimum sonsal (MAP (maximum a posteriori)) ve beklenilen sonsal (EAP (expected a posteriori)) olmak üzere üç kestirim tekniği kullanılabilir. Ayrıca flexMIRT toplam puanları MTK ölçek puanlarına dönüştüren tablolar oluşturabilmekte ve parametre kestirimine ilişkin çeşitli standart hata kestirim tekniklerinin kullanılmasını sağlamaktadır. Bu teknikler; tamamlayıcı EM, teorik bilgi fonksiyonu, Fisher bilgi fonksiyonu ve sandwich kovaryans matrisidir. Ayrıca çıktılarında çok sayıda uyum indeksi rapor edilmektedir. Bunlardan başlıcaları, Chen ve Thissen's χ^2 (Chen ve Thissen, 1997) Akaike Bilgi Kriteri (AIC), Bayesyen bilgi kriteridir (BIC) (Houts ve Cai, 2013).

BMIRT(Bayesian Multivariate Item Response Theory)

BMIRT programı Yao (2003) tarafından geliştirilmiştir. Programın en önemli avantajı ücretsiz olmasıdır. Programın araştırmacılar tarafından akademik çalışmalarında kullanılacağına ilişkin bir belge imzalamaları ve Java programını yüklemeleri yeterlidir. Programın kullanımına yönelik ayrıntılı bir pdf dosyası (Yao, 2013) ve program araştırmacılar tarafından Yao (2003) tarafından hazırlanmış bir web sitesinden indirilebilmektedir (http://www.bmirt.com/6271.html). Program kullanılmadan önce Java'nın bulunduğu yerin uzantısının bilgisayarın gelişmiş sistem ayarlarında bulunan ortam değişkenlerine yapıştırılması gerekmektedir. Program, .bat* uzantılı dosyalarla DOS komutuyla çalışmaktadır. Programı çalıştırmak için kontrol, ham veri ve uygulama olmak üzere üç farklı dosyanın hazırlanması gerekmektedir.

Bu program iki veya çok kategorili tek ve çok boyutlu modellere ilişkin parametreleri Metropolis-Hasting algoritmasını kullanan Markov zinciri Monte Carlo (Markov chain Monre Carlo (MCMC)) tekniği ile kestirmektedir. Monte Carlo yöntemleri veri simülasyonu için oldukça yaygın yöntemlerdir. MCMC kestirimleri için anahtar durum ise bilinen, uygun dağılımlardan üretilen örneklemler kullanılarak daha karmaşık dağılımlara ilişkin örneklem üretilmesidir. Bu durum kabul-red örneklemi (Chib ve Greenberg,1995) veya basitçe red örnekleme (Gamerman ve Lopes, 2006) olarak bilinmektedir.

BMIRT programı madde ve yetenek kestiriminin çok boyutlu ve çoklu grup veri setleri ile yapılmasına olanak tanımaktadır. Program hem açımlayıcı hem de doğrulayıcı faktör analizi, bir, iki ve üç parametreli lojistik modeller, Rasch modeli, genelleştirilmiş iki parametreli kısmi puan testlet modeli, aşamalı tepki modeli için parametre kestirimi yapabilmektedir (Yao, 2013).

DIMPACK V.2.0 (Parametrik olmayan Boyutluluk Değerlendirme Paketi, Nonparametric Dimensionality Assesment Package)

Program beş ayrı boyutluluk analizi yapan parametrik olmayan programı bir araya getirmiştir ve ücretsizdir. DIMPACK programı DIMTEST V.2.1, ATFIND V.1.3, DETECT V.2.1, CCPROX ve HAC programlarını bir araya getiren William Stout Enstitüsü (Nonparametric Dimensionality Assessment Package, 2006) tarafından geliştirilmiş bir paket programıdır. Programın içinde bulunduğu zip dosyasını indirmek ve programla ilgili yapılmış makalelere ulaşmak için bir web sitesi bulunmaktadır (http://psychometrictools.measuredprogress.org/dif2). Ayrıca programın kullanımına yönelik de teknik destek verilmektedir (stoutdist@stat.uiuc.edu).

Programın kurumundan önce araştırmacıların bilgisayarlarında "NET framework" programının kurulu olduğundan emin olmaları gerekmektedir. Programla analiz yapılırken kullanılabilecek maksimum madde sayısı 150 ve maksimum örneklem büyüklüğü 7000'dir. Programın bir ara yüzü olduğundan kullanıcı dostudur (Deng, Han ve Hambleton, 2013).

Program boyutluluk analizinde şartlı kovaryans matrisini kullanan parametrik olmayan boyutluluk değerlendirme paketidir. DIMTEST programı verinin tek boyutlu olmadığı varsayımına dayalıdır ve maddeleri alt testlere kümelemektedir, DETECT basit yapılı veri setleri için boyutluluk analizi yapmakta, CCPROX/HAC ise yakınlık matrisleri hesaplamakta ve hiyerarşik kümeleme analiziyle değerlendirme yapmaktadır. Sırasıyla bu programlar ayrı ayrı incelenmiştir.

DIMTEST

Program Stout (1987) tarafından oluşturulmuş ve daha sonra Stout, Froelich ve Gao (2001) tarafından geliştirilmiştir. DIMTEST tek boyutluluk ve yerel bağımsızlık varsayımını, parametrik olmayan hipotez ile test etmektedir (Nandakumar ve Stout, 1993; Stout, 1987; Stout ve ark., 2001). DIMTEST programının DIMPACK paketinde bulunan sürümünün yanı sıra DIMTEST_DOS şeklinde DOS sürümü de vardır. DIMTEST programında öncelikle aynı örneklemden alınan madde havuzu AT (assessment subtest, değerlendirme alt testi) ve PT (partitioning, bölümlenme alt testi) olarak adlandırılan iki ayrı alt teste ayrılmaktadır. DIMPACK paketinde bulunan DIMTEST madde setlerini ikiye ayırırken DOS komutlu sürümü AT1, AT2 ve PT olmak üzere üçe ayırmaktadır. Eğer araştırmacılar DIMPACK paketi yerine DOS komutlu sürümü kullandıysa, AT1'de oluşturulan maddelerin benzer yeteneğe duyarlı oldukları varsayılmaktadır. AT2 maddelerinin ise başka bir yeteneğe, AT1 tarafından ölçülen yeteneğe göre daha duyarlı oldukları varsayımı yapılmaktadır. Ancak hem AT1 hem de AT2'deki maddelere ilişkin doğru cevap oranları dağılımlarının gözlenen frekansları açısından benzer olduğu ifade edilmektedir. Ayrıca DIMPACK paketinde bulunan DIMTEST'in avantajlarından biri de AT maddelerinin otomatik olarak seçilmesinin mümkün olmasıdır.

DIMTEST programının performansında en belirleyici olan durumlardan biri AT maddelerinin seçimidir. DIMPACK paketinde bulunan DIMTEST progamında maddelerin AT ve PT şeklinde iki ayrı alt teste bölünmesi açımlayıcı veya doğrulayıcı yaklaşımla yapılabilmektedir. Doğrulayıcı yaklaşımda AT maddeleri test yapısına dayalı olan uzman görüşlerine göre oluşturulmaktadır. Açımlayıcı yaklaşımda AT maddeleri istatistiksel bir prosedürle (örneğin; madde tetrakorik korelasyonlarına ilişkin döndürülmemiş temel faktör analizi sonuçlarıyla) belirlenmektedir. DIMTEST, eğer bir test tek boyutlu ise, PT madelerinden ayrıldıktan sonra AT alt testinde bulunan iki madde seti arasındaki şartlı kovaryansın sıfıra eşit olması düşüncesine dayalıdır (Nandakumar ve Stout, 1993; Deng ve ark., 2013; Özkan ve Güvendir, 2014). AT maddelerinin belirlenmesi DETECT ve HAC/CCPROX programlarının birleşimlerini kullanan ATFIND isimli programla da mümkündür. Ayrıca alternatif olarak, araştırmacılar AT maddelerini kendileri DETECT veya HAC/CCPROX analizleri veya doğrusal olmayan faktör analizi kulanan NOHARM, TESTFACT ile veya tetrakorik korelasyon matrislerine dayalı doğrusal faktör analizi kullanarak belirleyebilirler (Nonparametric Dimensionality Assessment Package, 2006).

DIMTEST programının çıktılarında .dim* uzantılı bir dosya elde edilir. Bu dosyanın içinde AT ve PT maddeleri ile T istatistiği ve p değerleri verilmektedir. DIMTEST programıyla boyutluluk analizinde daha öncede ifade edildiği gibi bir nonparametrik hipotez vardır. Bu hipotez, "H_o: Test tek boyutludur, H_1: Test çok boyutludur" şeklinde yapılmaktadır. Bu hiptez DIMTEST tarafından hesaplanan Stout'un T istatistiği ile değerlendirilir. Eğer DIMTEST T istatistiği sıfıra yakınsa Ho hipotezi doğrulanmış ve elde edilen p değeri manidar değilse veri kümesinin tek boyutlu olduğu sonucuna varılmaktadır. Elde edilen p değerinin manidar çıkması ise verinin çok boyutlu olduğuna işaret etmektedir (Nonparametric Dimensionality Assessment Package, 2006; Deng ve ark., 2013).

DETECT(Dimensionality Evaluation to Enumarate Contributing Traits)

DETECT uygulaması Kim (1994) ile Zhang ve Stout (1999) tarafından geliştirilmiştir. Parametrik olmayan boyutluluk analizinde şüphesiz en önemli yaklaşımlardan biridir. DETECT tekniği diğer parametrik olmayan tekniklere göre avantajlara sahiptir. Bu avantajlar sırasıyla; veri setine ait boyut sayısını kestirebilmesi, çok boyutluluk için etki büyüklüğünü hesaplayabilmesi, her bir madde tarafından baskın şekilde ölçülen boyutu belirleyebilmesidir. DETECT tekniği özellikle çok boyutlu basit test yapısını belirlemek için geliştirilmiştir (Nonparametric Dimensionality Assessment Package, 2006).

Program iki ve çok kategorili verilerle uygulanabilmekte, kayıp veriyle analiz yapamamaktadır. Eğer çok boyutlu basit yapı yaklaşık olarak ortaya çıktıysa, DETECT tekniği öncelikle tüm olası madde çiftleri arasındaki şartlı kovaryansı hesaplayarak boyutluluk analizine başlamaktadır. DETECT daha sonra testi basit yapılı boyutlara ayırmayı sağlayan ortalama şartlı kovaryans matrislerini kullanarak kümeleme analizi yapmaktadır. DETECT tekniği ile boyutluluğun değerlendirilmesi için çeşitli istatistikler ve indeksler hesaplanmaktadır.

Boyutluluğun analizinde kullanılan indekslerden biri DETECT indeksidir. Bu indeksin hesaplanması için kullanılan formül aşağıda verilmiştir;

$$D(P) = \frac{2}{n(n-1)} \sum_{1 \le i \le j \le n} \delta_{ij} E[Cov(X_i, X_j) | \Theta_{TT}]$$

Formül incelendiğinde; n madde sayısını, P, k kümeye ayrılan n madde sayısını, Θ_{TT} test birleşimini, X_i ve X_j; i ve j maddelerine ait gözlenen puanları göstermektedir. δ_{ij} değeri eğer i ve j maddesi aynı küme içerisinde ise 1'e eşit çıkmakta değilse -1'e eşit çıkmaktadır. Teorik olarak eğer veri seti tek boyutlu ise $D(P)$ indeksinin sıfırın altında çıkması beklenmektedir. DETECT indeks değeri 1'den büyükse çok boyutluluk için güçlü bir kanıtın olduğu ifade edilmektedir. Eğer D(P) 0.40 ile 1.00 arasında ise çok boyutluluk için orta düzeyde bir kanıt söz konusudur. D(P), 0.20 ile 0.40 arasında ise çok boyutluluk olasılığı zayıftır ve ve 0.20'den küçükse verinin tek boyutlu olduğu yorumu yapılmaktadır (Zhang ve Stout, 1999; Roussos ve Ozbek, 2006; Svetina ve Levy, 2012).

DETECT tarafından hesaplanan indekslerden biri de r indeksidir. Bu indeks yaklaşık basit test yapısının yorumlanması için kullanılmaktadır. Bir çok istatistik programında testlerin basit yapıda olduğu varsayımı bulunmaktadır. Basit yapı her bir maddenin sadece tek bir boyutu ölçtüğü diğer boyutlarda hiç bir faktör yüküne sahip olmadığı çok boyutlu test yapısıdır. Ancak basit yapıya gerçek veri setlerinde çok ender rastlanmaktadır. Bu nedenle araştırmacılar tarafından yaklaşık test yapısı önerilmektedir. Yaklaşık test yapısı ise her maddenin bir boyutta baskın faktör yüküne sahip olduğu ancak diğer boyutlarda da faktör yüküne sahip olduğu çok boyutlu test yapısıdır. DETECT program ile yaklaşık test yapısı boyutluluğu incelenirken aşağıda verilen formül kullanılmaktadır:

$$r_{\max} = \frac{D(P^*)}{\hat{D}(P^*)}$$

Formül incelendiğinde, $\hat{D}(P^*)$ değeri tüm madde çiftleri boyunca şartlı kovaryans değerinden elde edilen toplam mutlak değerlerin maksimum olasılıklı değerini göstermektedir. Bu formülden elde edilen r_{max} değeri şöyle yorumlanmaktadır; eğer *r* indeksi 0.80 ve üstünde değerler aldıysa verinin yaklaşık basit çok boyutlu test yapısına sahip olduğu ifade edilmektedir (Kim, 1994; Roussos ve Ozbek, 2006; Zhang ve Stout,1999; Svetinave Levy, 2012; Jang ve Roussos, 2007).

CCPROX ve HAC

CCPROX ve HAC yaklaşımlarının ikisi birlikte açımlayıcı parametrik olmayan boyutluluk analizinde kullanılabilmektedir. CCPROX/HAC testin boyutluluk yapısına ilişkin istatistiksel bilgi sağlamaktadır. CCPROX/HAC yaklaşımından testin boyutluluk yapısını belirleme sürecinde yararlanıldığı gibi DIMTEST programında kullanılan AT madde setlerinin seçimi içinde yararlanılmaktadır.

CCPROX/HAC analizi iki bölümden oluşmaktadır. HAC analizinden önce CCPROX analizinin yapılması gerekmektedir, çünkü CCPROX analiziyle oluşturulan yakınlık matrisi HAC analizinde girdi dosyası (*.prx uzantılı kaydedilmekte) olarak kullanılmaktadır. Yakınlık matrisi dosyası otomatik olarak kaydedilmekte ve eğer CCPROX analizinden hemen sonra HAC analizi yapılacaksa, dosya otomatik olarak yüklenmektedir.

HAC

HAC hiyeraşik kümeleme analizi yapmaktadır. Hiyerarşik kümeleme analizi, n sayıda her bir madde bir kümede olacak şekilde başlamakta ve daha sonra minimum yakınlığa sahip maddelerin bir araya gelmesiyle ikinci aşamaya geçilmektedir. Bu şekilde maddelerin tümünün daha büyük bir kümede yer almasıyla analiz sonuçlanmaktadır. Yeni kümeler için yakınlıklar, diğer kümelere ve daha sonra hesaplanan küme analizi türlerine (program çok sayıda alternatif sunmaktadır; karmaşık bağlantı, tek bağlantı gibi) bağlı olarak değişmektedir. Analiz hangi hiyerarşinin ya da düzeyin doğru boyut sayısı olduğu ile ilgili bilgi verememekte ancak ek tanımlayıcı istatistiksel bilgi sunmaktadır.

CCPROX

HAC analizinin amacı, test maddelerinin boyutluluk yapısı ile ilgili bilgi sunmaktır. HAC analizinin testin boyutluluğu ile ilgili kullanışlı bilgiler sunabilmesi için, madde çiftleri için hesaplanan yakınlık değerlerinin maddeler arasındaki boyutluluk farklılıklarına duyarlı olması gerekmektedir. Bu nedenle de CCPROX analizinin amacı bir yakınlık matrisi oluşturmaktır. CCPROX analizinde kullanılan yakınlık kestirimleri her bir madde çifti arasındaki şartlı kovaryans kestirimlerine dayalıdır. Eğer test çok boyutlu ise ve iki madde aynı boyutu ölçüyorsa şartlı kovaryans pozitif olmakta ve eğer maddeler farklı boyutu ölçüyorlarsa bu iki maddeye ilişkin şartlı kovaryans negatif olmaktadır. Bu şartlı kovaryanslardan yararlanılarak birbirine benzer boyutu ölçen maddelerin aynı kümede toplanması amaçlanmaktadır (Roussos, 1992; Roussos, Stout ve Marden, 1998).

Velicer'in Map Testi Ve Horn'un Paralel Analizi:

Paralel analiz ve Velicer'in MAP testi, özellikle SPSS ve SAS gibi yaygın kullanılan mevcut paket programlarıyla uygulanan yaklaşımlar olmaması nedeniyle araştırmacılar tarafından boyutluluk analizinde yaygın kullanılamamaktadır. Ancak özellikle paralel analiz olmak üzere bu iki tekniğin boyut sayısı belirlemede oldukça etkili yaklaşımlar oldukları yönünde literatürde görüş birliği vardır (Hayton, Allen ve Scarpello, 2004; Watkins 2006; Storch, Murphy, Bagner, Johns, Baumeister ve Goodman, 2006; Munroe ve Pearson 2006; Nelson, Canivez, Lindstrom ve Hatt, 2007; Justicia, Pichardo, Cano ve Berben, 2008; Crawford, Green, Levy, Scott, Svetina ve Thompson, 2010; Garrido, Abad ve Ponsoda, 2011).

Bu iki tekniğin uygulanabilmesi için O'Connor (2000) tarafından SPSS, SAS, MATLAB, ve R programlarıyla uygulanabilir syntax (betik) dosyaları hazırlanmıştır. Bu dosyalar ücretsiz olup, dosyalara ve kullanımlarına yönelik ayrıntılı bilgilerin sunulduğu bir internet sitesi mevcuttur (https://people.ok.ubc.ca/brioconn/nfactors/nfactors.html). Ayrıca bu betik dosyaların SPSS programında kullanımına yönelik hazırlanmış ayrıntılı videolarda bulunmaktadır (https://www.youtube.com/watch?v=T908yGVgjPk).

Velicer'in Map Testi

Velicer'in MAP testi, Velicer (Minimum average partial, Velicer, 1976) tarafından geliştirilmiştir. Bu teknikle boyut sayısı belirlenirken kısmi korelasyon matrislerinden ve temel bileşen analizinden yararlanılmaktadır. Boyut sayısının belirlenmek istediği maddelerden bir korelasyon matrisi oluşturulmakta ve ilk basamakta bu korelasyon matrisinden ilk boyut ayrılmaktadır. Bu işlem sırasında orijinal matristen korelasyon katsayılarının karelerinin ortalaması hesaplanmaktadır. İlk boyut ayrıldıktan sonra kalan kısmi korelasyon matrisinin köşegen elemanlarından kısmi korelasyonların karelerinin ortalamaları hesaplanmaktadır. Bu işlem değişken sayısının bir eksik basamağına kadar sürdürülebilmekte ve her basamakta yeni bir kısmi korelasyon ortalaması elde edilmektedir. Daha sonra tüm basamaklardan elde edilen kısmi korelasyonların ortalama kareleri sıralanmaktadır. Orijinal matristen elde edilen korelasyon karelerinin ortalaması; en düşük kısmi korelasyon kareler ortalamasından daha küçük olduğu anda, artık matristen herhangi bir bileşenin/faktörün ayrılamayacağı ve boyut sayısının o değere kadar olan basamak sayısı kadar olması gerektiği önerilmektedir. Diğer bir deyişle analizde en düşük kısmi korelasyon ortalama karesinin basamak numarası bileşen sayısı olarak ifade edilmektedir (Velicer, 1976; Velicer, Eaton ve Fava, 2000; O'Connor, 2000; Watkins, 2006; Garrido ve ark., 2011).

Paralel Analiz

Paralel analiz, Horn (1965) tarafından boyut sayısını belirlemeye yönelik geliştirilen bir tekniktir. Paralel analizde gerçek veriyle aynı özelliklere sahip tesadüfi bir seri korelasyon matrisi oluşturulmakta ve hem bu rastgele oluşturulmuş korelasyon matrislerinden hem de gerçek veriden temel bileşen analiziyle özdeğerler hesaplanmaktadır. Boyut sayısı belirlenirken gerçek veriden hesaplanan öz değerlerin tesadüfi verilerden hesaplanan öz değerden büyük olduğu nokta referans olarak alınmaktadır. Diğer bir deyişle gerçek veriden hesaplanan i. sıradaki öz değer, tesadüfi veriden hesaplanan i. sıradaki öz değerden daha büyük olduğu basamak sayısı boyut sayısını göstermektedir (Zwick ve Velicer, 1986; O'Connor, 2000; Watkins, 2006; Ladesma ve Valero-Mora, 2007; Piconne, 2009).

Programlara Yönelik Yapılmış Araştırmalar

Knol ve Berger (1991) tarafından TESTFACT, NOHARM gibi çok boyutlu MTK programları ile beş farklı faktörleme tekniğini bir simülasyon çalışmasında karşılaştırmış ve farklı test koşullarında TESTFACT ile NOHARM programlarının benzer performansa sahip olduğunu ancak açık şekilde MAXLOG'a üstün olduğunu ifade etmişlerdir.

Gosz ve Walker (2001), basit ve karmaşık test yapılarıyla TESTFACT ve NOHARM programlarını simülasyon çalışmasıyla karşılaştırmışlardır. NOHARM programının TESTFACT'e göre madde performansı açısından daha iyi sonuçlar verdiğini ifade etmişlerdir.

Tate (2003), Mplus, TESTFACT ve NOHARM programlarıyla boyutluluk analizi yapmış; hem gerçek verilerle hem de simülasyon verileriyle bu programları karşılaştırmıştır. Hem gerçek veri setleriyle hem de üretilmiş veri setleriyle üç programın da benzer faktör yapılarını ortaya çıkardığı ifade edilmiştir.

Stone ve Yeh (2006), çoktan seçmeli test maddelerinin boyutluluğunun değerlendirilmesinde Mplus, NOHARM ve TESTFACT programlarını karşılaştırmıştır. Çalışmadan elde edilen sonuçlara göre şansla cevaplama olasılığına sahip veri setlerine yönelik üç programda benzer faktör çözümleri üretmiştir.

Finch ve Habing (2005), DETECT ve NOHARM programlarını bir simülasyon çalışmasıyla karşılaştırmışlardır. Farklı test koşulları (örtük yetenek dağılımları, test uzunluğu, örneklem büyüklüğü, boyutlar arasındaki korelasyon) kullanılarak bir simülasyon çalışmasıyla DETECT ve NOHARM programının maddeleri kümeleme, ve hangi maddenin hangi boyutta yer aldığını belirleme özellikleri araştırmacılar tarafından incelenmiştir. Araştırmacılar, NOHARM programının boyut sayısını daha doğru belirlediğini, ancak hangi maddenin hangi boyutta yer aldığını DETECT programına göre daha hatalı belirlediğini ifade etmişlerdir.

Finch ve Habing (2007) tarafından DIMTEST ve NOHARM programlarının sonuçları, farklı test uzunluğu, farklı örneklem büyüklükleri, şansla cevaplama olasılığının varlığı ve yokluğu, farklı yetenek dağılımları değişkenleri kapsamında incelenmiştir. DIMTEST ile NOHARM'a dayalı istatistiklerin birinci tip hatalarının düşük olduğu, benzer ki-kare istatistiği değerlerine sahip oldukları görülmüştür. Ancak şansla cevaplama olasılığının söz konusu olduğu durumlarda DIMTEST programının daha düşük birinci tip hataya sahip olduğu, diğer bir deyişle daha iyi sonuçlar ürettiği bulgusuna ulaşılmıştır.

Gessaroli ve Champlain (1996) tarafından, NOHARM ve DIMTEST programları farklı örneklem büyüklüğü (500, 1000) ve test uzunluğu (15, 30, 45) gibi test koşullarıyla karşılaştırılmıştır. Çalışmalarında test koşullarının çoğunda NOHARM programının ürettiği yaklaşık ki-kare istatistiğinin çoğu durumda DIMTEST programının ürettiği Stout'un T istatistiğine benzer sonuçlar verdiği, ancak küçük örneklem büyüküğü ve test uzunluğuyla NOHARM programının daha iyi sonuçlar ürettiği ifade edilmiştir.

Yeh (2007) tarafından şans başarısının mümkün olduğu çoktan seçmeli test maddeleriyle Mplus ve TESTFACT programları kullanılarak madde ayırt edicilik düzeyleri ve boyutlar arasındaki korelasyonda manipüle edilerek boyutluluğun analizinde kullanılan dört indeks (varyans oranı, paralel analiz, RMSR indirgeme katsayısı ve ki-kare farklılık testi) karşılaştırılmıştır. Çalışmada hem simülasyon hem de TIMSS 2003 verisiyle uygulama yapılmıştır. Araştırmada, TESTFACT programının şansla cevaplamanın söz konusu olduğu çoğu koşulda Mplus programına göre daha iyi sonuç verdiği, RMSR indirgeme indeksinin ve varyans oranının Mplus ile daha doğru kestirildiği ancak ki-kare testi ve paralel analizin TESTFACT ile daha doğru uygulandığı bulgusuna ulaşılmıştır. Çalışmada boyutluluk analizinde TESTFACT programının Mplus programına göre daha tutarlı sonuçlar ürettiği ifade edilmiştir.

Asparouhov ve Muthen (2012) tarafından, Mplus ve IRTPRO programları kullanılarak açımlayıcı ve doğrulayıcı faktör analizi için kestirim teknikleri (maksimum olabilirlik, MH-RM, Bayes tekniği) simülasyon çalışmasıyla karşılaştırılmıştır. Ortalama yanlılık ve hata değerlerinin iki program ve programlara ilişkin kestirim teknikleri açısından hem iki faktörlü doğrulayıcı model için hem de açımlayıcı faktör analiziyle birbirine oldukça yakın değerler aldığı bulgusuna ulaşılmıştır.

Yavuz (2014) tarafından, BMIRT programıyla madde parametresi kestiriminde kullanılan MCMC tekniği, flexMIRT program tarafından kullanılan BA-EM ve MH-RM teknikleri çeşitli test koşullarıyla (örneklem büyüklüğü, test uzunluğu, boyutlar arasındaki korelasyon ve boyut sayısı) karşılaştırılmıştır. Özellikle büyük örneklemlerle ve test uzunluklarıyla iki programında çok düşük hatalarla madde parametresi kestirimi yaptığı, boyut sayısı fazla olduğunda (örneğin 5 iken) flexMIRT ile uygulanan BA-EM tekniğinin madde parametresi kestirimi yapamadığı, MH-RM ve MCMC tekniğinin benzer doğrulukta kestirimler yaptığı bulgusuna ulaşılmıştır. Özellikle örneklem büyüklüğünün çok boyutlu verilerle madde parametresi kestiriminde önemli olduğu bulgusuna ulaşılmıştır.

Hattie, Krakowski, Rogers ve Swaminathan (1996) tarafından, DIMTEST programıyla hesaplanan Stout' un T istatistiği tek boyutlu ve çok boyutlu verilerle farklı test koşullarında incelenmiş ve bu istatistiğin telafi edici ve yaklaşık telafi edici çok boyutlu modellerde duyarlı ölçüm yaptığı bulgusuna ulaşılmıştır. Mearai Robin ve Sireci (2000) ise testteki madde sayısı az olduğunda boyutluluğun analizinde DIMTEST'in hatalı sonuçlar ürettiğini ifade etmiştir. DIMTEST programının 25'den küçük maddeye sahip testlerle ve 500'den küçük örneklem büyüklükleriyle kötü

sonuçlar verdiği farklı araştırmacılar tarafından da ifade edilmiştir (De Champlin ve Gessaroli, 1991; De Champlin ve Tang, 1993; Finch ve Habing, 2007).

Svetina (2011) tarafından, telafi edici ve telafi edici olmayan çok boyutlu madde tepki modellerini karmaşık test yapısı ile incelenmiş ve bu modellerin boyutluluğunu değerlendirmek için şartlı kovaryanslara dayalı DETECT ve faktör analitik yaklaşımlara dayalı NOHARM programlarını karşılaştırmıştır. Bu araştırmanın sonuçlarına göre iki ve üç boyutlu telafi edici modeller için DETECT temelli tekniklerin NOHARM'a göre daha iyi sonuç verdikleri ifade edilmiştir. Araştırmacı tarafından DETECT temelli tekniklerin özellikle boyutlar arasındaki korelasyon 0.60 ve altında olduğu durumlarda, model %30 ve daha düşük oranda karmaşık olduğunda daha iyi sonuç ürettiği bulgusuna ulaşmıştır. Ayrıca DETECT temelli tekniklerin örneklem büyüklüğü arttıkça daha iyi sonuçlar ürettiği ifade edilmiştir. Ancak modele ilişkin test yapısı karmaşık hale geldikçe DETECT programı ile kötü sonuçlar elde edildiği bulgusuna ulaşılmıştır. DETECT programının özellikle basit test yapısındaki maddelerin sınıflandırılmasında tutarlı sonuçlar ürettiği, karmaşık test yapısındaki maddelerin sınıflandırılmasında daha az tutarlı sonuçlar ürettiği ifade edilmiştir.

Meara, Robin ve Sireci (2000) tarafından, çok boyutlu ölçekleme yaklaşımı ile DIMTEST programı iki kategorili verilerin boyutluluğunu değerlendirmek üzere karşılaştırılmıştır. Test uzunlukları ve boyutlar arasındaki korelasyonlar farklılaştırılarak tek ve çok boyutlu veri setleri üretilmiş ve çalışmada elde edilen en önemli bulgulardan biri DIMTEST programının çoğu test koşulunda iyi sonuçlar verdiği ancak testteki madde sayısı az olduğunda hatalı sonuç ürettiğidir.

Deng ve Ansley (2000) tarafından telafi edici ve telafi edici olmayan çok boyutlu modellerin belirlenmesi için DIMTEST programı kullanılmıştır. Örneklem büyüklüğü arttıkça ve boyutlar arasındaki korelasyon azaladıkça telafi edici olmayan çok boyutluluğu daha doğru belirlediği ifade edilmiştir. Araştırmacılar DIMTEST'in telafi edici çok boyutlu veri setlerinin boyutluluğunu belirlemede hatalı sonuçlar ürettiğini ifade etmiştir

SONUÇLAR ve TARTIŞMA

Bu çalışmada çok boyutlu veri setleri ve modelleri ile boyutluluk analizini yapma olanağı tanıyan çeşitli paket programlar tanıtılmıştır. Programların ulaşılabilirlikleri, ne tür veri setleri ve modellerle analiz yapabildikleri, programların teknik özellikleri, çıktı dosyalarında bulunan indekslerin nasıl yorumlandıkları gibi programlara ait önemli özellikler açıklanmaya çalışılmıştır. Son olarak programların performanslarını karşılaştıran çeşitli araştrma sonuçlarına yer verilmiştir. İnceleme sonuçlarına göre programların çok sayıda özellik açısından farklılaştıkları, avantajlı ve dezavantajlı yönlerinin olduğu söylenebilir. Örneğin araştırmacıların ve uygulayıcıların veri setlerinde kayıp veri olması durumunda bazı programlarla analizler yapma olanağı varken diğerlerinde yoktur. Araştırmacıların programların maliyetleri konusunda da fikir sahibi olmaları önemlidir ve bazı programları (NOHARM, DIMPACK (DETECT, DIMTEST, CCPROX/HAC), BMIRT, Paralel Analiz, Velicer'in MAP Testi) ücretsiz edinmek mümkünken bazıları (flexMIRT, TESTFACT, Mplus, IRTPRO) ücretlidir. Ayrıca veri setlerin iki ve çok kategorili olup olmaması, örneklem büyüklüğü önemli değişkenlerdir. Eğer veri setleri parametrik olmayan özellikte iseler tercih edilmesi gereken programlar da parametrik olmayan yaklaşımlara dayalı olan programlardır (DETECT, DIMTEST, CCPROX/HAC). Program seçiminde önemli ölçütlerden biri de programların kullandıkları kestirim teknikleridir. Örneğin, eğer Bayesyen yaklaşıma dayalı bir kestirim tekniği tercih edildiyse (örneğin BMIRT programı ile uygulanabilen MCMC) bu tekniğin kestirim için daha uzun süreler gerektirdiği göz önünde bulundurulmalıdır. Öte taraftan programların kullanım koşulları da önemlidir, örneğin dos komutuyla çalışan veya yazılım dili zor olan ve ara yüzü olmayan bir programa aşina olmayan araştırmacılar programı kullanma konusunda zorluk yaşayabilirler, bu durumda ara yüzü olan bir programların kullanılması kolaylık sağlayacaktır. Son olarak hangi test koşullarında hangi programın performansı daha iyidir sorusunun cevabı çeşitli simülasyon çalışmalarında bulunmaktadır. Bu nedenle araştırmacıların test koşullarına yönelik yapılmış simülasyon çalışmalarına yönelik iyi bir alanyazın çalışması yapmaları katkı sağlayıcı olabilmektedir.

KAYNAKÇA

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13*, 113-127.

Ackerman, T. A. (1996) Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement, 20*, 311-329.

Ackerman, T. A., Gierl, M. J., & Walker, C.M. (2003) Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, *22*, 37–51.

Adams, R. J., Wilson, M. R., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1-23.

Ansley, R. A., & Forsyth, T. N. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, 9, 37-48.

Asparouhov, T. & Muthén, B. (2012). Using Mplus TECH11 and TECH14 to test the number of latent classes. *Mplus Web Notes*, *14*, 22.

Beguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some modelfit analysis of multidimensional IRT models. *Psychometrika*, *66*, 541–562.

Bentler, P.M., & Bonnet, D.C. (1980). Significance Tests and Goodness of Fit in the Analysis of Covariance Structures. *Psychological Bulletin*, *88* (3), 588-606.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Applications of an EM algorithm. *Psychometrik*a *46*, 443-459.

Bock, R. D., Gibbons, R., & Muraki, E. (1988) Full information item factor analysis. *Applied Psychological Measurement 12*, 261-280.

Bock, R. D., & Schilling, S. G. (2003). IRT based item factor analysis. In M du Toit (ed) *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*, 584-591. Scientific Software International, Lincolnwood, IL.

Bock, R. D, Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). *TESTFACT 4.0* [Computer software and manual]. Lincolnwood, IL: Scientific Software International.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement*, *29*, 395–414.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168.

Byrne, B.M. (1998). *Structural Equation Modeling with LISREL, PRELIS and SIMPLIS: Basic Concepts, Applications and Programming*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Cai, L., du Toit, S. H. C., & Thissen, D. (2011). *IRTPRO: Flexible, Multidimensional, Multiple Categorical IRT Modeling*. Scientfic Software International.

Cai, L. (2013). *flexMIRT version 2.00: A numerical engine for flexible multilevel multidimensional item analysis and test scoring. [Computer software]*. Chapel Hill, NC: Vector Psychometric Group.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265-289.

Chib, S. & Greenberg, E. (1995). Understanding the Metropolis Hastings Algorithm. *American Statistical Journal*, *49*, 327–335.

Crawford, A. V., Green, S. B., Levy, R., Lo, W. J., Scott, L., Svetina, D., & Thompson, M. S. (2010). Evaluation of parallel analysis methods for determining the number of factors. *Educational and Psychological Methods*, *70*, 885-901.

DeChamplain, A. F. & Gessaroli, M. E. (1991, April). *Assessing test dimensionality using an index based on non-linear factor analysis*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. 334 235).

DeChamplain, A. & Tang, L. (1993, April). *The effect of non-normal ability distributions on the assessment of dimensionality*. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

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

Deng, H. & Ansley, T. (2000, April). *Detecting compensatory and noncompensatory multidimensionality using DIMTEST*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA (ERIC Document Reproduction Service No. 445 029. Retrieved from http://catalogue.nla.gov.au/Record/5673279.

Deng, N., Han, K., T. & Hambleton, R., K., (2013). A Reviewof DIMPACK Version 1.0: Conditional Covariance Based Test Dimensionality Analysis Package. *Applied Psychological Measurement, 37* (2), 162-172.

Diamantopoulos, A. and Siguaw, J.A. (2000). *Introducing LISREL*. London: Sage Publications.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189-199.

du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International.

Embretson, S. E. (1997). Multicomponent response models. In W. J. Van der Linden R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305–321). New York: Springer Verlag.

Finch, H., & Habing, B. (2005). Comparison of NOHARM and DETECT in item cluster recovery: Counting dimensions and allocating items. *Journal of Educational Measurement, 42*, 149-169.

Finch, H., & Habing, B. (2007). Performance of DIMTEST- and NOHARM-based statistics for testing unidimensionality. *Applied Psychological Measurement, 31*, 292-307.

Fraser, C. (1988). *NOHARM II: A Fortran program for fitting unidimensional and multidimensional normal ogive models of latent trait theory* [Software]. Armidale, New South Wales: University of New England, Centre for Behavioral Studies.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23*, 267-269.

Gamerman, D. & Lopes, H. F. (2006) *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Second Edition. London: Chapman & Hall/CRC Press.

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2011). Performance of Velicer's Minimum Average Partial Factor Retention Method with Categorical Variables. *Educational and Psychological Measurement, 71* (3), 551-570.

Gelman, A., Meng, X. L. & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statist. Sinica 6*, 733–760.

Gessaroli, M. E., & De Champlain, A. F. (1996). Using an approximate chi-square statistic to test the number of dimensions underlying the responses to a set of items. *Journal of Educational Measurement, 33*, 157-192.

Gosz, J. & Walker, C. M. (2001). *An Empirical Comparison of Multidimensional Item Response Data Using TESTFACT and NOHARM*. Annual meeting of the American Educational Research Association, New Orleans, LA, USA.

Hattie, J., Krakowski, K., Rogers, J., & Swaminathan, H. (1996). An assessment of Stout's index of essential dimensionality. *Applied Psychological Measurement, 20*, 1-14.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7* (2), 191-205.

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations on the unidimensionality assumption. *Journal of Educational Statistics, 11*, 91–115.

Horn, J. L. (1965). A rationale and test for rhe number of factors in factor analysis. *Psychometrika, 30* (2). 179-185.

Houts, C. R., & Cai, L. (2013). *flexMIRT user's manual version 2.0: flexible multilevel multidimensional item analysis and test scoring.* Chapel Hill, NC: Vector Psychometric Group.

Hu, L.T., & Bentler, P.M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Structural Equation Modeling, 6* (1), 1-55.

Jang, E. E., & Roussos, L. A. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based non-parametric approach. *Journal of Educational Measurement, 44* (1), 1-21.

Jöreskog, K., & Sörbom, D. (1993). *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language*. Chicago, IL: Scientific Software International Inc.

Justicia, F., Pichardo, M. C., Cano, F., Berben, A. B. G., & De la Fuente, J. (2008). The revised two- factor study process questionnaire (R-SPQ-2F). Exploratory and confimatory factor analyses at item level. *Europen Journal of Psychological of Education, 23* (3), 355-372.

Kim, H.R. (1994). *New techniques for the dimensionality assessment of standardized test data*. (Doctoral Dissertation, University of Illinois at Urbana—Champaign).

Kenny, D.A., & McCoach, D.B. (2003). Effect of the Number of Variables on Measures of Fit in Structural Equation Modeling. *Structural Equation Modeling, 10* (3), 333-51.

Kline, R. B. (2005). *Principles and Practice of Structural Equation Modeling* (2nd ed.). New York: Guilford. 366 pp., ISBN 978-1-57230-690-5.

Knol, D. L. & Berger, M. P. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research*. *26* (3), 457-477.

Ladesma, R.D., & Valero- Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment. Research and Evaluation*, *12*, 1-11.

MacCallum, R.C., Browne, M.W., & Sugawara, H., M. (1996). Power Analysis and Determination of Sample Size for Covariance Structure Modeling. *Psychological Method*s, *1* (2), 130-49.

McDonald, R. P. (1962). A general approach to nonlinear factor analysis. *Psychometrika*, *27*, 398-415.

McDonald, R. P. (1967). *Nonlinear factor analysis [Psychometric Monographs, No. 15]*. Chicago: University of Chicago Press.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*, 100-117.

McDonald, R. P. (1997). *Normal-ogive multidimensional model*. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257-269). New York, NY: Springer-Verlag.

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement*, *24*, 99-114.

Meara, K., Robin, F., & Sireci, S.G. (2000). Using Multidimensional Scaling to Assess the Dimensionality of Dichotomous Item Data, *Multivariate Behavioral Research*, *35* (2), 229–259.

Munroe, A., & Pearson, C. (2006). The Munroe Multicultural Attitude Scale Questionnaire: A new instrument for multicultural studies. *Educational and Psychological Measurement*, *66*, 819-834.

Muthén, L.K. & Muthén, B.O. (1998-2012). *Mplus User's Guide. Seventh Edition.* Los Angeles, CA: Muthén & Muthén.

Nelson, J. M., Canivez, G. L., Lindstrom, W., & Hatt, C. V. (2007). Higher-order exploratory factor analysis of the Reynolds Intellectual Assessment Scales with a referred sample. *Journal of School Psychology*, *45*, 439–456.

Nandakumar, R. & Stout, W. (1993). Refinements of stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics, 18* (1), 41-68.

Nonparametric Dimensionality Assessment Package (2006)., Champaign, IL: William Stout Institute for Measurement.

O'Connor, B.P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, *32*, 396-402.

Özer Özkan, Y., & Acar Güvendir M. (2014). Türkiye'de Uygulanan Geniş Ölçekli Testlerin Çok Boyutluluğunun Analizi. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi, 29,* 31-47.

Piccone, A.V. (2009). *A comparison Of Three Computational Procedures for Solving the Number of factors problem in exploratory factor analysis*. (Doctoral dissertation. University of Northern Colorado).

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *27*, 25-36.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. Springer-Verlag, New York.

Roussos, L. A. (1992). *Hierarchical agglomerative clustering computer program user's manual*. (Doctoral Dissertation, University of Illinois, Urbana-Champaign).

Roussos, L. A., Stout, W. F., & Marden, J. I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement*, *35*, 1-30.

Roussos, L.A. & Ozbek, O. (2006). Formulation of the DETECT population parameter and evaluation of DETECT estimator bias. *Journal of Educational Measurement, 43,* 215-243.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the Reliability of Testlet-Based Tests. *Journal of Educational Measurement*, *28* (3), 237-247.

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

Steiger, J.H. (1990). Structural model evaluation and modification. *Multivariate Behavioral Research*, *25*, 214-12.

Steiger, J.H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, *42* (5), 893-98.

Stone, C. A., & Yeh, C. C. (2006). Assessing the dimensionality and factor structure of multiple-choice exams: An empirical comparison of methods using the Multistate Bar Examination. *Educationa and Psychological Measurement, 66*, 193–214.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589-617.

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

Stout, W. (2006). *DIMPACK 1.0.* Chicago, IL: Assessment Systems Corporation.

Storch, E. A., Murphy, T. K., Bagner, D. M., Johns, N., Baumeister, A., & Goodman, W. K. (2006). Reliability and Validity of the child behavior checklist obsessive-compulsive scale. *Psychistry Research*, *129*, 91-98.

Svetina, D. (2011). *Assessing Dimensionality in Complex Data Structures: A Performance Comparison of DETECT and NOHARM Procedures.* (Doctoral dissertation, Arizona State University). Retrieved from http://repository.asu.edu/attachments/56763/content.

Svetina, D., & Levy, R. (2012). An Overview of Software for Conducting Dimensionality Assessment in Multidimensional Models. *Applied Psychological Measurement, 36* (8), 659-669.

Tabachnick, B.G., & Fidell, L.S. (2007*). Using Multivariate Statistics (5th ed.)*. New York: Allyn and Bacon.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27*, 159-203.

Velicer, W.F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrka, 41*, 321-327.

Velicer, W. F., Eaton, C.A., & Fava, J.L. (2000). *Construct Explication through Factor or Component Analysis: A Review and Evaluation of Alternative Procedures for Determining the Number of Factors or Components*. In Goffin, R. D., & Helmes, E. (Eds.), *Problems and Solutions in Human Assessment: Honoring Douglas Jackson at Seventy.* Boston: Kluwer. (pp. 41-71).

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement, 40*, 255-275.

Watkins, M. W. (2006). Determining parallel analysis criteria. *Journal of Modern Applied Statistical Methods*, *5*, 344-346.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1986). *The effects of two-dimensional data on unidimensional IRT parameter estimates*. Annual meeting of the *American Educational Research Association*, San Francisco CA.

Wheaton, B., Muthen, B., Alwin, D., F., & Summers, G. (1977). Assessing Reliability and Stability in Panel Models. *Sociological Methodology, 8* (1), 84-136.

Yavuz, G. (2014). Çok Boyutlu Madde Tepki Kuramı Modelleri ve Programları için Karşılaştırmalı Analizler. (Yayımlanmamış Doktora tezi, Hacettepe Üniversitesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, Ankara).

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213.

Yao, L. (2003). *BMIRT: Bayesian multivariate item response theory*. CTB/McGraw-Hill, Monterey, CA.

Yao, L., & Boughton, K. A. (2009). Multidimensional Linking for Tests with Mixed Item Types. *Journal of Educational Measurement, 46* (2), 177-197.

Yao, L. (2013). *The BMIRT Toolkit*. Defense Manpower Data Center, DoD Center Monterey Bay,US.

Yeh, C. (2007). *The effect of guessing on assessing dimensionality in multiple-choice tests: A Monte Carlo study with application*. (Doctoral Dissertation. University of Pittsburg). Retrieved from http://d-scholarship.pitt.edu/7286/.

Zhang, J., & Stout, W. (1999).The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 213-249.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the numberof components to retain. *Psychological Bulletin, 99*, 432-442.

EXTENDED ABSTRACT

The assumption of unidimensionality tests and test items has become more questionable after 1980s (Ackerman, 1989; Ansley & Forsyth, 1985; Drasgow & Parsons, 1983; Harrison, 1986; Way, Ansley & Forsyth, 1986). Especially in recent years, there is substantial demand for multidimensional data sets and their various test applications (such as equating test scores, subtest score reporting, differential item functioning). There are various software currently available for assessment of dimensionality. The researchers, who want to make their analysis with multidimensional data sets and models, must to review the software packages in terms of their capabilities. There are important differences between software packages such as their technical details, their capabilities to conduct dimensionality

assessment, types of data, types of models, estimation techniques, estimation time, availability of them.

In this study, some popular software packages (Mplus, NOHARM, TESTFACT, IRTPRO, flexMIRT, BMIRT, DIMTEST, DETECT, CCPROX/HAC, Velicer's MAP test and Parallel Analysis) which are used for analyzing test dimensionality and used for various test applications with multidimensional data sets, was investigated. The software packages was investigated in terms of their availability, their capabilities of data sets, models, estimation techniques. The software packages' dimensionality analyses approach (confirmatory, exploratory) was investigated. Also assessing of the indices which can be gained from software packages' outputs was investigated. Finally the simulation studies in which the packages are compared in terms of their various properties was reviewed.

Mplus software was developed by Muthen (1998-2012). There are a pdf file and website which include using and availability of the software (https://www.statmodel.com/ ugexcerpts.shtml). Mplus offers researchers a wide choice of models, estimators, and algorithms in a program that has an easy-to-use interface and graphical displays of data and analysis results. Mplus allows the analysis of both cross-sectional and longitudinal data, single-level and multilevel data and data that come from different populations with either observed or unobserved heterogeneity Mplus has extensive capabilities for Monte Carlo simulation studies, where data can be generated and analyzed according to any of the models included in the program. Mplus can can be used in either exploratory and condirmatory factor analysis.

NOHARM (Normal Ogive Harmonic Analysis Robust Method) was developed by McDonald (1967) as a model and was developed as software package by Fraser (1998). The detailed information about using software can be accessed with the papers which has written by McDonald (1962; 1967; 1981; 1997; 2000). NOHARM can be used with unidimensional and multidimensional normal ogive models. NOHARM contains options for orthogonal and oblique rotations for exploratory factor solutions and it can be used in confirmatory factor analysis. NOHARM output for confirmatory factor analysis of dimensionality includes RMSR and Tanaka's goodness-of-fit index.

TESTFACT (Full Information Item Factor Analysis) was developed by Bock, Gibbons and Muraki (1998) as a model and the 4.0 version of TESTFACT package was developed by Gibons, Schiling, Muraki, Wilson and Wood (2003). TESTFACT supports analyses of polytomous and dichotomous data, missingness and proceeds by computing tetracoric correlation matrix. TESTFACT contains options for orthogonal and oblique rotations for exploratory factor solutions and it can be used in confirmatory factor analysis.

IRTPRO was developed by Cai, du Toit and Thissen (2011). IRTPRO offers exploratory and confirmatory analysis with a large number of unidimensional and multidimensional models (One, two and three parameter logistic models, Graded model, Generalized Partial Credit and Nominal) and data sets.

flexMIRT 2.0 was developed by Cai (2013). flexMIRT has some of the richest psychometric and statistical features but it is not free avaliable. flexMIRT can estimate item parameter for a wide range of unidimensional and multidimensional models (Rasch, one, two, three parameter logistic models, Graded model, Generalized Partial Credit and Nominal) with exploratory and confirmatory approach. flexMIRT fits a variety of unidimensional and multidimensional IRT models to single-level and multilevel data using maximum marginal likelihood or modal Bayes via Bock-Aitkin EM (with generalized dimension reduction) or MH-RM estimation algorithms.

BMIRT was developed by Yao (2003). One of the advantage of the software is that it is free avaliable. BMIRT (Yao, 2003) (Bayesian Multivariate Item Response Theory) that implements Markov Chain Monte Carlo (MCMC) methods using the Metropolis–Hastings sampling algorithm to estimate item, examinee, and population distribution parameters for a set of MIRT models for both dichotomously and polytomously scored test items. Both confirmatory and exploratory item factor analysis are possible. The program can perform unidimensional or multidimensional calibrations. It can operate on a single group or multiple groups. It can fit dichotomous or polytomous models (along with mixed models), including the three-parameter logistic model, the two-parameter logistic model, the Rasch model, the generalized two-parameter partial credit model, the testlet model, the graded response model, and the higher-order IRT model.

Nonparametric Dimensionality Assessment Package contains 5 different programs (DIMTEST v.2.1, ATFIND v.1.3, DETECT v.2.1, CCPROX and HAC) and it was developed by William Stout Institute (2006). DIMTEST is a nonparametric hypothesis testing procedure to test the assumptions of unidimensionality and local independence. When simple structure multidimensionality is even just approximately present, the DETECT procedure allows for a complete nonparametric description of the multidimensionality. The DETECT procedure begins by calculating the mean conditional covariance between all possible pairs of test items. CCPROX and HAC can be used together to conduct an exploratory nonparametric dimensionality analysis. CCPROX/HAC analysis computes proximity matrices and assesses dimensionality with hierarchical cluster analysis.

Altough they are highly recommended by researchers and recent studies, popular software packages (such as SPSS, SAS), can not permit analysis of dimensionality with paralellel analysis and Velicer's MAP (manimum average partial) test directly. Because of this the using of them are not very common with dimensionality analysis. The MAP test and parallel analysis can be implemented with the use of available syntax (O'Connor, 2000).

In summary, there are important differences between software package which can be used for assessing dimensionality, it appeared that each of the software packages had both advantages and disadvantages, and that users need to show care in making their choices. The researchers must know the software capabilities (such as estimation technique, type of data, models, availability, estimation technique and estimation time).

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2015; 6(2); 293-312

Ekler

Ek 1. *Programların Karşılaştırılması*

Bilgisayar Programları	Erişim	Web sitesi	Veri Türleri	Kestirim Teknikleri	Boyutluluk Analizi	Sınırlılığı
Mplus	Ücretli	https://www.statmodel.com/ugexcerpts.shtml	-Tek ve çok düzeyli veri, kayıp veri, her türlü ölçekten elde edilmiş veri setleri ve kombinasyonları -Monte Carlo Simülasyonu	-Bayesyen ve olabilirlik kestirimleri	-Açımlayıcı, Doğrulaycı	-Mplus için komut yazılması gerekmekte
NOHARM	Ücretsiz	http://noharm.software.informer.com/download/	-İki Kategorili	-Polinomiyal ve ağırlıklandırılmamış en küçük Kareler Yöntemi	-Açımlayıcı, Doğrulaycı	-Şans ve yetenek parametresi kestirimi yapamamaktadır, Çok kategorili ve kayıp verilerle çalışmamaktadır
TESTFACT	Ücretli	http://www.ssicentral.com/irt/index.html	- İki ve çok kategorili veriler, kayıp veriye sahip veri setleri	- Marjinal maksimum olabilirlik kestirimini ve ayrıca yetenek parametreleri için Bayesyen algoritması	-Açımlayıcı, Doğrulaycı (sadece iki faktörlü)	-Şans parametresi kestirimi yapamamaktadır
IRTPRO	Ücretli	http://www.ssicentral.com/irt/index.html	- İki ve çok kategorili veriler, kayıp veriye sahip veri setleri	-Maksimum olabilirlik, MAP, Bock Aitkin EM Bifactor EM, Generalized Dimension Reduction EM, MHRM	-Açımlayıcı, Doğrulaycı	-Doğulayıcı yaklaşımda sadece iki faktör model yapısı kullanılmakta
flexMIRT	Ücretli	https://www.vpgcentral.com/irt-software/	-İki ve çok kategorili veriler, kayıp ve erişilmemiş veriye sahip veri setleri	-BA-EM ve MH-RM teknikleri, Maksimum olabilirlik, MAP, EAP	-Doğrulayıcı	-Açımlayıcı yaklaşımla boyutluluk analizi yapamamakta

* Yrd. Doç.Dr., Adıyaman Üniversitesi, Eğitim Fakültesi, Adıyaman-Türkiye, e-posta: gyavuz2010@gmail.com
**Doç.Dr.,Hacettepe Üniversitesi, Eğitim Bilimleri Bölümü, Ankara-Türkiye, e-posta: nurid@hacettepe.edu.tr

Geliş Tarihi: 09.09.2015
Kabul Tarihi: 11.20.2015

			-Monte Carlo Simülasyonu			
BMIRT	Ücretsiz	http://www.bmirt.com/6271.html	- İki veya çok kategorili veri setleri	-MCMC	- Açımlayıcı, Doğrulaycı	-MCMC tekniğinin kestirim süresi uzundur, ara yüzü yoktur, DOS komutlarıyla calışır
DIMPACK (DIMTEST, DETECT, CCPROX/HCA)	Ücretsiz	http://psychometrictools.measuredprogress.org/dif2	- İki ve çok kategorili veriler, kayıp ve erişilmemiş veriye sahip veri setler	-Nonparametrik yaklaşımla	- Açımlayıcı, Doğrulaycı	- Maksimum madde sayısı 150 ve maksimum örneklem büyüklüğü 7000,
Velicer'in MAP Testi ve Paralel Analiz	Ücretsiz	https://people.ok.ubc.ca/brioconn/nfactors/nfactors.html	-Çok Kategorili	-Temel bileşenler analizi	-Açımlayıcı	-Programların kullanımı için SPSS veya SAS programlarıyla syntax (betik) yazılması gerekli

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2015; 6(2); 313-330

Applying Item Response Theory Models to Entrance Examination for Graduate Studies: Practical Issues and Insights

Madde Tepki Kuramının Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı'na Uyarlanması: Uygulamadaki Sorunlar ve Öneriler

Okan BULUT *

Abstract
Item response theory is a psychometric framework for the design, analysis, and scaling of standardized assessments, psychological instruments, and other measurement tools. Despite its increasing use in educational and psychological assessments across many countries around the world, it has not been applied to any large-scale assessment in Turkey. The purpose of this study is to investigate the fit of unidimensional item response theory models to the Entrance Examination for Graduate Studies which is a high-stake large-scale assessment in Turkey required for applying to graduate programs in Turkish universities. Model assumptions of item response modeling, such as unidimensionality, local independence, and measurement invariance, are examined. Also, model-specific assumptions, such as equal item discrimination and minimal guessing, are evaluated. Findings of this study suggest that the three-parameter IRT model shows the best model-data fit for the Entrance Examination for Graduate Studies. Also, the results of this study highlight potential issues that need to be addressed, such as high omit rates, speededness of the test, and aberrant guessing behaviors.

Key Words: Item response theory, large-scale assessment, test development, model fit.

Öz
Madde tepki kuramı standart testler, psikolojik envanterler ve diğer ölçme aletlerinin tasarımı, analizi ve ölçeklendirilmesinde kullanılan statistiksel bir modeldir. Dünyadaki birçok ülkede madde tepki kuramının ölçme ve değerlendirme alanındaki artan uygulamalarına karşın Türkiye'de bu yöntem geniş ölçekli sınavlara henüz uygulanmamıştır. Bu çalışmanın amacı tek boyutlu madde tepki kuramı modellerinin Türkiye'deki Akademik Personel ve Lisansüstü Eğitimi Giriş (ALES) sınavına uygulanmasını göstermektir. ALES sınavı Türk üniversitelerine yapılan yüksek lisans ve doktora başvuruları ve üniversitelerdeki akademik personelin belirlenmesi gibi birçok önemli alanda kullanılmaktadır. Madde tepki kuramının tek boyutluluk ve yerel bağımsızlık gibi temel varsayımlarının yanında belirli modellere özgü eşit madde ayırt edicilik gücü ve soruların minimum ölçüde tahmini gibi ek varsayımlar da incelenmiştir. Çalışmanın sonuçları üç parametreli lojistik modelin ALES için en uygun madde Tepki kuramı modeli olduğunu göstermiştir. ALES'te sınav süresinin yetersizliği, sınava girenlerin bazı soruları yüksek oranda cevapsız olarak geçmesi ve tipik olmayan soru tahmin davranışlarına dair sorunlara dikkat çekilmiştir.

Anahtar Kelimeler: Madde tepki kuramı, geniş ölçekli test, test geliştirme, model uyumu.

INTRODUCTION

Testing in education and psychology is mainly an attempt to measure a person's knowledge, intelligence, or other characteristics in a systematic and reliable way. Standardized testing has been the most useful evaluation method for measuring latent traits such as achievement, aptitude, and cognitive abilities. Standardized tests can provide decision-makers with useful information about applicants who apply for an undergraduate program in a university, try to obtain a driver's license, or apply for a job. In many testing situations, a complex measurement framework must be employed to define the

* Assistant Professor, University of Alberta, Faculty of Education, Edmonton, AB, CANADA, e-mail: bulut@ualberta.ca.

Geliş Tarihi: 11.09.2015
Kabul Tarihi: 18.11.2015

relationship between a latent trait and item responses, and generalize beyond the single situation in which a measurement is observed.

In educational testing, understanding what it takes to construct useful measures has only been applied in psychometrics (Wright, 1997). Initial methods to construct useful measures were based on the approach of counting concrete events. According to Thorndike (1904), someone who wants to measure a simple thing, such as spelling, is hampered by the fact that there exist no units in to measure. Even though one may observe the ability by the number of words from a list spelled correctly, the inequality of the units is still a serious issue. One might observe signs of spelling ability but would not have measured spelling (Engelhard, 1991). At this point, measurement models differentiate in terms of the use of raw data. There are two popular statistical frameworks for addressing measurement problems such as test development, test score equating, and the identification of biased items: classical test theory (CTT) and item response theory (IRT) (Hambleton & Jones, 1993).

CTT, also known as true score theory, was originally the leading framework for analyzing and developing standardized tests. Since the beginning of the 1970s, IRT has more or less replaced the role that CTT had and is now the major theoretical framework used in this scientific field (Crocker & Algina, 1986; Hambleton & Rogers, 1990; Hambleton, Swaminathan, & Rogers, 1991). The major advantage of CTT is its weak theoretical assumptions, which make CTT easy to apply in many testing situations (Hambleton & Jones, 1993). However, there are two major drawbacks of CTT compared to IRT. First, all item and person statistics derived from CTT are heavily dependent on the sample of test takers and the items used on the test. That is, depending on which test items are used and who takes the test, these statistics will dramatically change from one test administration to another. Second, because CTT focuses on the test-level information, it fails to explain the relationship between items and test scores. The lack of this information poses theoretical difficulties in measurement applications, such as test development, test equating, and test of measurement invariance. Unlike CTT, IRT primarily focuses on the item-level information based on the probabilistic distribution of examinees' success, and thus overcomes the technical issues that CTT has.

Item Response Theory

Item response theory, also known as latent trait theory, is not only a modern test theory, but also the most popular one in educational and psychological testing. IRT requires two major assumptions. First, the performance of an examinee on a test item can be predicted by a set of factors called traits, latent traits, or abilities. Second, the relationship between examinees' item performance and the traits underlying item performance can be described by a monotonically increasing function. In IRT, this function is called "item characteristic function" or "item characteristic curve" (ICC). Based on this function, as the level of latent trait increases, the probability of an examinee giving a correct response to an item increases as well.

An example of ICC is shown in Figure1. The horizontal axis shows the ability (latent trait) scale. The ability in IRT is symbolized by the Greek letter theta (θ). The vertical axis shows the probability of giving a correct response to the item. The difficulty parameter (b) sets the location of the curve on the horizontal axis; it shifts the curve from left to right as the item becomes more difficult. The location of b can be found by dropping a vertical line from the inflection point to the horizontal axis. The slope of the curve is called the item discrimination parameter (a). The a-parameter is found by taking the slope of the line tangent to the ICC at the b-parameter. The steeper the curve, the more discriminating the item is, and the greater its item-total correlation. As the a-parameter decreases, the curve gets flatter until there is virtually no change in the probability across the ability continuum. Items with very low a values are not appropriate for differentiating examinees with low and high abilities, just like items with very low item total correlations. The c parameter is the lower asymptote. It is the lowest point of the curve as it moves to negative infinity on the horizontal axis. The c parameter can be used to model guessing in multiple-choice items.

Model Assumptions in Unidimensional IRT

There are two major assumptions for unidimensional IRT models. These assumptions are unidimensionality and local independence. The unidimensionality assumption requires that there is a single latent trait underlying a set of items. Hambleton et al. (1991) state that this assumption cannot be strictly met because of several cognitive, personality-related, and test-taking factors, such as level of motivation, test anxiety, ability to work quickly, etc. Finding a dominant component or factor affecting test performance is required to meet this assumption.

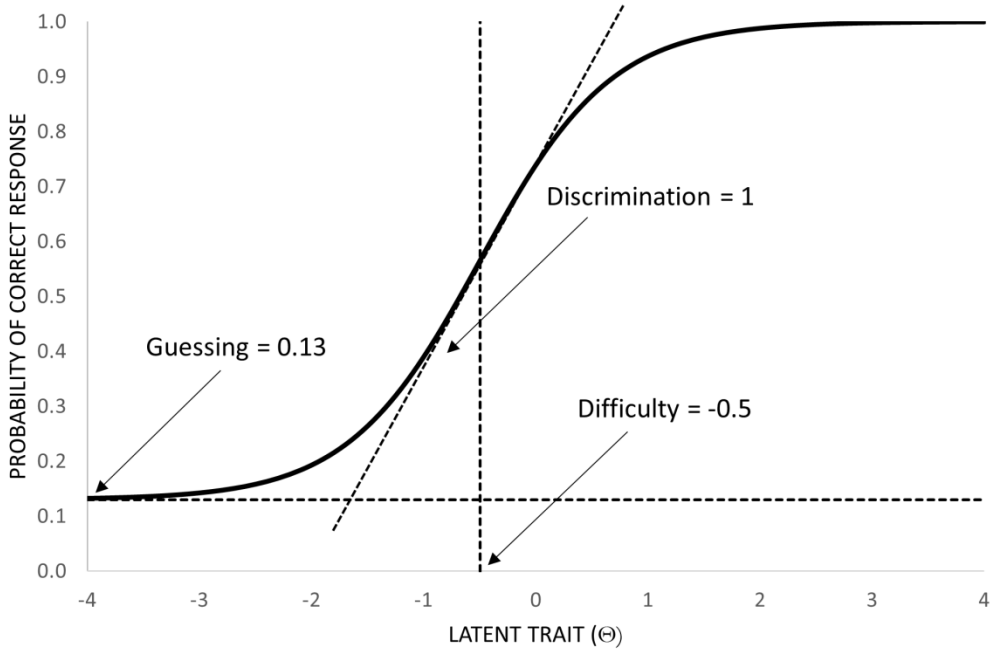


Figure 1. An Example of an Item Characteristic Curve in IRT

The local independence assumption requires the probability of a correct response by an examinee to an item not to be affected by responses given to other items in the test. In other words, after taking examinees' abilities into account, there should be no relationship between examinees' responses to different items. Therefore, high intercorrelations among the items are solely a result of the ability of the test-takers. When the trait level is controlled, local independence implies that no relationship remains between the items (Embretson & Reise, 2000).

When the assumption of unidimensionality is true, local independence is obtained. In this sense, the two concepts are equivalent (Lord, 1980). In addition to unidimensionality and local independence, there are other assumptions essential for both unidimensional and multidimensional IRT models. First, the ICC is a monotonically increasing function of the latent trait, continuous, and smooth (i.e., continuously differentiable), which results in an S-shaped curve (Hambleton et al., 1991; Raykov & Marcoulides, 2010, p. 270). Second, IRT models require invariance of item parameters and the latent trait. Item parameters are assumed to be invariant over different samples or subgroups of examinees from the population for whom the test is intended. Similarly, the latent trait needs to be invariant over different samples of test items from the population of items measuring the target ability (Baker, 1985; Hambleton et al., 1991). Third, the non-speeded test administration assumption requires that all examinees should have enough time to respond to all items in the test. The test cannot be a speeded test with binary scored items (Albanese & Forsyth, 1984). If some of these assumptions are not met, the selected IRT model is very likely not to fit to the item response data due to either poor item fit or poor person fit.

Unidimensional IRT Models

There are three frequently used IRT models for dichotomously scored test items: one-parameter, two-parameter and three-parameter IRT models. These models are the most commonly used IRT models, but there are many others – including models with the 4th parameter or upper asymptote, models that include a parameter for response time, models that include parameters for thresholds on partial credit or rating scale items, and others. The main difference between these models is the number of item parameters (a, b and c parameters as described earlier). Since the item parameters of the models are different, ICCs are also different.

The simplest IRT model is the one-parameter logistic IRT model (also known as the 1PL model). The 1PL model assumes that all of the items have the same item discrimination power and the lower asymptote (i.e., c parameters) is equal to zero for all items. The 1PL model can be shown as

$$P_i(\theta) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}}, \tag{1}$$

where $P_i(\theta)$ is the probability of an examinee with ability θ answering item *i* correctly, b_i is item difficulty parameter for item *i*, and e represents the base of the natural logarithm approximated at 2.178. The *D* in the equation represents a constant adjustment to the model in order to reduce the differences between the logistic IRT model and the normal ogive model to less than .01 (Crocker & Algina, 1986). The value of *D* is usually set to 1.7.

The two-parameter logistic model (also known as the 2PL model) has the same equation as the 1PL model. However, there is an additional parameter, a_i, which represents item discrimination for item *i*. It means that the item discrimination parameter varies across items, as does the item difficulty parameter. The 2PL can be shown as follows:

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}. \tag{2}$$

The three-parameter logistic model (also known as the 3PL model) has also a similar mathematical form. Differently from the first two models, the 3PL model includes a lower asymptote, c_i, which is the pseudo guessing level of item *i*. This additional parameter represents the probability of examinees with low ability giving a correct response to item *i* by chance. The mathematical form of the 3PL model can be written as follows:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}. \tag{3}$$

The three IRT models described above can be considered as variants of each other. Among the three models, the 1PL model is the most restricted model with fixed item discrimination and zero lower asymptote for all items; whereas the 2PL model is relatively less restricted with varying item discrimination parameters and zero lower asymptote for all items. When there is random guessing (i.e., $c_i > 0$), this may result in a certain degree of inflation in the probability of correct response. This type of guessing behavior is more likely to happen among individuals with the lowest ability. The 3PL is the only model that allows for the estimation of item discrimination, item difficulty, and lower asymptote for each item.

Previous IRT Model-Fit Studies

Model-data fit studies of IRT are crucial because they provide information about the appropriateness and the adaptability of the IRT models to psychometric measures such as tests, surveys, and scales. In the literature, there are two lines of research used for investigating model-data fit in IRT. The first one

is the comparison of CTT and IRT frameworks in terms of item and ability parameters by using real data and Monte Carlo simulation studies (Courville, 2005; Fan, 1998, Güler, Uyanık, & Teker, 2014; Hambleton & Jones, 1993; Progar, Sočan, & Peč, 2008). The second one is the model-data fit studies that solely focus on the application of IRT models to different instruments (e.g., tests, surveys, and scales) and provides in-depth investigations.

IRT models have been applied to various types of assessments including achievement tests, language assessments, personality inventories, and psychological instruments. One of the earliest IRT model-fit studies was conducted by Albenese and Forsyth (1984). They analyzed responses of examinees in grade 9 on five subtests of the Iowa Tests of Educational Development to compare the relative robustness of the 1PL, 2PL, and 3PL IRT models. The largest number of misfit items was observed in the 1PL model. Also, the results indicated that the modified 2PL model may provide the best representation of the data. Choi (1989) investigated the appropriateness of IRT models in language testing. The Certificate of English (FCE) from the University of Cambridge and the Test of English as a Foreign Language (TOEFL) were used for this study. The results showed that the listening subtest of the TOEFL did not meet the assumption of unidimensionality while the reading and vocabulary subtests of the FCE were convincingly unidimensional. Furthermore, the 3PL model indicated the best model-data fit for the FCE. Chernyshenko et al. (2001) compared the fit of the 2PL and 3PL IRT models to two personality assessment instruments, the US-English version of the Fifth Edition of the Sixteen Personality Factor Questionnaire and Goldberg's 50 item Big Five Personality measure. The findings of their study suggested that the 2PL and 3PL models fit some scales reasonably well but not others. The negative keyed questions in both personality assessments led to item misfit problems across several subscales.

There have been also several IRT model-fit studies in Turkey. In an early study, Berberoglu (1990) compared the 1PL and 3PL models using the Turkish version of the Group Assessment of Logical Test (GALT) consisting of 36 multiple-choice items. The results indicated that the GALT met the assumption of having a unidimensional latent trait. In addition to the unidimensionality assumption, other IRT assumptions required for the 1PL and 3L models were also met for the GALT. In a similar study, Kilic (1999) investigated the fit of the 1PL, 2PL, and 3PL models to the four subtests of the Student Selection Test (SST) in Turkey. SST is a very high-stakes test taken by high school graduates to enter an undergraduate program in a university in Turkey. The results of this study indicated that the 3PL model fit better than the other two IRT models.

Celik (2001) also conducted a similar study using the Secondary Education Institutions Student Selection and Placement Test in Turkey. The model-fit of the three unidimensional IRT models to this test was investigated. The results indicated that the 3PL model provided a better psychometric presentation of mathematics and science subtests. Önder (2007) investigated the fit of IRT models to the data obtained from ÖZDEBİR ÖSS 2004 D-II Exam Science Test. The result of this study suggested that the most appropriate model data fit was achieved by the 3PL model, followed by the 2PL model. The most recent IRT model-fit study in Turkey was conducted by Teker, Kelecioglu, and Eroglu (2013). They investigated the fit of IRT models to the 2009 administration of Seviye Belirleme Sınavı (SBS) that is a national exam for all 8th grade students in Turkey. Their results indicated that the 3PL model was the most appropriate model for the SBS data.

The review of the IRT model-fit literature shows that the 3PL is typically the best-fitting model for multiple-choice large-scale assessments. Also, the unidimensionality assumption is more prone to be violated than the local independence assumption. In light of findings of earlier studies, this study presents an empirical investigation of IRT model-fit to address the research questions described earlier.

Purpose of the Study

To date, despite their inevitable advantages over CTT, IRT models have not been operationally used in the analysis and decision-making processes of large-scale assessments in Turkey. This study aims to demonstrate the applicability of the IRT framework using a high-stakes assessment in Turkey. The

unidimensional IRT models were applied to the Entrance Examination for Graduate Studies (EEGS) that is a required examination to apply for a graduate program in Turkish universities. First, the assumptions of IRT models were examined. Then, the invariance of item and ability parameter estimates was investigated. Finally, the fit of IRT models for each subtest of the EEGS was examined to test whether the observed and theoretical distributions of IRT models overlap for the subtests of the EEGS.

METHOD

Sample

The data for this study come from the fall administration of the EEGS in 2010. The total number of examinees was 142,178. In this study, a sample of 5,000 examinees was randomly selected from the examinees that completed at least 25% of the EEGS. The descriptive statistics for the selected sample are presented in Table 1. 97.8% of the examinees completed an undergraduate program in a Turkish university and the remaining examined obtained their undergraduate degrees from a university outside of Turkey. Most of the examinees took the EEGS to apply for a graduate program or to be an academic staff in a university.

Table 1. Examinee Characteristics for the Selected Sample from the EEGS

Variable		Frequencies (f)	Percentages (%)
Gender			
	Male	2540	50.8
	Female	2460	49.2
University Location			
	Turkey	4890	97.8
	Foreign	110	2.2
Reason for taking the test			
	To apply for an academic position	872	17.4
	To apply for a graduate program	3924	78.5
	Other	204	4.1

Instrument

In this study, model-data fit analyses were carried out using the data from the 2010 administration of the EEGS. The EEGS is a large-scale assessment in Turkey that is administered twice a year by the Measurement, Selection, and Placement Center (also known as ÖSYM in Turkey). The scores from the EEGS are used for three purposes: 1) to start a graduate program in a university; 2) to determine candidates who will be sent to foreign countries for graduate education with a scholarship; and 3) to determine academic staff such as college instructors, graduate assistants, lecturers, and specialists.

The test is composed of 160 multiple-choice items with five response options. The EEGS consists of three subtests: Verbal, Quantitative 1, and Quantitative 2. The Verbal subtest includes 80 items that measure verbal reasoning abilities. The Quantitative 1 and the Quantitative 2 subtests consist of 40 items that measure mathematical and logical reasoning abilities. The Quantitative 2 subtest covers more advanced mathematical topics than the Quantitative 1 subtest.

Data Analysis

Data analysis of this study consists of three steps. The first step was preliminary data analysis. The purpose of preliminary analysis was to have an in-depth examination of the test items for any potential flaws or extreme values in the data. A CTT-based item analysis was carried out with the *psychometric* package (Fletcher, 2015) in R (R Development Core Team, 2015) for the three subtests of the EEGS. Descriptive statistics (item difficulties, point-biserial correlations, mean test scores, etc.) were obtained for the items and test scores across the subtests.

The second step was the evaluation of model assumptions. The main model assumptions, namely unidimensionality and local independence, were carefully investigated. The assumption of unidimensionality requires that the probability of successful performance by examinees on a set of items can be modeled by a mathematical model that has only one ability parameter (Dorans & Kingston, 1985). Although this is a very important assumption for IRT models, there is no simple way to assess the unidimensionality assumption. Stout's nonparametric DIMTEST (Stout, 1987), Humphreys and Montanelli's (1975) method of parallel analysis, and confirmatory factor analysis (CFA) are the most widely used methods for assessing scale unidimensionality.

In this study, the CFA approach was used to confirm the unidimensional latent structure of the Verbal, Quantitative 1, and Quantitative 2 subtests. A one-factor (i.e., unidimensional) CFA model was fit to each of the three subtests using Mplus 6 (Muthen & Muthen, 1998-2011). A robust weighted least squares (WLS) estimator with a diagonal weight matrix was used as the estimation method. The WLS estimator was selected because when dependent variables are identified as categorical, this estimator yields more accurate factor loading estimates than maximum likelihood and robust maximum likelihood estimators (Li, 2014).

Goodness-of-fit criteria, including root mean square error of approximation (RMSEA), Tucker-Lewis Index (TLI), and comparative fit index (CFI), were used to evaluate model-data fit of the one-factor CFA model for the three EEGS subtests. CFI and TLI are incremental fit indices that assess the relative improvement in fit of the selected model compared with a baseline model. Both indices range between 0.0 and 1.0 with values closer to 1.0 indicating good fit. RMSEA is an absolute fit index that is independent of sample size and thus performs well as an indicator of practical fit. For CFA models, Hu and Bentler (1999) suggested that for categorical data, RMSEA < .06, TLI > .90, and CFI > .90 indicate good fit. Based on these criteria, a satisfactory fit for the one-factor model would suggest that the test has a unidimensional structure.

There are also several methods for assessing local item dependencies in dichotomous data, such as Yen's Q_3 statistic (Yen, 1984) and the G^2 statistic (Bishop, Fienberg, & Holland, 1975; Chen & Thissen, 1997). Conditional inter-item correlations can be also used as a measure of local item independence (Ferrara, Huynh, & Baghi, 1997). In this study, to examine the assumption of local independence, inter-item correlation matrices were evaluated in a restricted range of abilities (i.e., high ability and low ability groups). For selecting low and high ability examinee groups, the 20^{th} and 80^{th} percentiles of total raw scores were used as the cut-off values in each subtest. The zero or close to zero off-diagonal elements of the variance-covariance or the correlation matrix for examinees within a restricted range of ability or test score scale indicate unidimensionality and that the test has met the assumption of local independence (Hambleton et al., 1991; McDonald, 1981).

To check the measurement invariance of the EEGS subtests between male and female examinees, a multi-group CFA framework (Meredith, 1993) was used. A one-factor CFA model for a dichotomous observed response, X_i, for item *i* can be written as follows:

$$X_i = \tau_i + \lambda_i \xi + \varepsilon_i, \quad (4)$$

where τ_i is the intercept for item *i*, λ_i is the factor loading for item *i*, ξ is the latent construct, and ε_i is the residual term for item *i*. To test measurement invariance across male and female examinees, a series of nested multiple group models was assessed.

Table 2 summarizes the four types of measurement invariance tests used in this study. For each test, a constrained model with fixed parameters across male and female examinees was tested against a less constrained model. The nested models were compared using a chi-square difference test as well as several model fit indices. Substantial decrease in goodness of fit between the two models indicates the violation of measurement invariance. Measurement invarance tests were conducted using the lavaan package (Rosseel, 2012) in R (R Core Team, 2015).

Table 2. Summary of Measurement Invariance Tests

Type	Constrained Parameters	Comparison Model
Configural invariance	-	-
Weak invariance	λ_i	Configural invariance
Strong invariance	τ_i, λ_i	Weak invariance
Strict invariance	$\tau_i, \lambda_i, \varepsilon_{ij}$	Strong invariance

After the unidimensionality, local independence, and measurement invariance assumptions were checked, the other model-specific assumptions were also carefully examined. The homogeneous distribution of item discrimination indices obtained from the preliminary item analysis was used to check the assumption of equal discrimination indices of the 1PL model. Performance of low-ability examinees on the most difficult questions was evaluated to check the minimal guessing assumption of the 1PL and 2PL models. The most difficult items were chosen based on the proportion-correct values obtained from the preliminary item analysis. The non-speeded test administration assumption was evaluated based on the percentages of examinees that completed the last five items of each subtest of the EEGS.

The final step of the data analysis was the comparison of model-data fit. The three subtests of the EEGS were calibrated and scored based on the 1PL, 2PL, and 3PL models, respectively. The IRT model estimation was implemented using marginal maximum likelihood estimation in Xcalibre 4.1 (Guyer & Thompson, 2011). The fit of the 1PL, 2PL, and 3PL models was compared using the Likelihood Ratio (LR) test, which is based on -2 times the difference in log-likelihoods from two nested models. The LR statistic can be computed as follows:

$$LR = -2\ln L_C - (-2\ln L_A), \tag{5}$$

where L_C is the log likelihood of the compact model (i.e., the model with fewer item parameters) and L_A is the log likelihoods of the augmented model (i.e., the model with more item parameters).

The LR statistic is approximately distributed as chi-square (χ^2) with degrees of freedom equal to the difference in the number of parameter estimates in the two models. The significant LR statistic indicates that the augmented model fits better than the compact model. Drasgow et al. (1995) suggested that the adequacy of model fit should be also evaluated using graphical methods. In this study, in addition to the LR test for model comparison, both model fit plots at the item and test levels as well as chi-square goodness of fit statistics for individual items were used to examine the fit of IRT models to the EEGS.

RESULTS

Results of Preliminary Analysis

The results of preliminary item analysis are presented in Table 3. The results indicated that Quantitative 2 was the most difficult subtest on average among the three subtests. The average item-total correlations (i.e., point-biserial correlations) demonstrate the discriminatory level of the three subtests between high ability and low ability examinees. Based on the results in Table 3, Quantitative 2 indicated a better discriminatory power than the other two subtests. In addition, the results indicated that all of the items functioned and discriminated well. Therefore, none of the items were excluded from the subsequent analyses. The EEGS indicated high test reliability based on coefficient alpha values obtained from each subtest.

Table 3. Summary Statistics for the Items in the EEGS Subtests

Subtest	*N*	Mean Difficulty	Mean Point-Biserial	Alpha
Verbal	80	0.73	0.43	0.96
Quantitative 1	40	0.76	0.39	0.90
Quantitative 2	40	0.67	0.46	0.93

The results of preliminary analysis at the subtest level are presented in Table 4. Descriptive statistics based on total raw scores are presented for the overall sample and for each gender group separately. The scores from the three subtests had negatively skewed distributions, suggesting that most examinees in the sample obtained high scores in the EEGS. Although the minimum and maximum raw scores did not differ across gender groups, female examinees performed better than male examinees in the Verbal subtest and the male examinees outperformed the female examinees in both Quantitative 1 and 2 subtests. Especially in the Verbal subtest, the distribution of raw scores was negatively skewed as most of the students obtained high test scores.

Table 4. Summary Statistics for the Total Raw Scores from the EEGS Subtests

Subtest	Group	N	M	SD	Min	Max	Skewness	Kurtosis
Verbal	Overall	5000	58.1	15.6	20	80	-0.63	-0.68
	Male	2540	56.9	16.2	20	80	-0.54	-0.87
	Female	2460	59.3	14.8	20	80	-0.72	-0.45
Quantitative 1	Overall	5000	30.4	7.1	10	40	-0.83	-0.07
	Male	2540	31.3	6.9	10	40	-1.03	0.40
	Female	2460	29.4	7.0	10	40	-0.66	-0.36
Quantitative 2	Overall	5000	26.8	8.8	10	40	-0.42	-1.12
	Male	2540	27.7	8.9	10	40	-0.55	-0.98
	Female	2460	25.8	8.7	10	40	-0.31	-1.20

Results of Model Assumptions

The CFA results indicated that all of the three subtests had acceptable levels of model-data fit based on the model-fit criteria described earlier (see Table 5). The satisfactory model-fit for the one-factor CFA model suggested that the unidimensionality assumption of the Verbal, Quantitative 1, and Quantitative 2 subtests was adequately met. However, it should be noted that the model fit indices presented in this study may not be robust against issues such as sample size or missing item responses in the data. Therefore, the use of alternative dimensionality tests is strongly encouraged for a more detailed analysis of the unidimensionality assumption.

Table 5. Model Fit Indices for the One-Factor CFA Model

Subtest	*N* of Items	CFI	TLI	RMSEA
Verbal	80	0.90	0.89	0.01
Quantitative 1	40	0.94	0.98	0.01
Quantitative 2	40	0.97	0.99	0.01

The means of inter-item correlations of high and low ability groups were close to zero across the three subtests (see Table 6). However, the results suggested that some items in the Verbal subtest had relatively higher inter-item correlations than the items in the other two subtests. These items were mostly linked to the same reading passages on the test, which suggests that some of the items may be problematic since they depend on the same content. Therefore, although the local independence assumption was assumed to be met based on inter-item correlations, the likelihood of having locally dependent items based on the reading passages remained as a potential concern for the Verbal subtest.

Table 6. Inter-Item Correlations Obtained from the Low and High Ability Groups

Subtest	Low Ability Group				High Ability Group			
	M	*SD*	Min	Max	*M*	*SD*	Min	Max
Verbal	0.155	0.098	-0.256	0.455	0.180	0.101	-0.299	0.484
Quantitative 1	0.071	0.098	-0.114	0.267	0.074	0.101	-0.194	0.289
Quantitative 2	0.053	0.071	-0.093	0.197	0.055	0.079	-0.101	0.203

To investigate the equal item discrimination assumption for the 1PL model, the frequency distributions of item-total correlations obtained from the preliminary item analysis were analyzed graphically. As seen in Figure 2, the item-total correlations were not homogenously distributed, suggesting that the items may not have an equal discrimination power and that the assumption of equal item discrimination may not be viable for the EEGS.

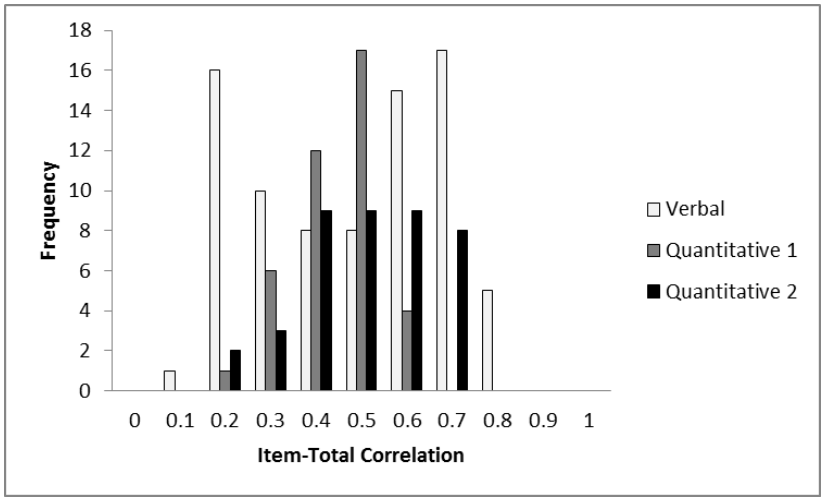


Figure 2. Frequency Distribution of Item-Total Correlations from the EEGS Subtests

The minimal guessing assumption for the 1PL and 2PL models was investigated by examining the performance of low-ability examinees (i.e., 20th percentile and below) on the most difficult items. The most difficult items were identified by selecting 10% of the items with the lowest proportion correct values. This procedure resulted in selecting eight items from the Verbal subtest, and four items from the Quantitative 1 and Quantitative 2 subtests. The results are presented in Table 7. The performance of low-ability examinees was worse than the performance of the overall sample on the difficult items. Low-ability examinees chose to skip the difficult items instead of randomly guessing. Most of the difficult items were mostly the last items on the tests and these items had high omit rates. This finding suggests that although minimal guessing assumption was met in the EEGS, high omit rates may still be a concern.

Table 7. Percentage of Correct Responses on Most Difficult Items by Low-Ability Examinees

Subtests	Items	*P*	Percent Correct	Percent Incorrect	Percent Missing
Verbal	Item 56	0.47	8.0	12.7	79.3
	Item 73	0.47	11.1	5.2	83.7
	Item 75	0.38	7.2	7.3	85.5
	Item 76	0.33	6.4	7.4	86.2
	Item 77	0.41	11.3	4.4	84.3
	Item 78	0.41	10.7	4.2	85.2
	Item 79	0.43	12.2	2.6	85.2
	Item 80	0.31	7.7	6.1	86.2
Quantitative 1	Item 30	0.44	25.6	55.6	18.8
	Item 31	0.46	15.2	26.1	58.7
	Item 32	0.38	9.4	28.8	61.8
	Item 33	0.47	16.2	17.4	66.4
Quantitative 2	Item 10	0.29	16.2	29.9	53.9
	Item 30	0.35	9.2	12.8	78.0
	Item 36	0.39	2.7	5.4	91.9
	Item 39	0.15	1.5	12.9	85.6

Note: *P* is the proportion of correct responses from the overall sample.

To check the non-speeded test administration assumption, percentages of omitted responses on the last five items in the Verbal, Quantitative 1, and Quantitative 2 subtests were examined. The results are presented in Table 8. The percentages of omitted responses on the last five items were significantly higher in the Verbal subtest than the other two subtests, regardless of difficulty levels of the items. The last five items in the Quantitative 1 subtest had low omit rates while the difficult items (item 36 and item 39) in the Quantitative 2 subtest indicated substantially higher omit rates than the other items. Those two items had high item-total correlations, suggesting that despite their high omit rates, the items discriminated high-ability and low-ability examinees very well.

Table 8. Descriptive Statistics for the Last Five Items on the EEGS Subtests

Subtest	Item	p-value	Item-total Correlation	Proportion Missing
Verbal	Item 76	0.33	0.48	0.51
	Item 77	0.41	0.50	0.53
	Item 78	0.41	0.52	0.54
	Item 79	0.43	0.51	0.54
	Item 80	0.31	0.45	0.56
Quantitative 1	Item 36	0.84	0.45	0.13
	Item 37	0.65	0.43	0.19
	Item 38	0.77	0.45	0.11
	Item 39	0.76	0.55	0.16
	Item 40	0.74	0.52	0.19
Quantitative 2	Item 36	0.39	0.61	0.53
	Item 37	0.85	0.44	0.12
	Item 38	0.80	0.49	0.18
	Item 39	0.15	0.36	0.52
	Item 40	0.66	0.61	0.29

Note: p-value is the proportion of correct responses.

One theoretical feature that makes IRT models superior over other psychometric frameworks is the invariance (i.e., equality) of item and examinee parameters from different examinee populations or measurement conditions (Rupp & Zumbo, 2006). Parameter invariance in IRT can be investigated when there are at least two examinee populations or two measurement conditions for parameter comparisons. In this study, measurement invariance of item parameters was investigated across male and female examinees using a multi-group CFA framework.

Table 9 shows the results of measurement invariance tests for the three subtests of EEGS. For all of the subtests, weak invariance was met, which suggests that the constructs indicated the same meaning across male and female examinees. In the context of IRT, fixing factor loadings across male and female examinees for testing weak invariance is analogous to fixing item discrimination parameters across male and female examinees. Weak invariance of the items in the EEGS shows that the discriminatory power of the items did not differ between male and female examinees. Strong invariance was ensured for the Quantitative 1 and Quantitative 2 subtests but not for the Verbal subtest. As explained earlier, strong invariance assumes that item intercepts are equal across groups. In the context of IRT, item intercepts are analogous to item difficulty parameters. If one group has higher or lower probability to respond to item correctly than the other group, then this affects the means of the observed item, hence affects the mean of the scale and the latent variable. In this study, significant χ^2 change in the Verbal subtest indicated non-invariance of intercepts (i.e., item difficulties) between male and female examinees. Therefore, it can be concluded that some items in the Verbal subtest were systematically easier or more difficult for one of the gender groups. Finally, strict invariance was met for none of the EEGS subtests. Strict invariance is particularly important for group comparisons based on the sum of observed item scores, because observed variance is considered as a combination of true score variance and residual variance. The violation of this invariance test suggests that the items in EEGS may not be equally reliable across male and female examinees. According to Meredith (1993),

strict invariance is necessary for a fair and equitable comparison across groups. Because none of the EEGS subtests indicated strict invariance in this study, it can be concluded that test scores from the three subtests of EEGS cannot be reliably and meaningfully compared between male and female examinees.

Table 9. Results of Measurement Invariance Tests for the EEGS Subtests

Subtest	Type of Invariance Test	$\Delta\chi^2$	Δdf	CFI	RMSEA
Quantitative 1	Configural invariance	-	-	0.75	0.05
	Weak invariance	48.56	39	0.72	0.05
	Strong invariance	52.14	39	0.72	0.05
	Strict invariance	664.14*	40	0.65	0.05
Quantitative 2	Configural invariance	-	-	0.75	0.06
	Weak invariance	51.23	39	0.73	0.06
	Strong invariance	56.25	39	0.73	0.06
	Strict invariance	283.28*	40	0.71	0.07
Verbal	Configural invariance	-	-	0.68	0.04
	Weak invariance	99.56	79	0.67	0.04
	Strong invariance	204.60*	79	0.66	0.04
	Strict invariance	687.43*	80	0.54	0.05

Note: $\Delta\chi^2$ = Difference in chi-square between the two consecutive models; Δdf = Difference in degrees of freedom between the two consecutive models; * p-value < .05

Results of Model-Fit Comparison

The advantages of IRT models can be achieved only if there is a satisfactory goodness-of-fit between the model and test data (Gao, 2011). In this study, the overall fit of 1PL, 2PL, and 3PL models was compared based on the Likelihood Ratio test. Table 10 presents the results of model-fit comparisons. For all of the EEGS subtets, the 3PL indicated the best model fit, the 2PL model was the second best-fitting model, and the 1PL model indicated the worst model fit. Especially in the Verbal subtest, the difference between -2 log likelihood values of the 1PL and 2PL models was very large. Because the restricted 1PL model cannot account for the variation among the Verbal test items as much as the other two models, the resulting model fit was very poor.

Table 10. Comparison of the Three IRT Models for the EEGS Subtests

Subtest	Comparisons	
	1PL vs. 2PL	2PL vs. 3PL
Quantitative 1	1659.063 (39)*	450.756 (40)*
Quantitative 2	4608.907 (39)*	1039.093 (40)*
Verbal	18835.897(79)*	2161.296 (80)*

Note: Each cell shows the difference in -2 log likelihood values and difference in the number of estimated item parameters. * p-value < .001

Although statistical tests of goodness-of-fit are widely used in the evaluation of model-data fit, they often provide inconclusive evidence for adequate model-data fit because of their sensitivity to sample size and their insensitivity to certain forms of model-data misfit (Chernyshenko et al., 2001; Van der Wollenberg, 1982). Therefore, it is important to use graphical fit plots of ICCs and item information functions (IIFs), in addition to chi-square goodness of fit tests for single items. IIF of an item can be expressed as:

$$I_i(\theta) = \frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)}, \quad (6)$$

where $P_i(\theta)$ is the probability of correctly responding to item i given θ, $Q_i(\theta)$ is equal to (1-$P_i(\theta)$), and $P_i'(\theta)$ is the first derivative of $P_i(\theta)$ given θ. In this study, IFFs from the 1PL, 2PL, and 3PL models were computed for each item in the Verbal, Quantative 1, and Quantitative 2 subtests. Misfit

items were identified based on chi-square goodness of fit statistics for the items and the evaluation of ICCs (see Table 11). The results suggested that the 3PL model provided the best model-fit in the Verbal and Quantitative 2 subtests. However, in the Quantitative 1 subtest, the 2PL model indicated better model-fit based on the fewer number of misfit items. A potential reason for the worse model-fit of the 3PL model in the Quantitative 1 subtest might be omitted responses and aberrant response patterns of examinees. In the context of IRT, some examinees may have unexpected guessing behaviors which may result in high guessing parameters ($c > 0.5$) for the items. To further investigate the extent to which examinees' response patterns are consistent with expectation, person fit statistics – such as the log-likelihood person-fit statistic (Levine & Rubin, 1979) and the standardized log-likelihood statistic (Drasgow, Levine, & Williams, 1985) – can be used.

Table 11. Number of Misfit Items in the Three Subtests of the EEGS

Subtest	N	IRT Models		
		1PL	2PL	3PL
Quantitative 1	40	20 (50%)	6 (15%)	10 (25%)
Quantitative 2	40	19 (45.5%)	11 (27.5%)	7 (17.5%)
Verbal	80	34 (42.5%)	12 (15%)	11 (13.8%)

In addition to the evaluation of model-fit based on item-fit statistics and model-fit plots, marginal test reliability, test information functions, and conditional standard error of measurement (CSEM) can be useful in selecting the best-fitting IRT model. Test information function (TIF) is basically the sum of all IIFs in a given test, which provides a visual depiction of where along the trait continuum a test is most discriminating (Reise & Waller, 2002).

Figure 3 shows TIF and CSEM plots from the 1PL, 2PL, and 3PL models across the three subtests of the EEGS. The results suggested that except for the Quantitative 1 subtest, the 3PL model provided more information and less measurement error than the other two models along the ability continuum. The 2PL model was evidently better than the 1PL and 3PL models in the Quantitative 2 subtest, which also supports the findings from item misfit analyses. The performance of the 1PL model was similar to the performances of the 2PL and 3PL models only on the tails of the distributions where very low-ability and high-ability examinees were located. The marginal reliability was above .90 for all of the IRT models across the three subtests of EEGS.

DISCUSSION AND CONCLUSION

The findings of this study suggest that despite the appealing practical features of IRT, fitting IRT models to large-scale assessments requires a comprehensive investigation of model assumptions, item fit, and overall model-fit. Compared to CTT, IRT requires much stronger model assumptions, such as unidimensionality and local independence of the items. Without adequate evidence supporting the integrity of those assumptions, IRT results from an operational assessment may not be credible.

As Chernyshenko et al. (2001) pointed out; there is a strong trade-off between searching for the most appropriate IRT model that adequately describes item responses and rejecting items that do not fit a chosen model. The findings of this study suggest that more complex models (e.g., the 3PL model) tend to fit the data from large-scale assessments better than the simple models (e.g., Rasch model, 1PL model). However, it should be noted that the selection of more complex models will increase the minimum sample size required for IRT analyses, as well as limit prospective applications of IRT in practical settings (Chernyshenko et al., 2001). Although sample size may not be a concern in large-scale assessments, other possible issues in large-scale assessments, such as high omit rates and random-guessing, still remain as potential threats to the estimation of complex IRT models.

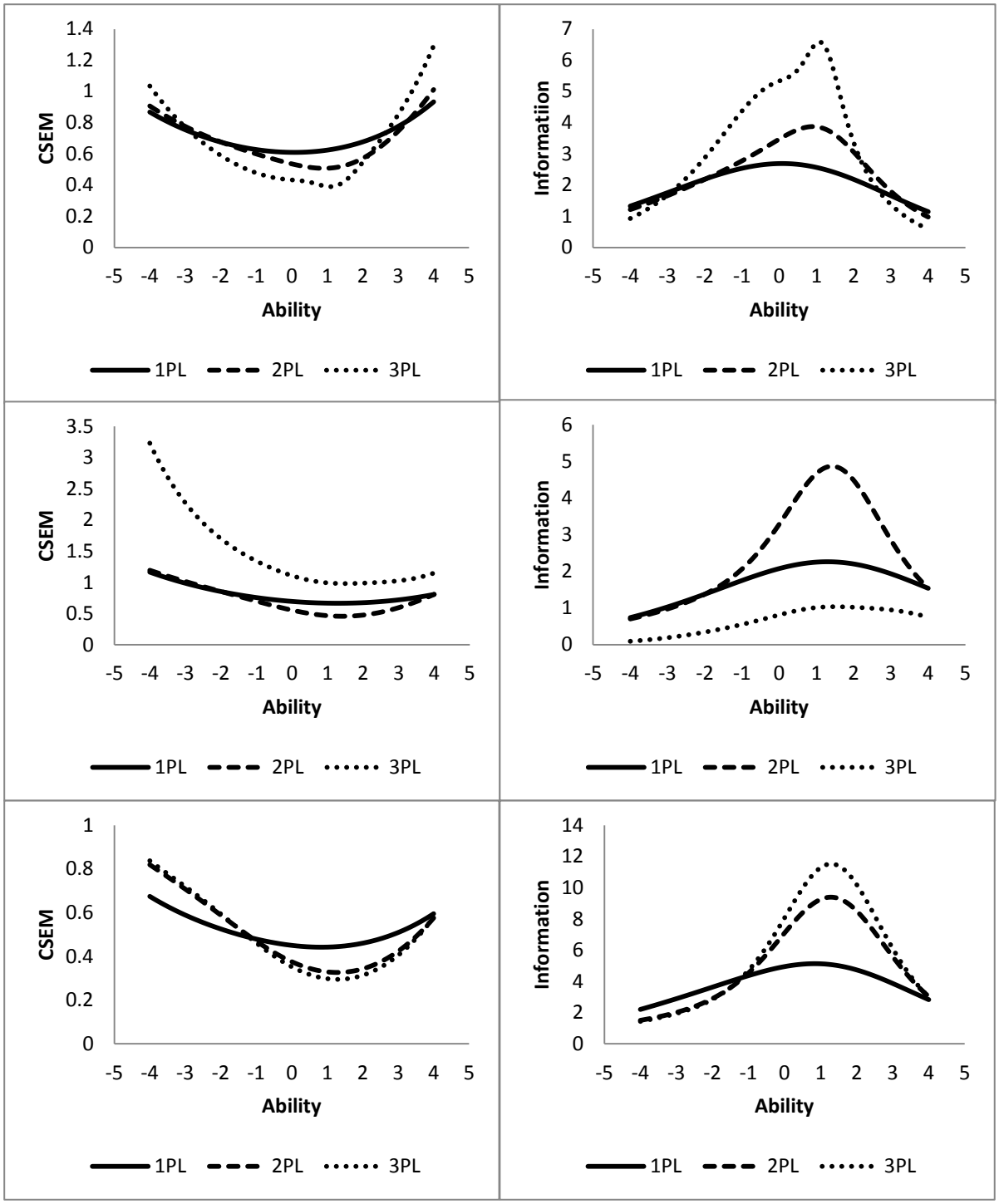


Figure 3. Conditional Standard Error of Measurement and Test Information Functions of the IRT Models for the Quantitative 1 (top), and Quantitative 2 (middle), and Verbal (bottom) Subtests

In this study, the estimation of guessing parameter was particularly problematic in the Quantitative 1 subtest because of higher difficulty levels of the items and higher omitted response rates. The assumption for the guessing parameter that every examinee has the same probability to guess an item correctly may not reflect the real guessing situation (De Ayala, 2008). Therefore, it is difficult to find

the reasons of the guessing problem in the EEGS without further investigation. Also, despite the fact that the 2PL model does not account for guessing in the items, it provided better model-fit than the 3PL model in the Quantitative 1 subtest. It is because the 3PL model can produce the most accurate item parameter and ability estimates when a moderate amount of guessing is assumed (Pelton, 2002). Guessing degrades the fit of the IRT models, because the empirical ICC cannot be expected to approach zero when the theta value is small (Progar, Sočan, & Peč, 2008).

The invariance property of IRT item and ability parameters has important implications for the application of IRT to large-scale assessments. First, assuming that item parameters and ability parameters are invariant regardless of who takes the test and which items are used, computerized adaptive testing (CAT), where each examinee responds to a different set of items from the precalibrated item bank, can be implemented. For instance, Bulut and Kan (2012) demonstrated the applicability of CAT in the EEGS. Their results suggested that CAT provided highly accurate ability estimates using much fewer test items than the paper-pencil form of the EEGS. Second, the invariance property of IRT models facilitates creating comparable scores on different forms of an assessment. A linear transformation of ability estimates can equate the test scores from groups of examinees with different abilities, such as students in different grades, or, from groups of examinees with similar abilities who take the test at different times. This feature would allow producing valid and compareable scores from the EEGS across multiple test administrations. Because test scores can be placed on a common scale through equating and linking procedures, the scores can be directly compared between the examinees who might take the test at different administrations of the EEGS.

In light of the findings of the present study, future studies should focus on the impact of omitted item responses on the validity and reliability of IRT-based test scores obtained from large-scale assessments. Furthermore, more comprehensive studies are needed for understanding the invariance of the item parameters between male and female examinees. Testing differential item functioning of the EEGS items can be helpful for understanding why male and female examinees differ on the Verbal, Quantitative 1, and Quantitative 2 subtests.

REFERENCES

Albenese, M. A., & Forsyth, R. A. (1984). The one-, two-, and modified two parameter latent trait models: An empirical study of relative fit. *Educational and Psychological Measurement, 44*(2), 229–246.

Baker, F. B. (1985). *The basic of item response theory*. Portsmouth, NH: Heinemann.

Berberoglu, G. (1990). Do the Rasch and three-parameter models produce similar results in test analyses? *Journal of Human Sciences, 10*, 7–16.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis*. Cambridge, MA: MIT Press.

Bulut, O., & Kan, A. (2012). Application of computerized adaptive testing to Entrance Examination for Graduate Studies in Turkey. *Eurasian Journal of Educational Research*, 49, 61–80.

Celik, D. (2001). *The fit of the one-, two- and three-parameter models of item response theory (IRT) to the ministry of national education secondary education institutions student selection and placement test data.* Unpublished master's thesis, Middle East Technical University, Ankara, Turkey.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265–289.

Chernyshenko, O. S., Stark, S., Chan, K. Y.,, Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: issues and insights. *Multivariate Behavioral Research, 36*(4), 523–562.

Choi, I. (1989). *An application of item response theory to language testing: Model-data fit studies* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Courville, T. G. (2005). *An empirical comparison of item response theory and classical test theory item/person statistics* (Unpublished doctoral dissertation). Texas A&M University.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston, Inc.

De Ayala, R. J. (2008). *The theory and practice of item response theory.* New York, NY: Guilfords Publications.

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement, 22*(4), 249–262.

Drasgow, F., Levine M. V., Tsien, S., Williams B. A., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19*, 143–165.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67–86.

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

Engelhard, G. (1991). Thorndike, Thurstone and Rasch: A comparison of their approaches to item-invariant measurement. *Journal of Research and Development in Education, 24*(2), 45–60.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person parameters. *Educational and Psychological Measurement, 58*, 357–381.

Ferrara, S., Huynh, H., & Bagli, H. (1997). Contextual characteristics of locally dependent open-ended item clusters on a large-scale performance assessment. *Applied Measurement in Education, 12*, 123–144.

Fletcher, T. D. (2015). psychometric: applied psychometric theory. [Computer software]. Available from http://CRAN.R-project.org/package=psychometric.

Gao, S. (2011). The exploration of the relationship between guessing and latent ability in IRT models. *Dissertations.* Paper 423.

Güler, N., Uyanık, G. K., & Teker, G. T. (2014). Comparison of classical test theory and item response theory in terms of item parameters. *European Journal of Research on Education, 2*, 1–6.

Guyer, R., & Thompson, N.A., (2011). *User's Manual for Xcalibre 4.1.* St. Paul MN: Assessment Systems Corporation.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response. *Educational Measurement: Issues and Practice, 12*(3), 38–47.

Hambleton, R. K., & Rogers, J. H. (1990). Using item response models in educational assessments. In W. Schreiber, & K. Ingenkamp (Eds.), *International developments in large-scale assessment* (pp. 155-184). England: NFER-Nelson.

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

Hu, L.T., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.

Humphreys, L. G., & Montanelli, R. G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research, 10*, 193–205.

Kilic, I. (1999). *The fit of one, two and three parameter models of item response theory to the student selection test of the student selection and placement center.* Unpublished master's thesis, Middle East Technical University, Ankara, Turkey.

Levine, M. V. & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269–290.

Li, C. H. (2014). *The performance of MLR, USLMV, and WLSMV estimation in structural regression models with ordinal variables* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. 1538039469)

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

McDonald, R. P. (1981). The dimensionality of test and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100–117.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*, 525–542.

Muthén, L.K., & Muthén, B.O. (1998-2011). *Mplus 6*. Los Angeles, CA: Muthén and Muthén.

Önder, İ. (2007). Model veri uyumunun araştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 32*, 210–220.

Pelton, T. W. (2002). *The accuracy of unidimensional measurement models in the presence of deviations for the underlying assumptions.* Unpublished doctoral dissertation, Brigham Young University, Department of Instructional Psychology and Technology.

Progar, Š, Sočan, G., & Peč, M. (2008). An empirical comparison of item response theory and classical test theory. *Horizons of Psychology, 17*(3), 5–24.

R Core Team (2015). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Raykov, T., & Marcoulides, G. A. (2010). *Introduction to psychometric theory*. New York, NY: Taylor & Francis.

Reise, S., & Waller, N. (2002). Item response theory for dichotomous assessment data. In Drasgow, F. and Schmitt, N. (Eds.), *Measuring and Analyzing Behavior in Organizations*. San Francisco: Jossey-Bass.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1–36.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66*(1), 63–84.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589–617.

Teker, G. T., Kelecioglu, H., & Eroglu, M. G. (2013). An investigation of goodness of model data fit. *Procedia – Social and Behavioral Sciences, 106*, 394–400.

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements*. New York: Teacher's College.

Van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47*, 123–140.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16*(4), 33–45.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125–145.

UZUN ÖZET

Giriş

Eğitim ve psikoloji alanındaki test uygulamaları kişilerin bilgi, beceri ve tutum gibi örtük değerlerini en güvenilir biçimde ölçmeyi amaçlamaktadır. Bu tarz uygulamalarda standart testler en çok tercih edilen ölçme aletleri olmuştur. Standart testler, üniversitelere giriş sınavlarından ehliyet sınavlarına kadar birçok önemli alanda kullanılmaktadır. Bu tarz testlerde madde analizi ve kişilerin puanlarının hesaplanmasında iki farklı yaklaşım kullanılmaktadır. Klasik test teorisi olarak bilinen ve Türkiye'deki sınavlarda da yaygın olarak kullanılmakta olan yaklaşım kişilerin başarı, tutum, ya da diğer yeteneklerini temsil eden ham puanların hesaplanmasına dayalıdır. Bu yöntem her ne kadar kolay bir şekilde uygulanabilmesi nedeniyle tercih edilse de elde edilen madde istatistikleri ve ham puanların sınavda kullanılan sorulara ve soruları cevaplayan kişilerin oluşturduğu örnekleme bağlı olmasından ötürü geri plana düşmüş ve yerine bu sorunları içermeyen madde tepki kuramı ortaya çıkmıştır. Madde tepki kuramına göre bilgi ve beceri gibi örtük özelliklerin sınavda kullanılan sorular ve bu soruları cevaplayan kişilerin seviyelerinden bağımsız olarak ölçülmesi amaçlanmaktadır.

Madde tepki kuramının geçmiş yıllarda yurtdışında Test of English as a Foreign Language (TOEFL) ve The Certificate of English, Türkiye'de ise Öğrenci Seçme Sınavı (ÖSS) ve ÖZDEBİR gibi geniş ölçekli testlere uygulanabilirliği incelenmiştir. Bu çalışma, madde tepki kuramının geniş ölçekli standart testlere uygulanmasına yönelik farklı bir empirik örnek sunmayı ve bu kapsamda madde tepki kuramının uygulanmasında ortaya çıkabilecek sorunları değerlendirmeyi amaçlamaktadır. Bu amaçla Akademik Personel ve Lisansüstü Eğitimi Giriş (ALES) Sınavı tek boyutlu madde tepki kuramı modelleri doğrultusunda incelenmiştir. ALES sınavı geniş ölçekli bir test olup Türk üniversitelerine yapılan yüksek lisans ve doktora başvuruları ve üniversitelerdeki akademik personelin belirlenmesi gibi birçok önemli alanda kullanılmaktadır. Madde tepki kuramının ALES sınavına uyarlanabilirliği benzer şekildeki klasik test teoremine dayalı diğer geniş ölçekli standart testlere yönelik de fikir sağlayacaktır.

Yöntem

Bu çalışmanın örneklemi olarak ALES sınavının 2010 yılı güz döneminden elde edilen veriler kullanılmıştır. Sınava giren 142.178 kişi arasından rastgele seçim yoluyla seçilen 5000 kişinin sınavda 160 soruya verdikleri cevaplar tek boyutlu madde tepki kuramı modellerine göre incelenmiştir. ALES sınavı Sayısal 1, Sayısal 2 ve Sözel olmak üzere üç alt testten oluşmaktadır. Sayısal 1 ve Sayısal 2 testleri 40, Sözel testi ise 80 soru içermektedir. Sınavdaki tüm sorular çoktan seçmeli olup cevaplar doğru ya da yanlış olarak değerlendirilmiştir. Sınavda kişilerin her soru için sadece bir seçenek işaretlemeleri gerekmektedir. Ayrıca kişilerin cevaplarını bilemedikleri soruları boş geçebilmelerine izin verilmiştir. Toplam sınav süresi 180 dakikadır.

ALES sınavı verilerinin madde tepki kuramı doğrultusunda incelenmesi üç aşamadan oluşmaktadır. Birinci aşamada sorulara verilen cevaplar ve sınavdaki ham puanlar tanımlayıcı istatistikler aracılığıyla incelenmiştir. Bu analizlerin amacı ALES'te kullanılmış soruların zorluk ve madde ayırıcılık indekslerini incelemek ve uygun olmayan soruları tespit etmektir. İkinci aşamada ise madde tepki kuramına dair model varsayımları değerlendirilmiştir. Bu temel varsayımlar sınavın tek boyutlu olması ve soruların yerel bağımsızlığıdır. Sayısal 1, Sayısal 2, ve Sözel testlerinin tek boyutluluğu her bir teste tek boyutlu doğrulayıcı faktör analizi uygulanarak incelenmiştir. Bu modelin her bir test için uygun model uyum indeksleri vermesi testlerin tek boyutluluğunu göstermektedir. Soruların yerel bağımsızlığı özelliği ise sınavda alt ve üst %20'lik dilimde bulunan kişilerin sorulara verdikleri cevaplar arasındaki korelasyona bakılarak incelenmiştir. Eğer yerel bağımsızlık varsayımı doğru ise bu iki gruptaki kişilerin cevapları arasında yüksek korelasyon bulunmaması gerekmektedir. Bu iki varsayım haricinde sınav süresinin yeterliliği, sorulara verilen cevaplardaki tahmin oranı ve bir parametreli model için madde ayırıcılık indekslerinin eşit oluşu gibi ikincil varsayımlar da incelenmiştir. Üçüncü aşamada ise bir parametreli, iki parametreli ve üç parametreli lojistik madde tepki kuramı modelleri Sayısal 1, Sayısal 2 ve Sözel testlere sırasıyla uygulanmış ve en uygun model tespit edilmeye çalışılmıştır. Madde ve testlerin uyumu belirtilen modellere uyumu hem istatistiksel hem de grafiksel yöntemlerle incelenmiştir.

Sonuç ve Tartışma

Çalışmanın sonuçlarına göre Sayısal 2 testinin diğer iki teste göre daha zor olduğu ve bu testin yüksek başarılı ve düşük başarılı öğrencileri daha iyi ayırt ettiği görülmüştür. Ayrıca genel olarak sınava giren erkek katılımcıların Sayısal 1 ve 2 testlerinde daha başarılı olduğu, bayan katılımcıların ise Sözel testinde daha başarılı oldukları görülmüştür. Sınavda kullanılan tüm soruların yeterince düzeyde madde ayırıcılık gücüne sahip oldukları belirlenmiştir.

Model varsayımlarının incelenmesinde tek boyutlu doğrulayıcı faktör modelinin Sayısal 1, Sayısal 2 ve Sözel testleri için yüksek model uyum indeksleri verdiği belirlenmiş ve bu sonuçlar doğrultusunda ALES alt testlerin tek boyutlu olduğu sonucuna varılmıştır. Soruların yerel bağımsızlığının incelenmesinde ise Sayısal 1 ve Sayısal 2 alt testlerinde yer alan sorular arasında yüksek korelasyonlara rastlanmamış ve yerel bağımsızlık varsayımının geçerli olduğu görülmüştür. Sözel testinde ise özelikle paragraf tipi sorularda aynı paragrafa dair cevaplanması gereken soruların bazıları arasında yüksek korelasyonlar görülmüştür. Sınav süresinin yeterliliği Sayısal 1, Sayısal 2 ve Sözel testlerinin son beş sorusundaki boş cevap oranlarına bakılarak incelenmiştir. Bu incelemeye gore Sözel testinin son sorularının soruların zorluk oranlarından bağımsız olarak yüksek oranda boş bırakıldığı, Sayısal 1 ve Sayısal 2 testlerinde ise boş bırakılan soruların genelde sınavdaki zor sorular olduğu tespit edilmiştir. Diğer iki varsayım incelendiğinde ise ALES sorularının eşit madde ayırıcılık indekslerine sahip olmadığı ve sınava katılanların soruların doğru cevaplarını tahmin etme konusunda daha çok soruları boş bırakma davranışına gittikleri görülmüştür.

Bir, iki, ve üç parametreli modellerin ALES alt testlerine uyumu incelendiğinde Sözel ve Sayısal 2 testlerine en uygun modelin üç parametreli model olduğu, fakat Sayısal 1 testinde iki parametreli modelin daha iyi uyum gösterdiği belirlenmiştir. Model parametreleri incelendiğinde ise bir ve iki parametreli modeldeki parametrelerin ve bu modellerden elde edilen sınav puanlarının daha stabil olduğu tespit edilmiştir. Üç parametreli modelde özellikle tahmin (guessing) parametresinin hesaplanması esnasında boş soru sayısının da yüksek olması nedeniyle beklenmedik değerler tespit edilmiştir.

Bu çalışmanın sonuçlarına göre madde tepki kuramının üstün istatistiki özelliklerine karşın sınavlara uygulanması aşamasında karşılaşılabilecek olası sorunlara dikkat çekilmiştir. ALES ve benzeri özellikteki geniş ölçekli testlere madde tepki kuramının uygulanabilmesi için sınavların bu doğrultuda önceden dikkatle tasarlanması gerektiği görülmüştür.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2015; 6(2); 331-350

Farklı Boyutluluk Özelliklerindeki Basit ve Karmaşık Yapılı Testlerin Çok Boyutlu Madde Tepki Kuramına Dayalı Parametre Kestirimlerinin İncelenmesi*

Examining the Parameter Estimations of Simple and Complex Structured Tests with Various Dimensionality Properties Based on Multidimensional Item Response Theory

Derya Çakıcı Eser** Selahattin Gelbal***

Öz
Bu çalışmada basit ve karmaşık yapıdaki iki ve üç boyutlu testlerin farklı test uzunluğu ve örneklem büyüklüğü koşullarında parametre kestirimleri yapılmış, kestirimlere ilişkin olarak elde edilen RMSE, yanlılık ve kestirilen parametreler ile gerçek parametreler arasındaki ilişki incelenmiştir. Çalışmanın verileri iki boyutlu basit, iki boyutlu karmaşık, üç boyutlu karmaşık ve üç boyutlu basit yapıya uygun olacak şekilde simülasyon yoluyla üretilmiştir. Veri setlerinin test uzunluğu 12 ve 48 madde; örneklem büyüklüğü 1000, 2000 ve 4000 olacak şekilde değiştirilmiştir. Bu şekilde ele alınan dört test yapısı ile 2 test uzunluğu ve 3 örneklem büyüklüğü koşulu çaprazlanarak (4x2x3) 24 koşul içeren desen oluşturulmuştur. Oluşturulan her koşula ilişkin olarak, madde ve birey parametreleri sabit tutulmak üzere, 25 tekrar yapılarak toplamda 600 veri seti oluşturulmuş ve analiz edilmiştir. Elde edilen sonuçlara göre madde parametrelerinin parametre kestirim iyiliği hem madde hem birey sayısındaki artışla artmaktadır. Ele alınan her bir koşulda d parametreleri aynı koşuldaki a parametrelerinden daha kararlı kestirilmiştir. Birey parametrelerinin kestirim iyiliği test uzunluğundaki artış ile artmaktadır. Yüksek korelasyon değerleri RMSE'nin düşük olduğu koşullarda elde edilmiştir. Ayrıca çalışmanın sonunda daha kararlı kestirim yapabilmek için gerekli test uzunluğu ve örneklem büyüklüğüne ilişkin öneriler geliştirilmeye çalışılmıştır.

Anahtar Kelimeler: Çok boyutlu madde tepki kuramı, basit yapı, karmaşık yapı, parametre kestirimi

Abstract
Under the scope of this study; parameter estimations of simple and complex structured two and three dimensional tests have been performed according to different test length and sample size conditions; according to the estimations, RMSE, bias and relation between estimated and actual parameters have been investigated. Research data has been generated by means of simulations according to requirements of two dimensional simple structure, two dimensional complex structure, three dimensional complex structure and three dimensional complex structure. The simulation conditions were test length (12 and 48 items) and sample size (1000, 2000 and 4,000). By intercrossing the simulation conditions (4 test structures x 2 test lengths x 3 sample sizes); 24 crossed conditions have been acquired. By fixing item and ability parameters and applying 25 replications; 600 data sets have been generated and analyzed. According to the results, parameter recovery of item parameters increases with the number of items and examinees. For each condition; d parameters have been estimated more accurately than a parameters on the same pattern. Parameter recovery of ability parameters is developed with the increase in test length. High correlation values acquired in the conditions where RMSE values are low. Additionally; suggestions on the required test length and sample size for accurate estimations are provided at the final stage of the study.

* Bu çalışma "Çok Boyutlu Madde Tepki Kuramının Farklı Modellerinden Çeşitli Koşullar Altında Kestirilen Parametrelerin İncelenmesi" başlıklı doktora tezinden oluşturulmuştur.

** Dr., Kırıkkale Üniversitesi, Eğitim Fakültesi, Kırıkkale-Türkiye, deryacakicieser@gmail.com
*** Prof. Dr. Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, sgelbal@gmail.com

Geliş Tarihi:13.10.2015
Kabul Tarihi: 09.11.2015

Keywords: Multidimensional item response theory , simple structure, complex structure, parameter estimation

GİRİŞ

Madde tepki kuramı (MTK) ilk olarak Lord ve Novick'in (1968) Statistical Theories Of Mental Test Scores kitabında Allan Birnbaum tarafından ele alınmıştır. Bu kuramda "yetenek" kavramı ile ilgilenilir ve yetenek θ (teta) simgesi ile gösterilir. Kurama göre bireyin bir testte gösterdiği performansın altında o test ile ölçülmeye çalışılan yeteneği yatar. Kuramın amacı bireyin testle ölçülen yetenek düzeyini tahmin etmek veya kestirmek için bir temel oluşturmaktır. Buna göre bir bireyin yetenek düzeyi, bireyin maddelere verdiği cevaplardan kestirilmeye çalışılır. Kuram ile bireyin bir testte ortaya koyduğu performans ile bu performansın altında yatan gözlenemeyen yeteneği veya özelliği arasında bir ilişki ortaya konur. Gözlenen ve gözlenemeyen bu özellikler arasındaki ilişki matematiksel fonksiyonlar ile açıklanır. Bu sebeple madde tepki modelleri matematiksel modellerdir. MTK modelleri oldukça güçlü varsayımlara dayanır (Hambleton ve Swaminathan, 1985; Embretson ve Reise, 2000). Bu varsayımlar tek boyutluluk ve yerel bağımsızlık başlıkları altında ele alınır. Varsayımlar aşağıda açıklanmıştır:

Tek boyutluluk: Tek boyutluluk test ile tek bir örtük özelliğin ölçülüyor olması şeklinde tanımlanır. Pratikte bunu karşılamak güçtür. Bu sebeple testin baskın bir faktörü ölçtüğünün gösterilmesi tek boyutluluğun sağlanması için yeterlidir (Hambleton ve Swaminathan, 1985; Embretson ve Reise, 2000).

Yerel Bağımsızlık: Yerel bağımsızlık aynı yetenek düzeyindeki bireylerin farklı maddelere verdiği tepkilerin istatistiksel olarak bağımsız olması olarak tanımlanır. Bu varsayım pratikte bireyin bir maddeye verdiği cevabın diğer bir maddeyi olumlu ya da olumsuz etkilememesi anlamına gelir (Hambleton ve Swaminathan, 1985; Embretson ve Reise, 2000).

Yukarıda yer alan varsayımlara ek olarak tüm MTK modellerinin sağlaması gereken koşul kullanılan testin hız testi olmamasıdır. Bir testin hız testi olması testin ilgilenilen yeteneğin yanında performans hızını da ölçtüğü anlamına gelir. Bu durumda testin tek boyutluluk varsayımı ihlal edilmiş olur (Hambleton ve Swaminathan; 1985).

Ancak MTK varsayımlarının karşılanmış olması doğrudan analize geçilebileceği anlamına gelmez. Bunun için öncelikle hangi modele göre analiz yapılacağına karar verilmesi gereklidir. MTK'da modeller cevap kategorisi sayısına göre belirlenir. MTK modelleri iki kategorili modeller ve çok kategorili modeller başlıkları altında toplanır. Bu çalışmanın konusu iki kategorili modeller olduğu için çok kategorili modellere değinilmemiştir.

İki kategorili MTK modelleri lojistik ve ogive modeller şeklinde ele alınır. Modelde yer alan parametrelere bağlı olarak; 1, 2 ve 3 parametreli modeller şeklinde adlandırılır. 3 parametreli lojistik model (3PLM)'de bireyin doğru cevap verme olasılığını etkileyen parametreler kişinin yetenek düzeyi ile beraber maddenin; güçlüğü, ayırıcılığı ve tahmin (şans) parametresidir. Bu modele göre bir maddenin doğru cevaplanma olasılığı tüm yetenek düzeyleri için her zaman sıfırın üzerindedir. 3 parametreli lojistik modeldeki bir maddeye ilişkin olasılık fonksiyonu eşitlik 1'de verilmiştir.

$$P(X_{is} = 1 \mid \theta_s, \beta_i, \alpha_i, \gamma_i) = \gamma_i + (1 - \gamma_i)\frac{\exp[\alpha_i(\theta_s - \beta_i)]}{1+\exp[\alpha_i(\theta_s - \beta_i)]} \qquad \text{(Eşitlik 1)}$$

X_{is} = s bireyin i maddesine verdiği cevap (1 veya 0)

θ_s = s bireyinin yetenek düzeyi

β_i = i maddesinin güçlüğü

α_i = i maddesinin ayırıcılığı

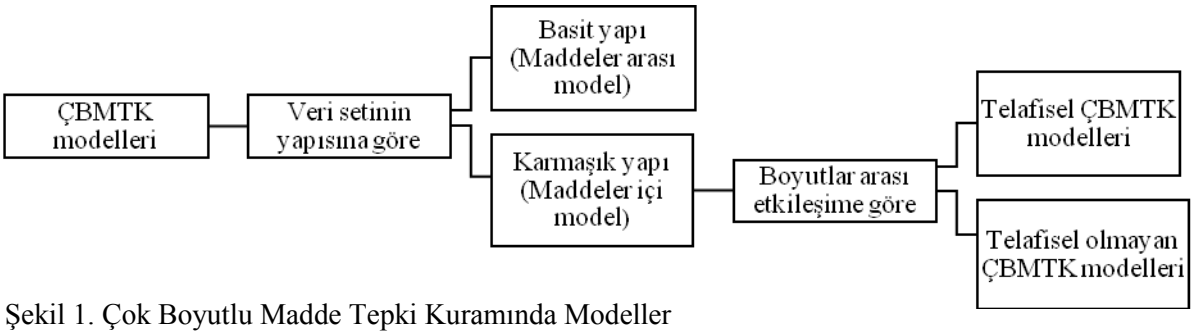
γ_i = i maddesinin en düşük asimptotu (şans parametresi)

Maddelerin şans parametresinin sıfır veya ihmal edilebilir olduğu, ancak maddelerin ayırt edicilik ve güçlük bakımından farklılaştığı durumda 2PLM kullanılır. 2PLM'nin olasılık fonksiyonu eşitlik 1'deki fonksiyondan şans parametresinin çıkartılmış (γi) halidir. Maddelerin sadece güçlük bakımından farklılaştığı, ayırt ediciliklerinin eşit olduğu durumda ise 1PLM kullanılır. 1PLM'nin özel bir durumu olarak Rasch modeli tanımlanır. Rasch modelinde madde güçlükleri farklılaşır, ancak madde ayırıcılıkları 1'e eşittir. Rasch modelinin lojistik fonksiyonunda madde parametrelerinden sadece güçlük parametresi yer alır.

MTK'da farklı özellikteki durumlara ilişkin birbirinden farklı modeller yer alsa da, varsayımlar karşılanmadığı takdirde bu modellerin kullanılması uygun olmayacaktır. Uygulamalarda en sık karşılaşılan durum, kullanılan testlerin tek boyutlu olmamasıdır. Bu durumda MTK'nın temel varsayımı olan tek boyutluluk ihlal edilmektedir. Bu ve buna benzer durumlarda tek boyutlu MTK modelleri yerine tek boyutlu MTK'nın genişletilmiş hali olan ve çok boyutlu veri setleri için geliştirilmiş olan çok boyutlu madde tepki kuramının (ÇBMTK) kullanılması daha uygundur.

Tek boyutlu madde tepki kuramı yapı veya boyut olarak adlandırılan gözlenemeyen bir değişken ile bireyin belli bir test maddesine doğru cevap verme olasılığı arasındaki ilişkiyi modellemek için kullanılır. Buna karşılık ÇBMTK iki ya da daha fazla gözlenemeyen değişken ile bireyin belli bir test maddesine doğru cevap verme olasılığı arasındaki ilişkiyi modellemek için kullanılır (Ackerman, Gierl ve Walker, 2003). Tek boyutlu MTK modellerinde doğru cevaplama olasılığı tek bir yeteneğe dayanırken, ÇBMTK'da doğru cevaplama olasılığı ölçülmeye çalışılan çok sayıda yeteneğin bir fonksiyonudur. Buna bağlı olarak da ÇBMTK'da ölçme modeli test sonucu olarak tek bir puan yerine bir yetenek profili ortaya koyar (Hartig ve Höhler, 2009). Bir veri setine ÇBMTK'yı uygulamak için veri setinin çok boyutlu olmasının yanında başka varsayımlara da ihtiyaç duyulur. Bu varsayımlardan biri yerel bağımsızlık diğeri monoton artmadır. Yerel bağımsızlık varsayımı tek boyutlu MTK'daki ile aynı anlamı taşıdığından burada tekrar anlatılmamıştır. Monoton artma varsayımı ise bireyin doğru cevap verme olasılığının, θ vektöründeki herhangi bir elemanın daha yüksek değer alması ile artmasıdır (Reckase, 2009). Buna göre bireyin doğru cevap verme olasılığı ile θ'sı arasında monoton artan bir fonksiyon ortaya konur. Monoton artma varsayımı tek boyutlu MTK'da karşılanabilmesine rağmen bu varsayıma ihtiyaç duymayan tek boyutlu MTK modelleri de mevcuttur. Bu sebeple tek boyutlu MTK'da bir varsayım olarak ele alınmaz (Roberts ve ark 2000; akt: Reckase, 2009).

ÇBMTK'da kullanılan farklı modeller, yetenek ile doğru cevaplama olasılığı arasında farklı istatistiksel ilişkileri varsayar. Buna ek olarak boyutlar ve maddeler arasındaki ilişki veri setinin yapısına göre tanımlanabilir. ÇBMTK'da modeller veri setinin karmaşıklığına göre ve boyutlar arası etkileşime göre sınıflandırılmıştır (Hartig ve Höhler, 2009). Bu sınıflandırma şekil 1'deki gibi verilebilir.



Şekil 1. Çok Boyutlu Madde Tepki Kuramında Modeller

ÇBMTK'da bir test birden fazla yeteneği içerir. Ancak bu durum iki farklı yapı ile elde edilebilir. Birinci yapı basit yapı adını alır. Basit yapılı bir test birden fazla alt testten oluşur. Her bir alt test tek

bir yeteneği veya özelliği ölçer ve kendi içinde homojendir. Her bir madde yer aldığı alt test ile ölçülmeye çalışılan tek bir özellik ile ilişkilidir. Basit yapılı testlerin ortaya koyduğu modele maddeler arası model adı verilir. Bu model genelde geniş ölçekli testlerde karşımıza çıkar. Örneğin TIMSS (Trends in International Mathematis and Science Study) matematik ve fen bilimleri alt testlerinden oluşmaktadır. Matematik testinde yer alan her bir alt test çeşitli konu ve bilişsel alanları içermektedir. Her bir konu alanı ham puanların rapor edildiği tek boyutlu bir alt test olarak ele alınır. Bunun yanında testin tamamının matematik yapısını ölçtüğü varsayılır. Bu türdeki çok boyutlu testlerin ortaya koyduğu modele çoklu-tek boyutlu model (multi-unidimensional model) adı verilmektedir (Sheng ve Wikle, 2007; Bulut, 2013).

Çoklu-tek boyutlu model ÇBMTK'nın özel bir durumudur. Modelde madde parametrelerinden güçlük ve şans parametreleri her madde için bir tane kestirilirken; ayırt edicilik parametresi, her madde için testte yer alan boyut sayısı kadar elemana sahip bir vektör şeklinde kestirilir. Buna göre m alt testten oluşan bir testteki j maddesinin ayırt ediciliği $\alpha j = (\alpha_{1j}, \alpha_{2j}, \ldots\ldots \alpha_{mj})$ vektörüne eşittir. Ancak her bir maddenin ayırt ediciliği ilişkili olmadığı alt testlerde sıfırdır. Bu sebeple vektör $\alpha j = (0,0,.., \alpha_{vj}, 0,\ldots 0)$ şeklinde sadeleştirilir. Üç alt testten oluşan basit yapılı 3PLM'deki bir testin madde parametrelerinin genel görünüşü tablo 1'deki gibidir.

Tablo 1.Basit Yapıda Parametreler

Alt test (i)	Madde (j)	α_{1j}	α_{2j}	α_{3j}	β_{ij}	γ_{ij}
1	1	α_{11}	0	0	β_{11}	γ_{11}
1	2	α_{12}	0	0	β_{12}	γ_{12}
1	3	α_{13}	0	0	β_{13}	γ_{13}
2	4	0	α_{24}	0	β_{24}	γ_{24}
2	5	0	α_{25}	0	β_{25}	γ_{25}
2	6	0	α_{26}	0	β_{26}	γ_{26}
3	7	0	0	α_{37}	β_{37}	γ_{37}
3	8	0	0	α_{38}	β_{38}	γ_{38}
3	9	0	0	α_{39}	β_{39}	γ_{39}

Çoklu-tek boyutlu 3 parametreli lojistik modelde i bireyinin v alt testinde yer alan j maddesine doğru cevap verme olasılığı eşitlik 2'deki gibidir. Modelden şans parametresi çıkartılırsa çoklu-tek boyutlu 2PL modele ilişkin olasılık fonksiyonu elde edilir.

$$\left(U_{vij}=1\middle|\theta_{vi},\alpha_{vj},\beta_{vj},\gamma_{vj}\right)=\gamma_{vj}+\frac{(1-\gamma_{vj})}{1+e^{[(\sum_{l=1}^{v}\beta_{jl}\theta_{il})+\alpha_{vj}]}} \qquad (Eşitlik\ 2)$$

U_{vij} = i bireyin v alt testindeki j maddesine verdiği cevap (1 veya 0)

θ_{vi} = i bireyinin v alt testine ilişkin yetenek düzeyi

β_{vj} = v alt testindeki j maddesinin güçlüğü

$\alpha_{vj,}$ = v alt testindeki j maddesinin ayırıcılığı

γ_{vj} = v alt testindeki j maddesinin en düşük asimptotu (şans parametresi)

ÇBMTK'da veri setinin yapısına dayalı olarak ele alınan ikinci yapı karmaşık yapıdır. Karmaşık yapı, diğer bir ifade ile basit olmayan yapı, maddeler içi model olarak da bilinir. Bu modelde maddeler tek bir yetenek ile ilişkili değildir. Hem testin bütünü hem de testteki maddeler birden çok

yetenek ile ilişkilidir. Bunun bir sonucu olarak maddeler ölçülmek istenen birden fazla yeteneğe faktör yükü verir (Sheng ve Wikle, 2007; Bulut, 2013).

Boyutlar arası etkileşime dayalı yapılacak sınıflandırmada öncelikli koşul, ele alınacak maddelerin maddeler içi modele sahip olmasıdır. Bir madde birden fazla özelliği ölçüyorsa, ölçtüğü özellikler arasındaki etkileşime dayalı olarak telafisel veya telafisel olmayan (kısmi telafisel) model başlığı altında ele alınabilir. Telafisel modelde; bir bireyin bir yetenekteki zayıflığı, maddenin ilişkili olduğu diğer boyutta sahip olduğu yüksek yetenek tarafından telafi edilir ve bu şekilde bireyin maddeyi doğru cevaplama olasılığı artar. Bu modelin matematiksel alt yapısında, θ-koordinatlarının doğrusal kombinasyonuna ogive veya lojistik fonksiyon eklemesi yapılır. Telafisel modele örnek olarak yabancı dil sınavında yer alan astronomi ile ilgili bir paragraf sorusu verilebilir. Böyle bir soruda hem astronomi bilgisi hem de yabancı dil bilgisi yoklanmaktadır. Ancak astronomi bilgisi yüksek olan bir birey yabancı dile ilişkin yeteneği düşük olsa da maddeyi doğru cevaplayabilir. Ya da bireyin astronomi bilgisi yeterli olmasa dahi sahip olduğu yüksek yabancı dil bilgisi ile maddeyi doğru cevaplayabilir. Burada maddenin ilişkili olduğu bu iki yetenek arasında -astronomi ve yabancı dil- telafi edicilik ilişkisi söz konusudur. Telafisel modele bir başka örnek olarak matematikte birden fazla çözüm yolu içeren ve hem cebirsel işlem, hem geometride alan hem de trigonometri ile ilgili bir madde verilebilir. Buna göre öğrenci yoklanan üç boyutta yüksek yeteneğe sahip olmasa dahi sadece bir ya da iki alanda yüksek yeteneğe sahip olarak maddeyi doğru cevaplayabilir.

Telafisel olmayan modelde bireyin maddeyi doğru cevaplayabilmesi için maddenin ilişkili olduğu tüm boyutlarda belli bir yetenek düzeyine sahip olması gereklidir. Buna göre bu modelde, maddeyi doğru cevaplamak için bir boyutta sahip olunması gereken minimum düzeyden yüksek yetenek, diğer boyutta sahip olunması gereken minimum düzeyden düşük yeteneği telafi edemez. Bir bireyin bu tipteki bir maddeyi doğru cevaplama olasılığı maddeyi cevaplamak için gerekli yetenek olasılıklarının çarpımıdır (Ackerman, 1996; Ackerman, Gierl ve Walker, 2003; Reckase, 2009). Telafisel olmayan modele örnek olarak kimya dersindeki bir laboratuvar uygulaması ele alınabilir. Burada öğrencinin hem kimya bilgisine sahip olması hem de deneyi nasıl yapacağını bilmesi gerekir. Öğrencinin kimya alanında bilgisi ne kadar fazla olursa olsun, başarılı olması için ilgili deneyin nasıl yapılacağına ilişkin olarak da belli bir bilgi düzeyine sahip olmalıdır. Yapılan az sayıda çalışmada modellerin kullanışlılığı, gerçek veri setlerini doğru biçimde ele alma ve temsil edebilme bakımından test edilmiştir (Bolt ve Lall, 2003; Babcock, 2009). Sonuçlar telafisel modelin daha iyi sonuçlar ürettiği ve bu bakımdan daha kullanışlı olduğunu göstermektedir (Reckase, 2009). Karmaşık model ortaya koyan telafisel özellikteki 3PLM bir testte j bireyinin i maddesine doğru cevap verme olasılığına ilişkin olasılık fonksiyonu aşağıdaki gibidir.

$$P(U_{ij}=1|\theta_j,\alpha_i,\gamma_i,d_i)=\gamma_i+(1-\gamma_i)\frac{e^{\alpha_i\theta_j'+d_i}}{1+e^{\alpha_i\theta_j'+d_i}} \qquad (Eşitlik\ 3)$$

U_{ij} = j bireyinin i maddesine verdiği cevap (1 veya 0)

θ_j = j bireyinin yetenek düzeyi

α_i = i maddesinin ayırıcılığı

γ_i = i maddesinin en düşük asimptotu (şans parametresi)

d_i = i maddesinin kesim noktası

Denklemde yer alan d parametresi α ve β parametrelerinin etkileşimi ile elde edilen kesim noktasıdır. Buna göre tek boyutlu lojistik modellerde yer alan (α(θ - β)) çarpımı yapılırsa (αθ - αβ) elde edilir. Burada yer alan – αβ terimi d ile yer değiştirildiğinde αθ+ d ifadesi elde edilmiş olur. Tek boyutlu model çok boyutlu modele genişletildiğinde; m boyutlu bir test için α parametresi 1*m

elemanlı ayırt edicilik vektörünü; θ parametresi 1*m elemanlı birey koordinatları vektörünü temsil eder. d parametresi αθ+ d denkleminin kesim noktasını verir. Modelde yer alan "e" teriminin üssünde yer alan ifade α ve θ parametrelerinin etkileştiğini göstermek üzere aşağıdaki şekilde genişletilebilir:

$$\alpha_i\theta_j' + d_i = \alpha_{i1}\theta_{j1} + \alpha_{i2}\theta_{j2} + \alpha_{i3}\theta_{j3} + \cdots\ldots\ldots + \alpha_{im}\theta_{jm} + d_i = \sum_{l=1}^{m} \alpha_{il}\theta_{jl} + d_i \qquad (Eşitlik\ 4)$$

(Reckase, 2009).

Yukarıda da bahsedildiği gibi MTK farklı test yapılarına özgü farklı modeller ve çıktılar sunan zengin bir kuramdır. Klasik test kuramı ile aynı alanlarda uygulanabilmesine rağmen çoğu zaman klasik test kuramından daha iyi sonuçlar sunmaktadır (Hambleton ve Jones, 1993; Harvey ve Hammer, 1999; Sinar ve Zickar, 2002; McDonald ve Paunonen, 2002). Temelleri 1980'lerde ortaya konan (Sympson, 1978; Doody-Bogan ve Hattie, 1981; Yen, 1983; akt: Ansley ve Forsyth, 1985) ÇBMTK ile ilgili Türkiye'de doktora düzeyinde tamamlanmış birkaç tez mevcuttur (Köse, 2010; Sünbül, 2011; Özkan, 2012, Koğar, 2014; Yavuz, 2014). Bu çalışmalardan Köse (2010), Sünbül (2011), Özkan (2012) ve Koğar'ın (2014) çalışmaları tek boyutlu MTK ve çok boyutlu MTK'nın telafisel modeldeki iki boyutlu testleri ile sınırlıdır. Yapılan en güncel çalışmalardan Yavuz'un (2014) tezinde boyut sayısı 3 ve 5 olarak değişen çok boyutlu testler ele alınmıştır. Ancak bu çalışmada da testin yapısı basit yapı ve iki-faktör modeli şeklinde sabitlenmiştir. Gerçek uygulamalarda ise boyut sayısı 2'nin üzerinde olan karmaşık yapılı testlerden de yararlanılmaktadır. Yapılan çalışmalarda bu durumu örnekleyecek herhangi bir desen ele alınmamıştır. Ancak bu tür çok boyutlu testler ile çalışırken doğru kestirimler yapabilmek için uygun örneklem büyüklüğü ve test uzunluğu koşullarının belirlenmesi gereklidir. Bu nedenle bu çalışma ile farklı boyutluluk (boyut sayısı 2 ve 3) ve farklı yapı özelliklerindeki (basit yapılı ve karmaşık yapılı) testlerin ÇBMTK çerçevesinde ele alınmasına, ele alınan yapılara ilişkin farklı örneklem büyüklüğü ve test uzunluğu sınamalarının yapılarak kurama dayalı kestirimlerin değişen koşullardan nasıl etkilendiğinin incelenmesine gerek olduğu görülmüştür. Bu doğrultuda çalışmada "basit ve karmaşık yapılı iki ve üç boyutlu testlerden çok boyutlu madde tepki kuramıyla yapılan madde ve birey parametresi kestirimleri örneklem büyüklüğü ve test uzunluğundan nasıl etkilenmektedir?" problem cümlesine cevap aranmıştır.

Bu problem cümlesine dayalı olarak araştırmanın alt problem cümleleri şu şekilde ele alınmıştır:

1. İki boyutlu karmaşık ve basit yapılı testler için;

1.1. Örneklem büyüklüğü 1000, 2000 ve 4000; test uzunluğu 12 ve 48 olarak ele alındığında madde parametreleri kestirimlerine ilişkin RMSE, yanlılık, gerçek ve kestirilen parametreler arasındaki korelasyonlar nasıl değerler almaktadır?

1.2. Örneklem büyüklüğü 1000, 2000 ve 4000; test uzunluğu 12 ve 48 olarak ele alındığında birey parametreleri kestirimlerine ilişkin RMSE, yanlılık, gerçek ve kestirilen parametreler arasındaki korelasyonlar nasıl değerler almaktadır?

2. Üç boyutlu karmaşık ve basit yapılı testler için;

2.1. Örneklem büyüklüğü 1000, 2000 ve 4000; test uzunluğu 12 ve 48 olarak ele alındığında madde parametreleri kestirimlerine ilişkin RMSE, yanlılık, gerçek ve kestirilen parametreler arasındaki korelasyonlar nasıl değerler almaktadır?

2.2. Örneklem büyüklüğü 1000, 2000 ve 4000; test uzunluğu 12 ve 48 olarak ele alındığında birey parametreleri kestirimlerine ilişkin RMSE, yanlılık, gerçek ve kestirilen parametreler arasındaki korelasyonlar nasıl değerler almaktadır?

Araştırmanın Amacı

Çalışmanın amacı basit ve karmaşık yapıda olan iki ve üç boyutlu testlerin farklı örneklem büyüklüğü ve test uzunluğu koşullarında madde ve birey parametreleri kestirimlerine ilişkin RMSE ve yanlılık durumlarını belirlemektir. Ayrıca çalışmada elde edilen parametre kestirimleri ile gerçek parametreler arasındaki ilişki de incelemiştir. Elde edilen sonuçlara bağlı olarak çalışılan test yapılarında minimum hataya ve yanlılığa sahip kestirimler yapabilmek için gerekli örneklem büyüklüğü ve test uzunluğuna ilişkin önerilerin geliştirilmesi amaçlanmaktadır.

Çok boyutlu madde tepki kuramına ilişkin olarak yapılan çalışmalar genel olarak iki boyutlu testler üzerine yoğunlaşmıştır. Yapılan pek çok çalışmada basit ve karmaşık yapı ele alınmıştır. Fakat aynı çalışma içerisinde iki modeli birlikte ele alan çalışmalara pek rastlanmamaktadır. Türkiye'de yapılan ÇBMTK çalışmaları iki boyutlu testleri ve karmaşık yapıyı ele almakta; üç boyutlu yapıları ele alan ÇBMTK çalışması bulunmamaktadır. Bu doğrultuda basit ve karmaşık yapılı üç boyutlu testler ile ilgili bir çalışma da henüz mevcut değildir. Bu çalışma ÇBMTK kapsamında iki ve üç boyutlu testleri ele alması bakımından önem taşımaktadır. Ele alınan çok boyutlu testlerin karmaşık yapının yanında basit yapıyı da kapsamasının araştırmayı önemli kıldığı ve basit yapı olarak adlandırılan çoklu-tek boyutlu modelin tanıtılmasının, farklı boyut sayısı, örneklem büyüklüğü, test uzunluğu koşullarında incelenmesinin alan yazına katkı sağlayacağı düşünülmektedir. Türkiye'de yapılan çalışmalarda farklı kestirim yöntemleri için farklı yazılımlar kullanılmıştır, ancak IRTPRO yazılımı kullanılarak yapılmış bir çalışma bulunmamaktadır. Bu çalışmada IRTPRO 2.1 kullanıldığından, çalışmanın sonucunda yazılım ile ilgili olarak araştırmacılara yönelik öneri geliştirilmesinin, araştırmanın önemini arttırdığı düşünülmektedir.

YÖNTEM

Araştırmanın Türü

Bu çalışma, farklı örneklem büyüklüğü, test uzunluğu ve boyut sayısı koşulları altında çok boyutlu madde tepki kuramının farklı modellerinden yapılan parameter kestirimlerine ilişkin hata, yanlılık ve yapılan kestirimler ile gerçek parametrelerin korelasyonlarını incelemesi, bu doğrultuda bir kuramın söz konusu modellerini test etme amacını taşımasından dolayı temel araştırma niteliğindedir.

Araştırmanın Verileri

Çalışma için gereken veriler simülasyon yoluyla elde edilmiştir. Yapılan bir çalışmada simülasyon verisi kullanmanın avantajı gerçek madde ve yetenek parametrelerinin bilinmesidir. Böylece araştırmacı gerçek değerler ile kestirilen değerler arasında sağlıklı karşılaştırmalar yapabilir. Ancak dikkat edilmesi gereken iki önemli koşul vardır. Bunlardan ilki araştırmacı veri setini oluşturmak için uygun bir yöntemi seçmelidir. İkinci koşul ise oluşturulan simülasyon veri setlerinin test maddelerine verilecek gerçek cevapları temsil ettiğinden emin olunması gerektiğidir (Way, Ansley ve Forstyh, 1988).

Yukarıda bahsedilen koşulları karşılamak amacıyla verilerin üretimi ÇBMTK'ya uygun veri üretmek üzere hazırlanmış SimuMIRT (2003) yazılımından faydalanarak yapılmış ve parametrelerin gerçekte var olan bir durumu yansıtacak şekilde olmasına özen gösterilmiştir. Bunun için parametre üretilmesinde çeşitli dağılım özellikleri göz önünde bulundurulmuş ve madde ve birey parametreleri ile iki kategorili veri setleri oluşturulmuştur.

Çalışmada çok boyutlu yapılar sabit olarak ele alınmıştır. Buna göre sabitler: iki boyutlu karmaşık, iki boyutlu basit, üç boyutlu karmaşık ve üç boyutlu basit yapı şeklindedir. Ackerman (1989) çalışmasında simülasyon yoluyla ürettiği verilerin thetaları arasındaki korelasyon arttıkça veri setinin tek boyutlu yapıya yaklaştığını raporlamıştır. Bu çalışmada da desenlerin tek boyutlu yapıya yaklaşmasının önüne geçmek için boyutlar arası korelasyon tüm koşullarda düşük korelasyonu göstermek üzere 0,3 olarak sabitlenmiştir. Test uzunluğu ve örneklem büyüklüğü araştırmanın bağımsız değişkenlerini oluşturmaktadır.

Test uzunluğu: Bağımsız değişkenlerden biri olan test uzunluğunun parametre kestirimi üzerine etkisini test etmek amacıyla kısa ve uzun testleri temsil edecek şekilde iki test uzunluğu koşulu ele alınmıştır. Bu doğrultuda madde sayısının ele alınan boyut sayılarının tam katı olması, basit yapılı testlerin her bir alt testinde en az 3 maddenin yer alması koşullarının sağlanması amacıyla kısa testleri temsil etmek üzere 12, uzun testleri temsil etmek üzere 48 maddeli testlerden yararlanılmıştır. Basit yapılı testlerin her bir alt testinde eşit sayıda madde yer almıştır.

Örneklem Büyüklüğü: Araştırmanın amacına yönelik olarak kestirimler üzerine örneklem büyüklüğünün etkisini test etmek ve araştırmanın sonucunda örneklem büyüklüğü önerisinde bulunmak amacıyla farklı örneklem büyüklükleri ile çalışılmıştır. Literatürde çok boyutlu yapılarda en az 1000 kişilik örneklemler ile çalışılması önerilmektedir (Bolt ve Lall, 2003; Yao ve Boughton, 2007; Lee 2012). Bu sebeplerle çalışmada aynı dağılım özelliklerine sahip 1000, 2000 ve 4000 olmak üzere 3 örneklem büyüklüğü koşulu ele alınmıştır.

İşlem

Çalışmada iki boyutlu karmaşık, iki boyutlu basit, üç boyutlu karmaşık ve üç boyutlu basit yapıda, test uzunluğu 12 ve 48; örneklem büyüklüğü 1000, 2000, 4000 şeklinde değişen 4 ÇBMTK modeli * 2 test uzunluğu * 3 örneklem büyüklüğü koşulu olmak üzere 24 koşul içeren desen oluşturulmuştur. Araştırmanın deseni tablo 2'de verilmiştir. Desene ilişkin verilerin oluşturulması için kullanılan SimuMIRT'ın (2003) basit ve karmaşık yapılı testler için farklı araçları mevcuttur. Veri üretimi önce madde parametreleri, sonra yanıtlar ve birey parametreleri olmak üzere iki aşamada gerçekleşmektedir. Madde parametrelerini üretmek için ele alınan koşuldaki test yapısı ile ilgili yazılım dosyasına madde sayısı, boyut sayısı, basit yapılı testler için her bir boyutta yer alan madde sayısı, madde güçlüğü, madde ayırt ediciliği için ortalama ve standart sapma değerleri girilmiştir. Ayırt edicilik parametrelerinin (a_i) üretilmesinde ortalaması 0,5 ve standart sapması 0,4 olan log-normal dağılım, güçlük parametrelerinin (d_i) üretilmesinde standart normal dağılım kullanılmıştır. Sonuçta ilgili koşula ilişkin madde parametreleri elde edilmiştir ve ikinci aşamaya geçilmiştir. Bu aşamada ilgili dosyaya ilk aşamada elde edilen madde parametreleri kopyalanmış; üzerine madde sayısı, birey sayısı, boyut sayısı, örneklem ortalaması, varyans kovaryans matrisinin alt üçgeni, cevap kategorilerinin sayısı ile birey parametreleri ve yanıtları oluşturmak üzere random seed değerleri girilmiştir. Sonuçta ilk adımda üretilen madde parametrelerine dayalı olarak iki kategorili yanıtlar ile birey parametreleri oluşturulmuştur. Yetenek parametreleri her bir birey için her bir boyuta ilişkin olarak elde edilmiştir ve oluşturulmasında standart normal dağılımdan faydalanılmıştır. Ayrıca boyutlar arası korelasyonun 0,3 olduğunu göstermek üzere varyans kovaryans matrisinin köşegen dışı elemanları 0,3 olarak girilmiştir.

Dağılım özellikleri sabit kalmak üzere boyut sayısı ve test yapısına göre veri üretiminde farklılıklar mevcuttur. Buna göre iki ve üç boyutlu basit yapılı ÇBMTK modellerinde her bir madde sadece bir boyut ile ilişkili olduğu için madde sayısı kadar ayırt edicilik ve güçlük parametresi üretilmiştir (bknz. Tablo 1). Üretilen her bir ayırt edicilik parametresi bir madde ve bir boyut ile ilişkilendirilmiş, ilişkilendirildiği maddenin diğer boyutlardaki ayırt ediciliği sıfır olarak alınmıştır. İki ve üç boyutlu karmaşık yapılı ÇBMTK modellerinde veri üretimi telafisel modele göre yapılmıştır. Bu modelde her bir madde tüm boyutlar ile ilişkili olduğundan her bir madde için boyut sayısı kadar ayırt edicilik parametresi ile bir tane güçlük parametresi üretilmiştir. Böylece karmaşık yapılarda boyut sayısı * madde sayısı kadar ayırt edicilik parametresi ve madde sayısı kadar güçlük parametresi elde edilmiştir.

Her bir koşula ilişkin olarak yapılan bu işlem madde ve birey parametreleri sabit kalmak üzere yanıtlar için girilen random seed değiştirilerek 25 kere tekrar (replikasyon) edilmiştir. Bu tip simülasyon çalışmalarında tekrar örneklem büyüklüğü ile eş anlamlıdır. MTK çalışmalarında tekrar sayısı kestirilen parametrelerin örnekleme hatasının varyansını ve simülasyon sonuçlarına dayalı istatistiklerin yeterli güce sahip olmasını etkilemektedir (Harwell vd. 1996). Bu çalışmada da alan yazın önerileri ve analizlerin uzun sürmesi gibi sınırlılıklar göz önünde bulundurularak bir koşul 25

tekrar ile sınırlandırılmıştır. Sonuçta 24 koşul * 25 tekrar olmak üzere 600 veri seti elde edilmiştir. Verilerin üretilmesinden sonra analize geçilmiştir.

Tablo 1: Araştırmanın Deseni

Sabitler	Değişkenler	
Yapı	Örneklem Büyüklüğü	Test Uzunluğu
2 Boyutlu Karmaşık Yapı	1000	12 madde
		48 madde
	2000	12 madde
		48 madde
	4000	12 madde
		48 madde
2 Boyutlu Basit Yapı	1000	12 madde
		48 madde
	2000	12 madde
		48 madde
	4000	12 madde
		48 madde
3 Boyutlu Karmaşık Yapı	1000	12 madde
		48 madde
	2000	12 madde
		48 madde
	4000	12 madde
		48 madde
3 Boyutlu Basit Yapı	1000	12 madde
		48 madde
	2000	12 madde
		48 madde
	4000	12 madde
		48 madde

Verilerin Analizi

Üretilen veri setlerinin madde ve birey parametreleri test yapısına uygun betikle IRTPRO 2.1 yazılımından faydalanarak kestirilmiştir. IRTPRO 2.1 yazılımında kestirimler Bock-Aitkin Expectation-Maximization (BA-EM), Adaptive Quadrature ve Metropolis-Hastings Robbins-Monro (MH-RM) algoritmaları ile yapılabilmektedir. Ancak MH-RM algoritmasının kullanımı boyut sayısının ikiyi veya üçü aştığı durumlarda önerilmekte, Adaptive Quadrature algoritmasıyla yapılan birey parametresi kestirimlerinde her bir birey için quadrature nod sayısı değişebilmektedir (IRTPRO guide, 2011). Çalışmada ele alınan veri setlerinde boyut sayısının üçü geçmemesi ve quadrature nod sayıların sabit tutularak birey parametrelerinin kestirilmek istenmesi sebebiyle, analizler BA-EM algoritmasından faydalanarak yapılmıştır. Analizlerin seri biçimde yapılabilmesi için yazılım batch modunda çalıştırılmıştır. Testlerin analizlerinde analiz için gereken süre özellikle karmaşık yapılarda madde ve boyut sayısının artması ile artmış, 3 boyutlu karmaşık yapılarda en yüksek değerini almıştır. Buna göre IRTPRO 2.1, 3 boyutlu karmaşık yapının 48 madde ve 4000 örneklem büyüklüğü koşulundaki bir veri setinin madde ve birey parametrelerini kestirmek için Intel Core i7-4500 CPU 1.80GHz 2.40 GHZ 8GB (RAM) 64 bit işletim sistemi olan bir bilgisayarda ortalama 22 saat çalıştırılmıştır.

Analizlerin tamamlanmasından sonra madde ve birey parametrelerinin kestirilen ve gerçek değerlerinden faydalanarak GOR istatistikleri (kestirim iyiliği - goodness of recovery) hesaplanmıştır. GOR istatistikleri farklı koşulların etkisini araştırmak ve kestirim kararlığını belirlemek amacıyla parametreler için hesaplanan istatistiklerdir (Maris, 1999; Turhan, 2006). Bu

çalışmada her bir parametre için RMSE ve yanlılık olmak üzere iki GOR istatistiği hesaplanmıştır. RMSE gerçek ve kestirilen parametre değerleri arasındaki farkın kareleri ortalamasının kareköküdür. Bu istatistik aşağıda yer alan formülden faydalanarak hesaplanmıştır.

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}\left(\bar{\tau}_{nj} - \tau_{nj}\right)^2}{N}} \qquad (Eşitlik\ 5)$$

RMSE formülündeki terimlerden N terimi testte yer alan parametre sayısını temsil etmektedir. Parametre sayısı, RMSE'nin madde parametreleri için hesaplandığı durumda testteki madde sayısı; birey parametreleri için hesaplandığı durumda testteki birey sayısı olarak alınmıştır. τ_{nj} n. maddeye/bireye ait j parametresinin gerçek değeri iken, $\bar{\tau}_{nj}$ n. maddenin/bireyin j parametresinin kestirilen değeridir. Her bir koşulda her bir tekrar için parametrelerin RMSE değeri hesaplanmıştır. Daha sonra koşulda yer alan 25 tekrardan elde edilen RMSE değerlerinin ortalaması, ilgili parametrenin RMSE değeri şeklinde raporlanmıştır. RMSE minimum sıfır olmak üzere her zaman pozitif değer almaktadır. İlgili koşulda elde edilen RMSE değerlerinin sıfıra yakın olması kestirim kararlılığının yüksek, sıfırdan uzak olması kestirim kararlılığının düşük olduğu şeklinde yorumlanmaktadır.

Yanlılık, parametrenin kestirilen ortalama değeri ile gerçek değeri arasındaki farktır. Çalışmada yanlılık değerleri her bir madde ve her bir birey parametresi için hesaplanmıştır. Bunun için aşağıda yer alan eşitlikten yararlanılmıştır.

$$Yanlılık\left(\tau_j\right) = \left(\frac{\sum_{r=1}^{R}\bar{\tau}_{jr}}{R}\right) - \tau_j \qquad (Eşitlik\ 6)$$

Eşitlik 6'da yer alan R tekrar sayısını temsil etmektedir. $\bar{\tau}_{jr}$ j parametresinin r. tekrardaki kestirilen değeri, τ_j j parametresinin gerçek değeridir. Formüle göre her bir parametrenin her tekrarda kestirilen değerlerinin ortalamasının gerçek parametreden farkı bulunmuştur. Böylece madde parametreleri için madde sayısı kadar, birey parametreleri için birey parametresi kadar yanlılık değeri elde edilmiştir. Elde edilen yanlılıkların madde/birey sayısına bölünerek hesaplanan değeri ilgili koşuldaki parametrenin yanlılığı olarak raporlanmıştır. Yanlılık hem pozitif hem de negatif değerler alabilmektedir. Sıfır ve sıfıra yakın yanlılık değerleri parametre kestiriminin yansız yapıldığı şeklinde yorumlanmaktadır.

Araştırmada gerçek parametreler ile kestirilen parametreler arasındaki ilişkiyi görmek için korelasyonlar incelenmiştir. Bunun için her bir tekrardan kestirilen parametreler ile gerçek parametreler arasındaki Pearson Momentler Çarpımı Korelasyon katsayısı hesaplanmıştır. Desende yer alan 25 tekrar için bu işlem yapılmış ve ortalaması hesaplanarak rapor edilmiştir. Korelasyon katsayılarının hesaplanmasına dair izlenen bu süreç hem madde hem birey parametreleri için aynı şekildedir.

BULGULAR

Bu başlıkta araştırmanın bulguları alt problemler halinde verilmiştir:

1. Birinci alt problem için bulgular

1.1. İki boyutlu karmaşık ve basit yapılı testlerden kestirilen madde parametrelerine ilişkin bulgular:

İki boyutlu karmaşık ve basit yapılı testlerden kestirilen madde parametrelerine ilişkin bulgular tablo 3'teki gibidir:

İki boyutlu karmaşık yapılı testlerden yapılan madde parametresi kestirim sonuçlarına göre RMSE'ler 0,048 ile 0,323 arasında değişen değerlere sahiptir. Basit yapılı testlerde madde parametrelerine ait RMSE'ler 0,067 ile 0,174 arasında değişen ve genel itibari ile aynı koşullarda karmaşık yapıda ortaya çıkan değerlerden daha düşük değerler almıştır. Özellikle karmaşık yapılı testlerin kısa test koşulunda madde parametrelerinden elde edilen RMSE değerleri, uzun test koşulundaki RMSE değerlerinden belirgin biçimde yüksektir. Her iki yapıda da tüm parametreler örneklem büyüklüğünün artması ile daha düşük hatalı kestirilmiştir. Ayrıca d parametresi, her iki test yapısında da test uzunluğu ve örneklem büyüklüğüne ilişkin koşulların çoğunda a parametrelerinden daha düşük RMSE değerleri ile kestirilmiş; d parametresinin a1 ve a2 parametresine göre daha kararlı davrandığı gözlenmiştir.

Tablo 3: İki Boyutlu Testler İçin Madde Parametrelerine Ait Bulgular

Karmaşık Yapı										
Madde sayısı	Örneklem	RMSE			Yanlılık			Korelasyon		
		a1	a2	d	a1	a2	d	a1	a2	D
12 madde	1000	0,317	0,323	0,106	0,124	-0,151	-0,008	0,805	0,905	0,995
	2000	0,226	0,275	0,065	0,124	-0,169	0,006	0,891	0,945	0,998
	4000	0,212	0,262	0,048	0,140	-0,197	-0,016	0,917	0,966	0,999
48 madde	1000	0,135	0,139	0,090	0,000	-0,026	-0,020	0,946	0,946	0,997
	2000	0,093	0,096	0,061	-0,010	-0,020	-0,006	0,972	0,983	0,999
	4000	0,068	0,071	0,049	-0,013	-0,020	-0,019	0,985	0,991	0,999
Basit Yapı										
Madde sayısı	Örneklem	RMSE			Yanlılık			Korelasyon		
		a1	a2	d	a1	a2	d	a1	a2	d
12 madde	1000	0,171	0,174	0,120	-0,012	-0,008	0,013	0,959	0,967	0,983
	2000	0,132	0,131	0,094	-0,013	-0,003	0,017	0,969	0,976	0,985
	4000	0,097	0,102	0,076	-0,009	-0,021	0,013	0,978	0,981	0,987
48 madde	1000	0,122	0,114	0,132	-0,015	0,003	-0,003	0,974	0,979	0,956
	2000	0,092	0,084	0,111	-0,023	-0,003	0,003	0,981	0,985	0,957
	4000	0,073	0,067	0,096	-0,012	-0,001	-0,007	0,982	0,988	0,958

İki boyutlu karmaşık yapılı testlerde madde parametrelerine ait yanlılık değerleri -0,197 ile 0,140 arasında değişmektedir. İki boyutlu basit yapılı testlerde yanlılığın -0,023 ile 0,017 arasında değişen değerler aldığı görülmektedir. Aynı koşullar altında basit ve karmaşık yapıdan hesaplanan yanlılık değerleri karşılaştırıldığında basit yapının daha yansız kestirim sonuçlarına sahip olduğu görülmektedir. Yanlılık sonuçlarına göre örneklem büyüklüğü ve madde sayısındaki artış ile hem a parametrelerinin hem de d parametresinin yanlılığında düzenli bir azalış ya da artışın olmadığı ifade edilebilir.

Kestirilen madde parametreleri ile gerçek parametrelerin korelasyonlarının ortalaması iki boyutlu karmaşık yapılı testlerin tüm koşullarında 0,805 ve üzerinde korelasyon değerlerine sahiptir. İki boyutlu basit yapılı testlerde korelasyonlar 0,956 ile 0,988 arasında değişen mükemmel değerlerdir. RMSE ve yanlılık bakımından en düşük değerlere sahip olan d parametresi her iki test yapısında da tüm koşullarda gerçek parametrelerle 0,956 ve üzerinde pozitif yönde, mükemmel düzeyde ilişki göstermiştir. a parametrelerine ait korelasyonlar her iki test yapısında da örneklem büyüklüğünün artması ile daha yüksek değere ulaşmıştır.

1.2. İki boyutlu karmaşık ve basit yapılı testlerden kestirilen birey parametrelerine ilişkin bulgular:

İki boyutlu karmaşık ve basit yapılı testlerden kestirilen birey parametrelerine ilişkin bulgular tablo 4'teki gibidir:

İki boyutlu karmaşık yapılı testlerden kestirilen yetenek parametrelerine ilişkin RMSE değerleri incelendiğinde değerlerin kısa testte 0,715 ile 0,621; uzun testte 0,540 ile 0,489 arasında değiştiği görülmektedir. İki boyutlu basit yapılı testlerin birey parametrelerine ilişkin RMSE değerleri madde sayısının artması ile hem teta1 hem teta2 parametresi için 0,7 değeri etrafında iken 0,5 değeri etrafına gerilemiştir. Tablonun geneli incelendiğinde yetenek parametrelerine ilişkin RMSE değerlerinin örneklem büyüklüğünün artması ile ciddi bir değişime uğramadığı görülmektedir. Ancak her iki test yapısında da kısa testten uzun teste geçilmesi ile yetenek parametrelerinin RMSE değerlerinde düşme gözlenmiştir. Bu durum, birey parametrelerinin test uzunluğundan etkilendiği; örneklem büyüklüğünden etkilenmediği şeklinde yorumlanabilir.

Tablo 2: İki Boyutlu Testler İçin Birey Parametrelerine Ait Bulgular

Karmaşık Yapı							
Madde Sayısı	Örneklem	RMSE		Yanlılık		Korelasyon	
		θ1	θ2	θ1	θ2	θ1	θ2
12 madde	1000	0,715	0,634	-0,016	0,040	0,686	0,774
	2000	0,704	0,621	-0,210	0,022	0,693	0,786
	4000	0,708	0,623	-0,008	0,030	0,696	0,788
48 madde	1000	0,540	0,496	-0,005	0,039	0,835	0,876
	2000	0,531	0,489	-0,012	0,022	0,838	0,880
	4000	0,533	0,491	0,005	0,030	0,840	0,880
Basit Yapı							
Madde Sayısı	Örneklem	RMSE		Yanlılık		Korelasyon	
		θ1	θ2	θ1	θ2	θ1	θ2
12 madde	1000	0,761	0,749	0,027	0,072	0,631	0,668
	2000	0,756	0,741	0,022	0,054	0,630	0,676
	4000	0,761	0,744	0,040	0,061	0,635	0,677
48 madde	1000	0,499	0,526	0,022	0,043	0,865	0,865
	2000	0,492	0,523	0,015	0,025	0,863	0,862
	4000	0,495	0,524	0,033	0,033	0,865	0,864

Karmaşık yapılı iki boyutlu testlerden hesaplanan yanlılık sonuçlarına göre birey parametreleri, -0,210 ile 0,040 değerleri arasında değişen bir yanlılıkla kestirilmiştir. Basit yapılı iki boyutlu testlerden hesaplanan yanlılık değerleri ele alınan tüm koşullarda pozitiftir ve 0,015 ile 0,072 arasında değişen değerler almaktadır. Yine basit yapıda tüm koşullarda birey sayısı arttıkça yanlılık değerleri önce azalmış sonra artmıştır. Karmaşık yapılı testte teta2 için aynı durum söz konusu iken teta1 için tam tersi gerçekleşmiştir. Yanlılık değerleri genel olarak incelendiğinde değerlerin örneklem büyüklüğü ve test uzunluğundaki artış ile aynı yönde olmayan bir örüntüye sahip olduğu söylenebilir.

Her bir tekrarda kestirilen birey parametreleri ile gerçek birey parametrelerinin korelasyonlarının ortalamalarını veren tablo 4'te görüldüğü gibi yetenek parametrelerine ait korelasyon değerleri karmaşık yapıda 0,686 ile 0,880 arasında; basit yapıda 0,630 ile 0,865 arasında değişmektedir. Her iki yapıda da genel olarak testteki madde sayısı arttıkça teta parametrelerine ait korelasyon değerleri artmıştır. Hem basit hem karmaşık yapıda RMSE'si daha düşük değer alan teta parametrelerinin aynı koşullarda gerçek parametrelerle gösterdiği ilişki daha yüksek değerler almaktadır. Madde sayısı sabit kalmak üzere örneklem büyüklüğü arttırıldığında korelasyon değerinde ciddi bir artış meydana

gelmemiştir. Bu bulgu RMSE değerleri ile birlikte ele alındığında birey parametrelerinin kestiriminin örneklem büyüklüğünden etkilenmediği bulgusunu destekler niteliktedir.

2. İkinci alt problem için bulgular

2.1. Üç boyutlu karmaşık ve basit yapılı testlerden kestirilen madde parametrelerine ilişkin bulgular:

Üç boyutlu karmaşık ve basit yapılı testlerden kestirilen madde parametrelerine ilişkin bulgular tablo 5'te verilmiştir. Tablo 5'e göre üç boyutlu karmaşık yapılı testlerden kestirilen madde parametrelerine ilişkin RMSE değerleri, iki boyutlu desenlerde elde edilen RMSE değerlerinden farklılık göstermekte ve karmaşık yapıda hesaplanan RMSE'ler oldukça yüksek değerler almaktadır. RMSE'ler 0,054 ile 1,508 arasında değişen geniş bir aralığa sahiptir. En yüksek hata ile kestirilen parametre 12 maddeli test ve 1000 örneklem büyüklüğü koşulunda kestirilen a3 parametresi olmuştur. Üç boyutlu karmaşık yapılarda elde edilen yüksek RMSE değerlerine karşılık üç boyutlu basit yapılarda RMSE değerleri 0,241 ile 0,042 arasında değerler almaktadır. Üç boyutlu basit ve karmaşık yapılarda kısa testten uzun teste geçildiğinde madde parametrelerine ait RMSE değerlerinde düşme gözlenmiştir. Hem basit hem de karmaşık yapının uzun test koşulunda a1, a2 ve a3 parametrelerinin RMSE değerleri birbirine çok yakındır. İki test uzunluğu koşulunda da her iki yapıda en düşük RMSE değerlerine sahip parametre d parametresi olmuştur. Bu bulgu iki boyutlu desenlerde elde edilen bulgulara paralellik göstermektedir. Bu sonuca dayalı olarak üç boyutlu basit yapılarda da en kararlı kestirilen madde parametresinin d parametresi olduğu yorumuna gidilebilir.

Tablo 3: Üç Boyutlu Testler İçin Madde Parametrelerine Ait Bulgular

Karmaşık Yapı													
Madde sayısı	Örneklem	RMSE				Yanlılık				Korelasyon			
		a1	a2	a3	d	a1	a2	a3	d	a1	a2	a3	d
12 madde	1000	0,446	1,289	1,508	0,602	0,002	0,210	0,247	-0,174	0,836	0,358	0,435	0,899
	2000	0,474	1,350	1,290	0,530	0,172	0,193	0,077	-0,152	0,888	0,336	0,601	0,912
	4000	0,371	0,624	0,640	0,181	0,173	0,051	-0,163	-0,042	0,903	0,562	0,615	0,977
48 madde	1000	0,298	0,325	0,279	0,118	-0,011	-0,150	0,092	-0,042	0,857	0,884	0,889	0,996
	2000	0,273	0,294	0,254	0,078	-0,039	-0,159	0,110	-0,015	0,880	0,911	0,915	0,998
	4000	0,274	0,273	0,245	0,054	-0,008	-0,149	-0,113	-0,014	0,872	0,928	0,913	0,999
Basit Yapı													
Madde sayısı	Örneklem	RMSE				Yanlılık				Korelasyon			
		a1	a2	a3	d	a1	a2	a3	d	a1	a2	a3	d
12 madde	1000	0,241	0,206	0,186	0,103	0,008	-0,037	0,025	-0,016	0,955	0,942	0,980	0,993
	2000	0,137	0,134	0,132	0,072	-0,014	-0,043	0,012	-0,007	0,984	0,972	0,990	0,996
	4000	0,119	0,092	0,089	0,049	0,004	-0,024	-0,019	0,002	0,987	0,986	0,995	0,998
48 madde	1000	0,114	0,110	0,109	0,082	0,004	-0,038	-0,008	-0,015	0,990	0,987	0,987	0,996
	2000	0,075	0,078	0,086	0,060	-0,001	-0,031	-0,007	-0,004	0,996	0,994	0,992	0,998
	4000	0,058	0,053	0,061	0,042	0,005	-0,011	-0,019	-0,005	0,997	0,997	0,996	0,999

Üç boyutlu karmaşık yapılı veri setlerinde yapılan en yanlı kestirim testin kısa ve örneklem büyüklüğünün 1000 olduğu koşulda a3 parametresine (0,247); en yansız kestirim ise testin kısa ve örneklem büyüklüğünün 1000 olduğu koşulda a1 parametresine (0,002) aittir. Diğer taraftan üç boyutlu basit yapılı veri setlerinden kestirilen madde parametrelerinin yanlılıkları 0,043 ile 0,025 arasında değişen, birbirine ve sıfıra yakın değerler almıştır. Basit yapının tüm koşulları göz önünde bulundurulduğunda en yanlı kestirilen parametre, kısa test ve 2000 örneklem büyüklüğü koşulunda a2 parametresi iken, en yansız kestirilen parametre kısa test ve 4000 örneklem koşulunda kestirilen d parametresi olmuştur. Üç boyutlu veri setlerinden kestirilen madde parametrelerinin yanlılıkları

daha önce incelenen iki boyutlu yapılara benzer biçimde madde sayısının ve örneklem büyüklüğünün artması ile düzenli bir artma ya da azalma göstermemiştir.

Tablo 5'te yer alan kestirilen madde parametreleri ile gerçek madde parametreleri arasındaki korelasyon değerlerine göre üç boyutlu karmaşık yapılı testler için korelasyon ortalamaları 0,336-0,999 arasında; basit yapılı testlerde ise korelasyonlar 0,942 ile 0,999 arasında değişen mükemmel değerlere sahiptir. Bu sonuçlar RMSE sonuçları ile paralellik göstermektedir. RMSE'nin 1'den büyük olarak hesaplandığı koşullarda korelasyonlar orta düzeyde çıkmıştır. Örneklem büyüklüğü ve test uzunluğu arttıkça her bir parametre için elde edilen korelasyon değerleri de artmıştır. d parametresi hem basit hem de karmaşık yapıdaki her iki test uzunluğunda ve tüm örneklem büyüklüklerinde 0,899 ve üzerindeki değerlerle gerçek parametrelerle mükemmel düzeyde ilişkiye sahiptir. Bu bulgu d parametresinin daha kararlı kestirildiği bulgusuna paraleldir.

2.2. Üç boyutlu karmaşık ve basit yapılı testlerden kestirilen birey parametrelerine ilişkin bulgular:

Üç boyutlu karmaşık ve basit yapılı testlerden kestirilen birey parametrelerine ilişkin bulgular tablo 6'daki gibidir. Üç boyutlu testlerden kestirilen birey parametrelerine ilişkin RMSE değerleri karmaşık yapıda 0,542 ile 0,848 arasında; basit yapıda 0,553 ile 0,838 arasında değişmektedir. Her iki yapıdan aynı koşullarda kestirilen RMSE değerleri birbirine yakın değerler almakla beraber; uzun testte boyut başına düşen madde sayısının artması ile basit yapıda genel olarak karmaşık yapıdan daha düşük RMSE değerleri görülmektedir. Örneklem büyüklüğünün artması parametrelerin RMSE değerlerini ciddi bir biçimde etkilememiş, test uzunluğu sabit tutulmak üzere örneklem büyüklüğü arttırıldığında bazı yetenek parametrelerinin RMSE'lerinde düşüş, bazılarında artış gerçekleşmiştir. Testte yer alan madde sayısı arttığında her iki test yapısında da elde edilen RMSE'lerde fark edilir bir düşüş olmuştur.

Tablo 4 : Üç Boyutlu Testler İçin Madde Parametrelerine Ait Bulgular

Karmaşık Yapı										
Madde Sayısı	Örneklem	RMSE			Yanlılık			Korelasyon		
		θ1	θ2	θ3	θ1	θ2	θ3	θ1	θ2	θ3
12 madde	1000	0,721	0,730	0,848	0,043	0,038	0,067	0,695	0,626	0,526
	2000	0,753	0,753	0,838	0,007	0,034	0,070	0,718	0,603	0,531
	4000	0,686	0,753	0,819	0,010	0,036	0,046	0,731	0,649	0,556
48 madde	1000	0,619	0,542	0,583	0,009	0,006	0,046	0,785	0,821	0,823
	2000	0,608	0,546	0,575	-0,029	-0,001	0,044	0,793	0,825	0,830
	4000	0,605	0,558	0,576	-0,021	0,003	0,029	0,797	0,833	0,823
Basit Yapı										
Madde Sayısı	Örneklem	RMSE			Yanlılık			Korelasyon		
		θ1	θ2	θ3	θ1	θ2	θ3	θ1	θ2	θ3
12 madde	1000	0,827	0,832	0,768	0,044	0,045	0,085	0,542	0,480	0,650
	2000	0,817	0,838	0,763	0,015	0,037	0,084	0,574	0,493	0,655
	4000	0,819	0,819	0,752	0,024	0,042	0,068	0,578	0,515	0,655
48 madde	1000	0,567	0,580	0,612	0,054	0,018	0,064	0,828	0,803	0,803
	2000	0,553	0,587	0,611	0,016	0,010	0,062	0,833	0,804	0,805
	4000	0,554	0,592	0,605	0,024	0,014	0,047	0,835	0,815	0,802

Üç boyutlu testlerden elde edilen birey parametrelerine ilişkin yanlılık değerleri karmaşık yapıda -0,029 ile 0,070 arasında; basit yapıda 0,085 ile 0,010 arasında değişen değerler almıştır. Basit yapıda ele alınan tüm koşullarda; karmaşık yapıda ise neredeyse tüm koşullarda birey parametresi kestirimleri pozitif yanlılık ile yapılmıştır. Her iki test yapısında da birey parametrelerine ait yanlılık, örneklem büyüklüğü ve madde sayısından ciddi biçimde etkilenmemiştir. Ele alınan koşullarda

hesaplanan yanlılık değerlerinin yakın olduğu görülmektedir. Buna göre yetenek parametrelerinin tüm koşullarda benzer yanlılıkla kestirildiği sonucuna ulaşılabilir.

Üç boyutlu yapılardan birey parametreleri için elde edilen korelasyon ortalamaları genel itibari ile RMSE değerleri ile paralel bulgulara sahiptir. Buna göre RMSE'nin yüksek olduğu kısa test koşulunda madde parametreleri için korelasyonlar düşük; RMSE'nin daha düşük değerler aldığı uzun test koşulunda korelasyonlar daha yüksek değerler almıştır. Korelasyon ortalamaları karmaşık yapıda 0,526 ile 0,833; basit yapıda 0,480 ile 0,835 arasında değişmektedir. Testlerde yer alan madde sayısının artması korelasyon değerlerinin artmasını sağlamıştır. Genel olarak 12 maddeli testlerde örneklem büyüklüğünün artışı, 48 maddeli testlerde örneklem büyüklüğü artışından daha etkili olmuştur.

SONUÇLAR ve TARTIŞMA

Bu çalışmada iki ve üç boyutlu yapıların madde ve birey parametresi kestirimlerinin farklı örneklem büyüklüğü ve test uzunluklarından nasıl etkilendiği RMSE, yanlılık ve korelasyonlar ölçüt alınarak incelenmiştir. Örneklem büyüklüğünün ve test uzunluğunun parametre kestirimi üzerinde çok etkili olduğu bilinmektedir. Bu etkilerin niteliği ve niceliği pek çok çalışmanın konusu olmuştur. (Stone, 1992; Sinar ve Zickar, 2002; Bolt ve Lall, 2003; De Mars, 2003; de la Torre ve Patz, 2005; Sheng ve Wikle, 2007; Babcock, 2009; Finch, 2010; Finch, 2011; Sünbül, 2011; Lee, 2012; Kieftenbeld and Natesan, 2012; Zhang, 2012). Bu çalışmanın bulgularına ait sonuç ve öneriler madde ve birey parametrelerine ilişkin olmak üzere iki başlık altında verilmiştir.

1. Madde Parametrelerine İlişkin Tartışma ve Yorum

Çalışmadan elde edilen bulgulara göre örneklem büyüklüğünün artması hem iki hem de üç boyutlu testlerde madde parametrelerine ilişkin RMSE değerlerinin düşmesine sebep olmuştur. Bu bulgu Bolt ve Lall (2003), Finch (2010) ve Zhang'ın (2012) elde ettiği sonuçlar ile paralellik göstermektedir. Zhang (2012) yaptığı çalışma ile boyutlar arası korelasyon ve madde sayısı sabit iken birey sayısı arttıkça parametrelerin RMSE değerlerinin azaldığını ifade etmiş ve birey sayısının artmasının çok boyutlu yapıların kestirimleri için iyi sonuçlar oluşturduğunu ifade etmiştir. Finch (2010) ise madde ve örneklem büyüklüğünün yüksek olduğu durumda düşük hatalı ve kararlı kestirimler yapılabileceğini belirtmiş ve çalışmasında madde ayırıcılığına ait RMSE değerlerinin ve güçlük parametresine ilişkin standart hataların birey ve madde sayısı arttıkça düştüğünü, ancak yanlılığın büyük örneklemlerde daha yüksek değerlere sahip olduğunu rapor etmiştir. Bu sebeple araştırmacı, çalışmasında, çok sayıda birey ve fazla sayıda madde kullanmanın her zaman doğru bir ÇBMTK parametre kestirimini sağlamayacağını ifade etmiştir. Bu çalışmada elde edilen bulgular Finch'in (2010) çalışmasından elde edilen bulgulara paralellik göstermektedir. Çalışmadan elde edilen sonuçlara göre madde parametrelerine ilişkin yanlılık değerleri örneklem büyüklüğünün ve/veya madde sayısının artmasına bağlı olarak düzenli bir artış veya azalış sergilememiştir. Örneğin 3 boyutlu basit yapılı kısa testten kestirilen a1 parametresinin yanlılık değeri örneklem büyüklüğü 1000 iken 0,008; 2000 iken -0,014 ve 4000 iken 0,004 olarak hesaplanmıştır. Buna göre bu parametrenin kestirim yanlılığı örneklem büyüklüğünün ilk kez artışı (1000'den 2000'e) ile artmış, ikinci kez artışı (2000'den 4000'e) ile azalmıştır. Benzer durum farklı koşullarda diğer parametreler için de geçerlidir. Bu çalışmada ele alınan yapılara göre madde parametrelerinin hesaplanan mutlak yanlılık değerlerinin alt ve üst sınırları incelenirse, boyut sayısı arttıkça ve yapı basitten karmaşığa geçtikçe yanlılık değerlerinin arttığı görülmektedir. Bu değerlere göre yanlılık değerlerinin artışı boyut sayısı ve yapının karmaşık olup olmaması ile ilişkilidir. Bu bulgu Lee'nin (2012) ifadeleriyle paralellik göstermektedir. Lee (2012) çalışmasında ÇBMTK modelinin parametre kestirimini etkileyen faktörlerin artmasının kestirim yanlılığının artışına sebep olduğunu belirtmiştir.

Çalışmadan elde edilen bir diğer bulgu da elde edilen yanlılık ve RMSE değerlerinin aynı boyut sayısı, test uzunluğu ve örneklem büyüklüğü koşullarındaki basit yapılı ve karmaşık yapılı testlerde farklılık göstermesidir. Buna göre iki boyutlu basit yapılarda a1 ve a2 parametreleri için elde edilen

yanlılık ve RMSE değerleri tüm test uzunluğu ve örneklem büyüklüğü koşullarında karmaşık yapıdan daha düşük değerlerdedir.

İki boyutlu karmaşık yapılı testlerde d parametresi daha kararlı kestirilmiştir ve bu parametreye ilişkin RMSE değerleri tüm koşullarda, yanlılık değerleri ise iki koşul (uzun testteki 1000 ve 4000 örneklem koşulları) dışındaki tüm koşullarda basit yapıdan daha düşük değerlere sahiptir. Bu bulguları doğrular biçimde madde parametrelerine ilişkin elde edilen korelasyon ortalamaları da a1 ve a2 parametreleri için basit yapılı testte, d parametresi için karmaşık yapılı testte daha yüksek değerlerde olma eğilimindedir. Üç boyutlu yapılarda tüm örneklem büyüklüğü ve test uzunluğu koşullarında basit yapıdan elde edilen RMSE ve yanlılık değerleri, karmaşık yapıdan elde edilen RMSE ve yanlılık değerlerinden daha düşüktür. Bu durum sadece kısa test ve 1000 örneklem büyüklüğü koşulunda a1 parametresinin yanlılık değeri için istisna teşkil etmektedir. Üç boyutlu yapılardan elde edilen madde parametrelerinin korelasyonlarının ortalaması ise yukarıdaki bulguyu destekler biçimde basit yapılarda daha yüksek değerlere sahiptir. Uzun test koşulunda kestirilen d parametrelerinin gerçek d parametreleri ile korelasyonlarının ortalaması hem basit hem karmaşık yapıda bire bir aynı değerlere sahiptir. Basit ve karmaşık yapıların parametre kestirimlerinin RMSE, yanlılık ve korelasyonların bu şekilde farklılık göstermesi Finch (2011) ile tutarlık göstermektedir. Finch (2011) yaptığı çalışmada basit yapıda olmayan maddelerin basit yapıya göre daha yüksek yanlılık ve standart hata ile kestirildiğini, basit yapıdaki maddelerin daha kararlı kestirildiğini ifade etmiştir.

Bu çalışmadan madde parametrelerine ilişkin olarak elde edilen diğer bir bulgu ele alınan koşulların büyük bir kısmında d parametresine ait RMSE değerlerinin aynı koşullarda kestirilen a parametrelerinden daha düşük, d parametresine ait korelasyon ortalamalarının a parametrelerininkinden daha yüksek olmasıdır. Bu bulgu Zhang (2012) ile Sheng ve Wikle (2007) tarafından yapılan çalışmalar ile örtüşmektedir. Zhang (2012) d parametrelerinin daha kararlı kestirildiğini; Sheng ve Wikle (2007) ise ele aldıkları koşullarda d parametrelerine ait RMSE'lerin a parametrelerininkinden daha düşük olduğunu, dolayısıyla d parametrelerinin a parametrelerinden daha kararlı kestirildiğini ifade etmişlerdir.

Elde edilen sonuçlara dayanarak yapılacak parametre kestirimlerinde boyut sayısı arttıkça madde parametresi kestirimlerinin iyileştirilmesi için madde ve birey sayısı arttırılmalıdır. BA-EM ile yapılacak kestirimlerde madde parametrelerinin RMSE değerlerinin düşük olması için iki boyutlu basit yapılarda en az 12 maddelik testler ve en az 2000 kişilik örneklemin kullanılması önerilmektedir. Ancak iki boyutlu karmaşık yapılar için 12 maddelik testlerde tatmin edici sonuçlara ulaşılamamaktadır. Bu yapılarda 48 maddenin yer aldığı testler ve 1000 kişilik örneklem büyüklüğü kullanıldığında kabul edilebilir düzeyde RMSE sonuçlarına ulaşılabilir. Üç boyutlu basit yapılarda 4000 örneklem ve üstünde 12 maddelik kısa testlerden faydalanılması önerilmektedir. Üç boyutlu karmaşık yapılarda ise 48 madde ve 4000 örneklemin üstüne çıkılmasına ihtiyaç vardır.

2. Birey Parametrelerine İlişkin Tartışma ve Yorum

Bu çalışmada ele alınan koşullarda madde parametrelerinin yanı sıra birey parametreleri de kestirilmiştir. Buna göre; ele alınan tüm boyutluluk koşullarında madde sayısı sabit tutulduğunda birey sayısının artması birey parametrelerinin (tetaların) kestirim iyiliğini önemli biçimde etkilememiş, teta parametrelerinin RMSE değerlerinde ciddi bir fark meydana getirmemiştir. Buna paralel olarak gerçek teta parametreleri ile kestirilen teta parametreleri arasındaki korelasyonlarının ortalaması da örneklem büyüklüğü artışından etkilenmemiştir. Ancak birey parametrelerinin kestirim iyiliği test uzunluğunun artması ile değişim göstermiştir. Buna göre 12 maddeli testten 48 maddeli teste geçildiğinde birey parametresi kestirimine ilişkin RMSE değerlerinde ciddi bir azalma meydana gelmiştir. Kirisci, Hsu ve Yu (2001), de la Torre ve Patz (2005), Köse (2010) ve Kieftenbeld ve Natesan (2012) çalışmalarında örneklemin büyümesinin birey parametrelerinin daha iyi kestirilmesine bir etkisinin olmadığını, ancak test uzunluğunun artmasıyla teta kestirim iyiliğinin geliştirilebileceğini belirtmişlerdir. Bu bakımdan bu çalışmadan elde edilen sonuçlar literatür ile paralellik göstermektedir.

Bu çalışmaya göre madde sayısının artmasıyla gerçek birey parametreleri ile kestirilen birey parametreleri arasındaki korelasyonlar yükselmiştir. Bu bulgu de la Torre ve Patz'ın (2005) bulguları ile tutarlıdır. de la Torre ve Patz (2005) yaptıkları çalışmada test uzunluğu ve boyut sayısı koşulları tek başına arttığında elde edilen korelasyonların da arttığını göstermişlerdir. Sheng ve Wikle (2007) ise yaptıkları çalışmada boyutlar arası korelasyonu değişen iki boyutlu basit yapıların teta parametrelerini incelemişlerdir. Buna göre bu çalışmaya en yakın koşulda teta1 ve teta2 parametrelerinin gerçek ve kestirilen değerleri arasındaki korelasyonu 0,80 ve 0,86 olarak bulmuşlardır. Eldeki çalışmada da Sheng ve Wikle'ın (2009) koşullarına benzeyen iki boyutlu basit yapının uzun test koşulunda elde edilen korelasyonlar 0,86 değerindedir. Bu açıdan elde edilen korelasyon değerleri ilgili alan yazın ile tutarlık göstermektedir. Madde parametrelerine benzer biçimde birey parametrelerinde de boyut sayısı arttıkça ve yapı basitten karmaşığa geçtikçe aynı koşullar için elde edilen RMSE değerleri artmış ve korelasyon değerleri azalmıştır.

Birey parametresi için basit yapılı ÇBMTK kestirimleri sonunda pozitif yanlılıklar elde edilmiş diğer bir ifade ile birey parametreleri için üst kestirimde bulunulmuştur. Boyut sayısı arttıkça hesaplanan yanlılık değerlerinin ranjı genişlemiştir. Bu sebeple birey parametresi için kestirimi etkileyen faktör sayısının artmasının kestirim iyiliğini düşürdüğü söylenebilir. Bu bulgu Lee'nin (2012) ÇBMTK'da madde parametrelerinin kestirim yanlılığı ile ilgili bulgularıyla paralellik göstermektedir.

Elde edilen sonuçlara göre boyut sayısı arttıkça birey parametresi kestirimlerinin iyileştirilmesi için madde sayısı arttırılmalıdır. Karmaşık yapılı testlerde kararlı kestirimler yapmak için, aynı boyut sayısına sahip basit yapıdaki testlerde kararlı kestirim yapılmasını sağlayan madde ve birey sayısından daha fazla madde ve bireye gerek olduğu göz önünde bulundurulmalıdır. Birey parametresi kestirimlerini iyileştirmek için, hesaplanan korelasyon değerleri göz önünde bulundurulduğunda, iki boyutlu (basit ve karmaşık) yapılar ve üç boyutlu basit yapılar için en az 12 maddelik, üç boyutlu karmaşık yapılarda ise en az 48 maddelik testlerin kullanılması önerilmektedir.

KAYNAKÇA

Ackerman, T.A. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement, 20*(4), 311-329.

Ackerman, T.A., Gierl, M.J., & Walker, C.M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37-51.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, *9*(1), 37-48.

Babcock, B. G. E. (2009). *Estimating a noncompensatory IRT model using a modified Metropolis Algorithm.* Unpublished Doctoral Dissertation. University Of Minnesota Faculty Of The Graduate School.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov Chain Monte Carlo. *Applied Psychological Measurement, 27*(6), 395-414.

Bulut, O. (2013). *Between-person and within-person subscore reliability: Comparison of unidimensional and multidimensional IRT models.* Unpublished doctoral dissertation, University Of Minnesota Faculty Of The Graduate School.

de la Torre, J., & Patz, R. L. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics, 30*(3), 295-311.

DeMars, C.E. (2003). Sample size and the recovery of nominal response model item parameters. *Applied Psychological Measurement, 27*(4), 275-288.

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

Finch, H. (2010) Item parameter estimation for MIRT model: Bias and precision of confirmatory factor analysis based models. *Applied Psychological Measurement, 34*(1), 10-26.

Finch, H. (2011) Multidimensional item response theory parameter estimation with nonsimple structure items. *Applied Psychological Measurement, 35*(1), 67-82.

Hambleton, R. K., & Jones, R. W. (1993). An NCME module on comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38–47.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory principles and applications*. Kluwer-Nijhoff Publishing. Boston-USA.

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation, 35*(2009), 57–63

Harvey, R.J., & Hammer, A. L. (1999). Item response theory. *The Counselling Psychologist, 27*(3), 353-383.

Harwell, M., Stone, C.A., Hsu, T.C., & Kirisci L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101-125.

IRTPRO 2.1 for Windows (Item Response Theory for Patient-Reported Outcomes). *Scientific Software International.*

IRTPRO: Users Guide (2011). *Scientific Software International.*

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov Chain Monte Carlo Estimation. *Applied Psychological Measurement, 36*(5), 399-419.

Kirisci, L., Hsu, T., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25*(2), 146-162.

Koğar, H. (2014). *Madde tepki kuramının farklı uygulamalarından elde edilen parametrelerin ve model uyumlarının örneklem büyüklüğü ve test uzunluğu açısından karşılaştırılması.* Yayımlanmamış doktora tezi, Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü.

Köse, İ.A. (2010). *Madde tepki kuramına dayalı tek boyutlu ve çok boyutlu modellerin test uzunluğu ve örneklem büyüklüğü açısından karşılaştırılması.* Yayımlanmamış doktora tezi, Ankara Üniversitesi Eğitim Bilimleri Enstitüsü.

Lee, J. (2012). *Multimensional item response theory: An investigation of interaction effects between factors on item parameter recovery using Markov Chain Monte Carlo.* Unpublished doctoral dissertation, Michigan State University Measurement and Quantitative Methods.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*(2), 187-212.

McDonald, P., & Paunone, S. V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement, 62*(6), 921-943.

Özkan, Y. Ö. (2012). *Öğrenci başarılarının belirlenmesi sınavından (ÖBBS) klasik test kuramı, tek boyutlu ve çok boyutlu madde tepki kuramı modelleri ile kestirilen başarı puanlarının karşılaştırılması.* Yayımlanmamış doktora tezi, Ankara Üniversitesi Eğitim Bilimleri Enstitüsü.

Reckase, M. D. (2009). *Multidimensional item response theory (Statistics for social and behavioral sciences)*. New York: Springer.

Sheng Y. and Wikle C. K. (2007). Comparing multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement*, *68*(3), 413-430.

SimuMIRT (2003). Software. *Lihua Yao.*

Sinar, E.F., & Zickar, M. J. (2002). Evaluating the robustness of graded response model and classical test theory parameter estimates to deviant items. *Applied Psychological Measurement, 26*(2), 181-191.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement*, *16*(1), 1-16.

Sünbül, Ö. (2011). *Çeşitli boyutluluk özelliklerine sahip yapılarda, madde parametrelerinin değişmezliğinin klasik test teorisi, tek boyutlu madde tepki kuramı ve çok boyutlu madde tepki kuramı çerçevesinde incelenmesi.* Yayımlanmamış doktora tezi, Mersin Üniversitesi Eğitim Bilimleri Enstitüsü.

Turhan, A. (2006) *Multilevel 2PL item response model vertical equating with the presence of differential item functioning.* Unpublished doctoral dissertation. Florida State University.

Way, W, D., Ansley, T. N., & Forsyth, R. A. (1988). Unidimensional IRT estimates the comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement, 12*(3), 239-252.

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31(2),* 83-105.

Yavuz, G. (2014). *Çok boyutlu madde tepki kuramı modelleri ve paket programları için karşılaştırmalı analizler.* Yayımlanmamış doktora tezi, Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü.

Zhang, J. (2012). Calibration of response data using MIRT models with simple and mixed structures. *Applied Psychological Measurement, 36*(5), 375-398.

EXTENDED ABSTRACT

Introduction

Through item response theory (IRT), a relationship between the performance of an examinee in the test and the observed ability or trait of that examinee which underlie this performance is presented. The relation between these observable and unobservable traits is explained by means of mathematical functions. Therefore, item response models are mathematical models. These models are based on rather significant assumptions (Hambleton and Swamington, 1985; Embretson and Reise, 2000). Assumptions of IRT are assessed under unidimensionality and local independence titles. Even though there are different models for different conditions in IRT, these models are not suitable for usage if the assumptions are not addressed. In terms of application, the situation is influenced due to the fact that tests that have been used are not unidimensional. Therefore, unidimensionality, one of the most fundamental assumptions of IRT, is violated. Therefore, in these and other similar cases; multidimensional item response theory (MIRT) shall be used instead of IRT; as MIRT is the expanded version of IRT which has been developed for multidimensional data sets.
Various models that are used in MIRT; assumes various statistical relationships between ability and probability for correct response. Additionally, the relation between dimensions and items can be defined according to the structure of the data set. The first structure is referred as "simple structure". A simple structured test includes multiple sub-tests. Each of these sub-tests measure a single ability or trait and has a homogenous dispersion. Each item is related to a single trait which is tried to be measured with the sub-test. Complex structure is the second type of structure than can be considered for a MIRT data set. Complex structure, which is also referred as a non-simple structure, is also known as within item model. The test itself, as well as the items of the test, is related to multiple abilities (Sheng and Wikle, 2007; Bulut, 2013).
IRT often provides better results than the classical test theory despite its applicability in the same fields (Hambleton and Jones, 1993; Harvey and Hammer, 1999; Sinar and Zickar, 2002; Macdonald and Paunonen, 2002). MIRT; fundamentals of which have been set forth in 1980s (Sympson, 1978; Doody-Bogan and Hattie, 1981; Yen, 1983; akt:1978 Ansley and Forsyth, 1985); have been addressed by various studies that have been performed outside Turkey. On the other hand, number of studies which have been performed in Turkey on MIRT is rather low. There are a number of doctorate dissertation studies which have been performed on the subject (Köse, 2010; Sünbül, 2011; Özkan, 2012, Koğar, 2014, Yavuz, 2014). These studies include either simple or complex structures. None of them include both simple and complex test structures with 2- and 3-dimensional tests. Accordingly; the study is structured to seek the answer for the problem of "What is the influence of sample size and test length on item and ability parameter estimations of simple and complex structured two and three dimensional tests that have been performed by means of multidimensional item response theory?"

Method

The data which is required for the study is acquired via simulation. SimuMIRT (2003) software has been used for generating data. Multidimensional structures have been considered as constants under the scope of the study. Therefore constants are as: two dimensional complex, two dimensional simple, three dimensional complex and three dimensional simple structures. Test length (12 and 48) and sample size (1000, 2000, 4000) are independent variables of the study. At the end a test design consists of 24 conditions (4 MIRT models * 2 test length * 3 sample size) is obtained. For each condition 25 replication is used. In total 600 data sets (24 conditons* 25 reps) are analyzed by IRTPRO 2.1. Following the data analysis; RMSE and bias statistics have been calculated by using estimated and actual values of item and ability parameters. In additon Pearson Product-Moment Correlation have been assessed to evaluate the relation between actual parameters and estimated parameters.

Results and Discussion

According to the results of the study; increase of the sample size has caused decrease in RMSE values of item parameters in both two and three dimensional tests. This finding is in compliance with results of the studies performed by Bolt and Lall (2003), Finch (2010) and Zhang (2012). A regular increase or decrease of bias, which is directly related to increase of item number or sample size, has

not been observed. But bias values showed that; the increase in bias is directly related to number of dimensions and complexity of the structure. This finding is in compliance with the assumptions of Lee (2012). Mean correlation values that have been acquired for item parameters have been measured to be higher for a1 and a2 parameters in simple structured tests; and for d parameter in complex structured tests. Increase in number of examinees haven't had a significant influence on estimation quality of ability parameters (thetas) when number of items is kept the same; and haven't caused a significant difference in terms of RMSE values under all dimensional conditions. Accordingly; mean correlation between actual teta parameters and estimated teta parameters hasn't been influenced by the increase of population. On the other hand; estimation accuracy of ability parameters has also shown a difference when test length is changed. In this context; RMSE values have decreased significantly in terms of estimation of ability parameters when 12-item tests have been expanded to 48 item tests. Kirisci, Hsu and Yu (2001), de la Torre and Patz (2005), Köse (2010) and Kieftenbeld and Natesan (2012) have indicated that; expansion of the sample size does not have any influence on estimation accuracy of ability parameters, yet, theta estimation success may be developed if the test length is increased. Under this perspective, it can be said that findings of this study are in compliance with the current literature.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2015; 6(2); 351-364

Comparison of Different Estimation Methods for Categorical and Ordinal Data in Confirmatory Factor Analysis*

Doğrulayıcı Faktör Analizinde Sınıflama ve Sıralama Düzeyindeki Veriler için Farklı Kestirim Yöntemlerinin Karşılaştırılması

Hakan KOĞAR ** Esin YILMAZ KOĞAR ***

Abstract
In confirmatory factor analysis (CFA), which is used quite often for scale development and adaptation studies, the selected estimation method, affects the results obtained from the data. Because of the selected estimation method, the model parameters and their standard errors, and the model data fit values may alter the results substantially. So that, the purpose of this research is to compare the performance of different estimation methods for CFA. Maximum likelihood (ML), unweighted least squares (ULS) and diagonally weighted least squares (DWLS) are used in this research as estimation methods. These methods are applied in data sets and regression coefficients and their standard errors, t values, fit indexes and iteration numbers obtained from these estimation methods are examined. As a result, ULS method can converge with the minimum number iterations and it seems to be the more accurate method for estimating the parameters.

Keywords: Confirmatory factor analysis, weighted least square, unweighted least square, diagonally weighted least square

Öz
Ölçek geliştirme ve uyarlama çalışmalarında oldukça sık kullanılan doğrulayıcı faktör analizinde (DFA), hangi kestirim yönteminin kullanılması gerektiğine doğru karar vermek, çalışmadan elde edilen sonuçları etkilemektedir. Çünkü kullanılan kestirim yöntemi, model parametreleri ve onların standart hataları ve uyum indeksi değerleri gibi sonuçlar üzerinde etkiye sahiptir. Bu nedenle bu araştırmada, DFA'da kullanılan farklı kestirim yöntemlerinin performanslarını karşılaştırmak amaçlanmıştır. Araştırmada, maksimum olabilirlik maximum likelihood – ML), ağırlıklandırılmamış en küçük kareler (unweighted least square – ULS) ve diyagonal en küçük kareler (diagonally weighted least squares – DWLS) kestirim yöntemleri kullanılmıştır. Sınıflama ve sıralama düzeyinde olan veri setlerine bu kestirim yöntemleri uygulanmış ve bu kestirim yöntemlerinden elde edilen regresyon katsayıları ve bu katsayıların standart hatalar, t değerleri, uyum indeksleri ve tekrar sayıları incelenmiştir. Araştırma sonucunda ULS tekniğinin tüm veri setlerinde en az sayıda tekrar ile parametreleri tahmin ettiği ve ilgili örneklem verisine ait parametreleri tahmin etmek için en uygun teknik olduğu belirlenmiştir.

Anahtar Kelimeler: Doğrulayıcı faktör analizi, ağırlıklandırılmış en küçük kareler, ağırlıklandırılmamış en küçük kareler, diyagonal ağırlıklandırılmış en küçük kareler

INTRODUCTION

The social sciences and behavioral sciences rather focus on the latent variables that are not directly visible, and it is attempted to take decisions on latent variables through these variables. The Structural Equation Modelling (SEM) is widely used to identify the relationship between observable variables and latent variables (Jöreskog & Sörbom, 1996a; Sammel, Ryan & Legler, 1997). SEM is a set of statistical methods that allows describing the relationship between one or more continuous or

* This study was presented in "International Congress on Education for the Future: Issues and Challenges (ICEFIC 2015)".
** Ass. Prof. Dr., Akdeniz Üniversity, Faculty of Education, Antalya, Turkey, e-posta: hkogar@gmail.com
*** Res. Asst., Hacettepe Üniversity, Faculty of Education, Ankara, Turkey, e-posta: esinyilmazz@gmail.com

Geliş Tarihi: 27.10.2015
Kabul Tarihi: 24.12.2015

categorical independent variables and one or more continuous or categorical dependent variables (Tabachnick & Fidell, 2007, p. 676). In other words, SEM is a comprehensive statistical approach for testing hypothese based on the relationship between observable variables and latent variables (Hoyle, 1995).

The Confirmatory Factor Analysis (CFA), a customized version of SEM, is often used for studies developing and adapting scales. Brown (2006) states that CFA can be used (1) to psychometrically evaluate measurements, (2) to validate the structure, (3) to test the effect of the method, and (4) to indicate invariance of test measurements. To perform a CFA on a data set, first a theory must exist or there must be a predetermined factor structure. Then, a model is created based on this information and the model is tested through the observed data set (Raykoy & Marcoulides, 2000, p. 95). In brief, the CFA model aims to assess whether data set supports the assumed relationship between a group of measured variables.

In social sciences, data are often collected by scoring multiple-choice items in two categories and ordinal items with three or more categories, but not by an equal-interval scale (Bandalos, 2014). Since there is a relationship of a < b < c < d between ordinal variables in Likert scales and a relationship of a < b between data that is scored in two categories, it is referred to as ordered categorical data while there is no order between coding in non-ordered categorical data (Cai, 2008). Bollen (1989) emphasizes that ordered categorical data are treated as continuous data due to technical limitations on measuring tools. However, this approach to categorical data would result in biased estimation of parameters and unfavourable standard errors (Babakus, Ferguson & Joreskog, 1997; DiStefano, 2002; Rigdon & Ferguson, 1991). And it is important to match between the assumptions underlying the statistical model and the empirical characteristics of the data to be analyzed (Flora & Curran, 2004).Therefore, different estimation methods have been developed according to the data structure.

Estimation Methods

A right decision on what estimation method to use for statistical analyses has a direct influence on results derived from a study. The most common estimation method in SEM is the maximum likelihood (ML) method because it is selected in default in many software packages. This method is capable to make consistent and unbiased estimations on properly defined models, large sample sizes, normally distributed independent, continuous and multivariate data sets (Kline, 2005). It was deduced in the literature that the use of ML as an estimation method particularly for non-normally distributed data sets with a few number of answer categories resulted in bias in factor loadings, standard errors, statistics for chi-square test, and goodness of fit indexes (Babakus et al., 1987; Bollen, 1989; Green, Akey, Fleming, Hershberger, & Marquis, 1997; Hutchinson & Olmos, 1998). However, ML would not give significantly biased results if the number of categories of ordered data is high, the size of the sample is large, and the observed items are almost distributed normally (Mîndrilă, 2010).

For ordered data, it is assumed that there is a continuous variable such as y_i^* under a variable measured sequentially, such as y_i, in factor analysis, and this continuous variable ranging from –∞ to +∞ indicates characteristics underlying responses at y_i^* order level (Forero, Maydeu-Olivares & Gallardo-Pujol, 2009; Jöreskog, 1990; Lee, Poon & Bentler, 1990; Muthén, 1984). Tetrachoric or polychoric correlations should be used in SEM when categorical and/or orded data are used to determine the underlying continuous latent variable (Muthén, 1984; Muthén & Kaplan, 1985). It is because a high level of bias occurs in the estimation of parameters, standard errors, and factor loads based on Pearson Product-Moment Correlation Coefficient (PCC) (DiStefano, 2002). So to solve this problem, Jöreskog and Sörbom (1996b) suggested that polychoric correlations as the most consistent and robust estimator. The bias of estimation of parameters, standard errors, and factor loads is reduced by using polychoric correlation instead of PCC matrix (Babakus et al., 1997; Rigdon &

Ferguson, 1991). While tetrachoric correlation is used for data with two categories, polychoric correlation is used for data with more than two categories.

Some of estimation methods developed for ordinal data are weighted least square (WLS), unweighted least square (ULS) and diagonally weighted least squares (DWLS). Babakus (1985) suggests using polychoric correlation for performing CFA on ordinal data instead of Pearson correlation. In all of these methods, asymptotic covariance matrix is used that is derived from polychoric correlation matrix estimated from the observed categorical variables (Katsikatsou, Moustaki, Yang-Wallentin, & Jöreskog, 2012). The weight matrix of ULS and DWLS is the diagonal form of asymptotic covariance matrix; and the weight matrix of WLS is the reverse of asymptotic covariance matrix. Only diagonal elements of asymptotic covariance matrix are used for DWLS (Yang-Wallentin, Jöreskog, & Luo, 2010).

WLS can be an alternative method for ordinal data in particular that is not distributed normally, highly skewed or kurtic, or both (Muthén, 1993). However, WLS estimation converges to asymptotic features very slowly, therefore its performance on a small sample size is not good (Katsikatsou et al., 2012). ULS, on the other hand, has some features such as lack of distributional assumption and ability to estimate all parameters at a time. But this method requires all observed variables to have a same level of scale (Kline, 2005; 159). Recently, the use of DWLS estimation method has become popular for factor analysis of ordinal data. This popularity is attributed to ability of using DWLS as a method to determine measurement invariance in case of using continuous variables and to ability of DWLS to estimate variances smaller than ULS can do (Forero et al., 2009).

Katsikatsou et al. (2012) point out that the DWLS and ULS estimation methods are more preferable than WLS and both of these methods display a similar performance in small sample sizes. Likewise, Yang-Wallentin et al. (2010) established in their simulation study that WLS performs poorly under symmetric and non- symmetric 2, 5 and 7 categories conditions compared to ULS, DWLS, and ML and ULS had a remarkable performance. Forero et al. (2009) compared ULS and DWLS methods in their simulation study. They found that ULS estimated parameters more accurately and displayed less variability as well as showing more accurate standard error values and better convergence. Mîndrilă (2010) compared estimation methods DWLS and ML. They found that ML estimated parameters more accurately in continuous and normally distributed data and DWLS estimated parameters more accurately in data sets not normally distributed. The literature has many studies that compared different estimation methods which have effects on results obtained (DiStefano, 2002; Forero et al, 2009; Hu, Bentler & Kano, 1992; Lei, 2009; Muthen & Kaplan, 1985; Rigdon & Ferguson, 1991; Yang-Wallentin et al., 2010). However, there are only limited number of studies performed on real data sets (Katsikatsou et al., 2012). So this study is focused on real data sets.

Taking a correct decision on what estimation method to use when performing a CFA as in all other statistical analyses influences the results obtained from the study because the estimation method used has an influence on model parameters and their standard error values and fit index. When performing a CFA, ML method is widely used which is utilized for continuous variables and assumes that observable variables have a multivariate normal distribution. However, most of variables used for social sciences and psychology are not continuous but categorical / ordinal (Yang-Wallentin et al., 2010), and it is not appropriate to use methods developed for continuous variables regardless of data structure for accuracy of results.

Purpose of the Study

The overall objective of this research was to compare performances of different estimation methods used for CFA. For this, the following research questions were addressed:

In ordinal set-1 and set-2, and categorical set-3,

1. What are regression coefficients obtained from ML, ULS and DWLS estimation methods and their standard errors?

2. What are t values obtained from ML, ULS and DWLS estimation methods?

3. What are fit indexes obtained from ML, ULS and DWLS estimation methods?

4. What is the number of iterations that ML, ULS and DWLS estimation methods do for convergence (achieving parameter estimation)?

METHOD

This is a theoretical research which compared different estimation methods for CFA:

Data Collection Tools

Ordinal variables used for this study were obtained from first and fourth year students of primary school teaching in 2007-2008 academic year at seven different universities in seven regions of Turkey using Epistemological Belief Scale. The total number of participants was 548. This instrument was developed by Schommer (1990), adapted by Deryakulu and Büyüköztürk (2002) and revised by Deryakulu and Büyüköztürk (2005). The scale was a three-factor scale including "Belief that Learning Depends on Effort" (17 items), "Belief that Learning Depends on Skills" (8 items) and "Belief that there is only one correct" (9 items), and 34 items in total. The respond category was a five-point Likert scale.

Beck Hopelessness Scale was used for categorical data set. The data set obtained by Dinler-İçöz (2014) from 200 children for their master thesis was used for this study with their permission. Beck Hopelessness Scale contains 20 items. Questions 1, 3, 5, 6, 8, 10, 12, 13, 15 and 19 have 1 point each for the response "no" and the questions 2, 4, 7, 9, 11, 14, 16, 17, 18 and 20 have 1 point each for the response "yes". They are instructed to choose "correct" for statements they consider suitable, and "wrong" for statements they consider unsuitable. The score from the scale can vary from 0 to 20, and a high score indicates hopelessness whereas a low score indicates hope in children (Savaşır & Şahin, 1997).

Data Analysis

To compare sequential findings obtained from different sample sizes, two different data sets were created from a sample size of 548 persons as 250 (set-1) and 500 (set-2) randomly selected. Univariate and multivariate test of normality was performed on each data set. For this, first univariate normality assumption was tested for each item. West, Finch and Curran (1995) stated that if skewness value was greater than 2 and kurtosis value was greater than 7, they impaired univariate normality assumption of items. In the present study, the absolute value of skewness values for the data set of 250 persons ranged from 0.042 to 6.936, and kurtosis values ranged from 0.123 to 14.489. The skewness value of 26 items on the scale was greater than 2, the kurtosis value of 9 items was greater than 7, and all p values of chi-square values obtained had a significance level of 0.01 which was significant. For other data set, i.e., data set of 500 persons, the absolute value of skewness values ranged from 0.004 to 10.302 and kurtosis values ranged from 0.388 – 24.470. The skewness value of 28 items on the scale was greater than 2, kurtosis value of 2 items was greater than 7, and all p values of chi-square values obtained had a significance level of 0.001 which was significant. Based on these findings, not all of the items in both data sets displayed univariate normal distribution.

Mardia (1970) multivariate normality test results are listed in Table 1.

Table 1. The Findings of Multivariate Statistical Tests

Data Sets	Skewness			Kurtosis			Skewness and Kurtosis		
	Chi-square	Z score	p	Chi-square	Z score	p	Chi-square	Z score	p
Set-1	257,50	26,07	0,000	1375,94	14,15	0,000	879,71	0,000	257,50
Set-2	155,40	39,36	0,000	1417,09	22,32	0,000	2047,36	0,000	155,40

According to Table 1, two ordinal data sets in this study were determined that not multivariate normality.

The model must be identified or over identified to allow estimation of parameters in CFA (Brown, 2006; Kline, 2005). The model is referred to as unidentified if the number of unknown parameters is higher than available information in the model, as fully identified if the number of unknown parameters is equal to available information, and as over identified if the number of unknown parameters is less than available information (Brown, 2006). T-rule (Byrne, 1998) was used to identify the model. T-rule tests whether there is adequate degree of freedom to calculate and compare fit indexes, and there is adequate information for required estimation of parameters (Byrne, 1998).

Table 2. Model Identification with T Rules

Variable	Available Information	Estimation Information	Degree of Freedom
Ordinal (Set-1 and Set-2)	595*	71 Error: 34 Regression: 34 Covariance of latent variable: 3	524**
Categorical (Set-3)	210*	54 Error: 20 Regression: 34	156**

** v(v+1)/2 v: number of items*

*** degree of freedom, difference between available information and estimation information.*

According to Table 2, ordered and categorical data sets are over identified and so, CFA can be applied these data sets.

For categorical and ordinal data sets, unweighted least square and DWLS were used as estimation method in CFA. For this, an asymptotic covariance matrix was derived from all data sets. Regression values and their standard error values, t values and fit indexes were obtained and compared with this matrix. If the fit index values are $\chi 2/df<3$, $0<RMSEA<0.05$, $0.97\leq NFI\leq 1$, $0.97\leq CFI\leq 1$, $0.95\leq GFI\leq 1$ and $0.95\leq AGFI\leq 1$, this indicates a perfect fit, if they are $4<\chi 2/df<5$, $0,05<RMSEA<0.08$, $0.95\leq NFI\leq 0.97$, $0.95\leq CFI\leq 0.97$, $0.90\leq GFI\leq 0.95$ and $0.90\leq AGFI\leq 0.95$, this indicates an acceptable fit (Kline, 2005; Sümer, 2000). The method “weighted least squares" was excluded because there were non-positive elements in the asymptotic covariance matrix and a high sample size ($\geq$1000) was not achieved in this study, which is necessary for this method to estimate parameters (Hoogland & Boomsma, 1998). In addition, parameters were estimated by maximum likelihood method even though data sets were not distributed normally.

FINDINGS

The regression coefficients and standard error values which obtained from ML, ULS and DWLS estimation methods are shown in Table 3 and Table 4.

Table 3. Regression Coefficients and Standard Error Values (Set-1 and Set-2)

Items	Set-1			Set-2		
	ML (SE)	ULS (SE)	DWLS (SE)	ML (SE)	ULS (SE)	DWLS (SE)
I1	0,72 (0,07)	0,67 (0,05)	0,67 (0,05)	142,29 (7,88)	0,73 (0,03)	0,74 (0,03)
I2	0,69 (0,06)	0,65 (0,05)	0,66 (0,05)	0,79 (0,04)	0,72 (0,03)	0,72 (0,03)
I3	1,09 (0,07)	0,83 (0,04)	0,83 (0,04)	1,10 (0,05)	0,86 (0,02)	0,86 (0,02)
I4	57,24 (5,82)	0,56 (0,07)	0,57 (0,07)	20,55 (1,36)	0,60 (0,04)	0,61 (0,04)
I5	36,33 (2,75)	0,72 (0,05)	0,73 (0,05)	74,51 (3,87)	0,74 (0,03)	0,74 (0,03)
I6	12,72 (1,37)	0,56 (0,06)	0,56 (0,07)	13,35 (1,00)	0,56 (0,05)	0,56 (0,05)
I7	0,66 (0,06)	0,59 (0,06)	0,60 (0,06)	0,63 (0,04)	0,62 (0,04)	0,62 (0,04)
I8	0,47 (0,06)	0,45 (0,06)	0,45 (0,06)	0,43 (0,04)	0,45 (0,05)	0,45 (0,05)
I9	23,62 (1,78)	0,73 (0,04)	0,73 (0,04)	28,05 (1,52)	0,71 (0,03)	0,72 (0,03)
I10	0,42 (0,06)	0,47 (0,07)	0,47 (0,07)	0,60 (0,06)	0,47 (0,05)	0,48 (0,05)
I11	74,38 (5,46)	0,75 (0,04)	0,75 (0,04)	73,32 (3,90)	0,73 (0,03)	0,74 (0,03)
I12	1,01 (0,06)	0,86 (0,03)	0,86 (0,03)	80,03 (3,74)	0,82 (0,03)	0,82 (0,03)
I13	0,60 (0,06)	0,60 (0,05)	0,61 (0,05)	0,63 (0,04)	0,62 (0,04)	0,63 (0,04)
I14	0,67 (0,06)	0,62 (0,06)	0,64 (0,06)	0,62 (0,04)	0,65 (0,04)	0,65 (0,04)
I15	25,64 (1,78)	0,78 (0,05)	0,78 (0,05)	15,55 (0,87)	0,72 (0,04)	0,72 (0,04)
I16	0,60 (0,07)	0,51 (0,06)	0,52 (0,06)	0,62 (0,06)	0,46 (0,04)	0,47 (0,04)
I17	21,89 (1,85)	0,68 (0,05)	0,69 (0,05)	24,24 (1,62)	0,63 (0,04)	0,63 (0,04)
I18	0,85 (0,17)	0,62 (0,09)	0,64 (0,09)	0,49 (0,09)	0,62 (0,08)	0,62 (0,07)
I19	0,52 (0,06)	0,56 (0,07)	0,56 (0,07)	1,03 (0,10)	0,44 (0,05)	0,45 (0,05)
I20	0,47 (0,05)	0,51 (0,07)	0,51 (0,07)	0,88 (0,06)	0,58 (0,05)	0,60 (0,05)
I21	0,41 (0,04)	0,54 (0,08)	0,55 (0,08)	0,81 (0,07)	0,31 (0,06)	0,35 (0,06)
I22	0,50 (0,05)	0,52 (0,08)	0,53 (0,08)	1,00 (0,06)	0,66 (0,05)	0,69 (0,05)
I23	0,40 (0,04)	0,45 (0,09)	0,46 (0,09)	0,76 (0,06)	0,48 (0,06)	0,51 (0,05)
I24	0,51 (0,06)	0,50 (0,07)	0,50 (0,07)	1,09 (0,08)	0,53 (0,06)	0,53 (0,05)
I25	0,46 (0,04)	0,65 (0,06)	0,66 (0,06)	0,72 (0,05)	0,61 (0,05)	0,61 (0,05)
I26	0,73 (0,19)	0,39 (0,06)	0,40 (0,07)	0,68 (0,12)	0,43 (0,07)	0,44 (0,09)
I27	0,71 (0,14)	0,41 (0,20)	0,43 (0,23)	0,66 (0,09)	0,39 (0,28)	0,38 (0,26)
I28	0,64 (0,08)	0,16 (0,18)	0,17 (0,22)	0,61 (0,05)	0,11 (0,22)	0,10 (0,20)
I29	1,05 (0,13)	0,02 (0,15)	0,02 (0,17)	0,88 (0,08)	0,06 (0,14)	0,07 (0,12)
I30	0,44 (0,07)	0,22 (0,19)	0,24 (0,20)	0,47 (0,05)	0,22 (0,24)	0,21 (0,22)
I31	1,11 (0,15)	0,25 (0,22)	0,27 (0,23)	0,94 (0,09)	0,29 (0,27)	0,28 (0,25)
I32	0,57 (0,09)	0,07 (0,12)	0,06 (0,15)	0,64 (0,07)	0,02 (0,16)	0,02 (0,15)
I33	0,77 (0,11)	0,40 (0,25)	0,42 (0,27)	0,82 (0,07)	0,40 (0,33)	0,39 (0,30)
I34	0,85 (0,12)	0,23 (0,20)	0,23 (0,21)	0,89 (0,08)	0,15 (0,21)	0,14 (0,19)

In Table 3, regression coefficient for set-1 ranged from 0.40 to 74.38 in estimations made by maximum likelihood (ML) and from 0.02 to 0.86 in estimations made by unweighted least squares (ULS) and diagonally weighted least squares (DWLS). More stable regression coefficients were achieved in methods ULS and DWLS. In evaluation of standard errors, standard error values varied between 0.04 and 5.82 in estimations made by ML; between 0.03 and 0.25 in ULS method, and between 0.03 and 0.27 in DWLS method. The mean of standard error values was 0.677 for ML, 0.090 for ULS and 0.096 for DWLS. The ULS method estimated parameters with least error for set-1. This finding is similar to that in the study by Forero et al. (2009).

Regression coefficient for set-2 ranged from 0.43 – 142.29 in estimations made by ML, from 0.02 to 0.86 in estimations made by ULS and DLWS. In evaluation of standard error values, standard error

ranged between 0.04 and 7.88 in estimations made by ML, between 0.02 and 0.33 in ULS method, and between 0.03 and 0.30 in DWLS method. The mean of standard error values was 0.806 for ML, 0.089 for ULS and 0.084 for DWLS. The DWLS method estimated parameters with least error for set-2. This finding is similar to that in the study by Mîndrilă (2010).

It appears that similar regression coefficients and standard error values were obtained from ULS and DWLS methods. Katsikatsou et al. (2012) and Yang-Wallentin et al. (2010) found similar findings in their study. As sample size increased, similar findings were obtained in standard error values of parameter estimates from the ULS method whereas standard error values were lower in parameter estimates from DWLS. ULS and DWLS methods estimated parameters with less error as compared to the ML method.

Table 4. Regression Coefficients and Standard Error Values (Set-3)

Items	Set-3		
	ML (SE)	ULS (SE)	DWLS (SE)
I1	0,72 (0,10)	0,72 (0,08)	0,73 (0,07)
I2	0,57 (0,10)	0,57 (0,09)	0,57 (0,10)
I3	0,26 (0,11)	0,28 (0,12)	0,31 (0,12)
I4	0,38 (0,11)	0,38 (0,10)	0,38 (0,11)
I5	0,61 (0,10)	0,62 (0,08)	0,61 (0,08)
I6	0,89 (0,10)	0,88 (0,05)	0,90 (0,05)
I7	0,89 (0,10)	0,89 (0,05)	0,92 (0,05)
I8	0,49 (0,10)	0,48 (0,09)	0,49 (0,10)
I9	0,29 (0,05)	0,62 (0,09)	0,62 (0,10)
I10	0,37 (0,11)	0,36 (0,12)	0,38 (0,13)
I11	0,83 (0,10)	0,83 (0,06)	0,83 (0,06)
I12	0,52 (0,10)	0,51 (0,09)	0,54 (0,09)
I13	0,71 (0,10)	0,71 (0,07)	0,72 (0,08)
I14	0,72 (0,10)	0,72 (0,06)	0,75 (0,06)
I15	0,88 (0,10)	0,88 (0,05)	0,88 (0,06)
I16	0,52 (0,10)	0,53 (0,10)	0,55 (0,10)
I17	0,92 (0,09)	0,92 (0,04)	0,94 (0,04)
I18	0,87 (0,10)	0,87 (0,06)	0,87 (0,06)
I19	0,15 (0,03)	0,55 (0,20)	0,59 (0,13)
I20	0,57 (0,08)	0,69 (0,11)	0,70 (0,10)

In Table 4, regression coefficient for set-3 ranged from 0.15 to 0.92 in estimations mad by ML, from 0.28 to 0.92 in estimations made by ULS method, and from 0.31 to 0.94 in estimations made by DWLS. In evaluation of standard errors, standard error values were between 0.03 – 0.11 in estimations made by ML, between 0.05 and 0.20 in ULS method, and between 0.04 – 0.13 in DWLS method. The mean of standard error values was 0.094 for ML, 0.086 for ULS, and 0.085 for DWLS. The DWLS method estimated parameters with least error for set-3. Although similar regression coefficients were obtained from all parameter estimation methods, it appears that the ULS and DWLS methods estimated parameters with less error as compared to the ML method.

The t values which obtained from ML, ULS and DWLS estimation methods are shown in Table 5.

Table 5. t Values for Regression Coefficients

	Set-1			Set-2			Set-3		
Items	ML	ULS	DWLS	ML	ULS	DWLS	ML	ULS	DWLS
I1	10,29	13,40	13,40	18,06	24,33	24,67	7,20	9,00	10,43
I2	11,50	13,00	13,20	19,75	24,00	24,00	5,70	6,33	5,70
I3	15,57	20,75	20,75	22,00	43,00	43,00	2,36	2,33	2,58
I4	9,84	8,00	8,14	15,11	15,00	15,25	3,45	3,80	3,45
I5	13,21	14,40	14,60	19,25	24,67	24,67	6,10	7,75	7,63
I6	9,28	9,33	8,00	13,35	11,20	11,20	8,90	17,60	18,00
I7	11,00	9,83	10,00	15,75	15,50	15,50	8,90	17,80	18,40
I8	7,83	7,50	7,50	10,75	9,00	9,00	4,90	5,33	4,90
I9	13,27	18,25	18,25	18,45	23,67	24,00	5,80	6,89	6,20
I10	7,00	6,71	6,71	10,00	9,40	9,60	3,36	3,00	2,92
I11	13,62	18,75	18,75	18,80	24,33	24,67	8,30	13,83	13,83
I12	16,83	28,67	28,67	21,40	27,33	27,33	5,20	5,67	6,00
I13	10,00	12,00	12,20	15,75	15,50	15,75	7,10	10,14	9,00
I14	11,17	10,33	10,67	15,50	16,25	16,25	7,20	12,00	12,50
I15	14,40	15,60	15,60	17,87	18,00	18,00	8,80	17,60	14,67
I16	8,57	8,50	8,67	10,33	11,50	11,75	5,20	5,30	5,50
I17	11,83	13,60	13,80	14,96	15,75	15,75	10,22	23,00	23,50
I18	5,00	6,89	7,11	5,44	7,75	8,86	8,70	14,50	14,50
I19	8,67	8,00	8,00	10,30	8,80	9,00	5,00	2,75	4,54
I20	9,40	7,29	7,29	14,67	11,60	12,00	7,13	6,27	7,00
I21	10,25	6,75	6,88	11,57	5,17	5,83	-	-	-
I22	10,00	6,50	6,63	16,67	13,20	13,80	-	-	-
I23	10,00	5,00	5,11	12,67	8,00	10,20	-	-	-
I24	8,50	7,14	7,14	13,63	8,83	10,60	-	-	-
I25	11,50	10,83	11,00	14,40	12,20	12,20	-	-	-
I26	3,84	6,50	5,71	5,67	6,14	4,89	-	-	-
I27	5,07	2,05	1,87	7,33	1,39	1,46	-	-	-
I28	8,00	0,89	0,77	12,20	0,50	0,50	-	-	-
I29	8,08	0,13	0,12	11,00	0,43	0,58	-	-	-
I30	6,29	1,16	1,20	9,40	0,92	0,95	-	-	-
I31	7,40	1,14	1,17	10,44	1,07	1,12	-	-	-
I32	6,33	0,58	0,40	9,14	0,13	0,13	-	-	-
I33	7,00	1,60	1,56	11,71	1,21	1,30	-	-	-
I34	7,08	1,15	1,10	11,13	0,71	0,74	-	-	-

In evaluation of t values for regression coefficients, t values for set-1 ranged between 3.84 and 16.83 in estimations made by ML; between 0.13 and 28.67 in estimations made by ULS, and between 0.12 and 28.67 in estimations made by DWLS. The mean of these values was 9.64 for ML, 8.89 for ULS and 8.88 for DWLS. T values for set-2 ranged from 5.44 to 22.00 in estimations made by ML, from 0.43 to 43.00 in ULS method, and from 0.50 to 43.00 in estimations made by DWLS. The mean of these values was 13.66 for ML, 12.25 for ULS and 12.49 for DWLS.

For set-1 and set-2, higher t values were obtained in parameter estimations made by ML. Although standard error values were lower in estimations made by ULS and DWLS, the reason why estimations made by ML had a higher t value was to obtain very low regression values in estimations made by ULS and DWLS for several items.

For set-3, t values ranged between 2.36 and 10.22 in estimations made by ML, between 2.33 and 23.00 in estimations made by ULS method, and between 2.58 and 23.50 in estimations made by DWLS. The mean of these values was 6.48 for ML, 9.54 for ULS and 9.56 for DWLS. For set-3, higher t values were obtained in parameter estimations made by ULS and DWLS.

Fit indexes results which obtained from ML, ULS and DWLS estimation methods are shown in Table 6.

Table 6. Fit Index Values

		Set-1	Set-2	Set-3
ML	χ^2/df	2,91	4,32	1,04
	RMSEA	0,09	0,08	0,01
	CFI	0,88	0,91	1,00
	NFI	0,84	0,89	0,91
	GFI	0,74	0,79	0,92
	AGFI	0,70	0,76	0,90
ULS	χ^2/df	2,00	3,78	1,26
	RMSEA	0,06	0,08	0,04
	CFI	0,95	0,93	0,99
	NFI	0,90	0,91	0,97
	GFI	0,92	0,90	0,96
	AGFI	0,91	0,88	0,95
DWLS	χ^2/df	1,98	3,76	1,30
	RMSEA	0,06	0,07	0,04
	CFI	0,95	0,93	0,99
	NFI	0,90	0,91	0,97
	GFI	0,93	0,91	0,98
	AGFI	0,92	0,90	0,97

In evaluation of fit index values, for set-1 and set-2, by comparing with other methods, data sets are more fit to DWLS method. On the other hand, the fit indices were obtained by ULS method is higher than ML method. Fit index values were very similar to each other that were obtained from estimations made by ULS and DWLS. Katsikatsou et al. (2012) and Yang-Wallentin elt al (2010) obtained similar findings from their study. The model created with estimations made by ML fell below the acceptable level.

For set-3, estimations made by ML had higher match in fit indexes χ2/df and RMSEA, and estimations made by ULS and DWLS methods had higher match in fit indexes NFI, GFI and AGFI. Similar values were obtained from all three parameter methods in fit index CFI.

The iteration numbers of ML, ULS and DWLS estimation methods for convergence are shown in Table 7.

Table 7. Iteration Numbers

	Set-1	Set-2	Set-3
ML	28	20	8
ULS	20	18	5
DWLS	24	29	9

To have standard error values at an acceptable level, estimation of parameters continues iteratively. When iteration ends, this means that no significant changes will occur in estimations of parameters. If there is only few number of iterations, this means that relevant parameter estimation method better matches with the data set of sample. The reason is that different parameter estimation methods have different distributional assumptions (Marsh & Grayson, 1995). In evaluation of number of iterations, ULS method estimated parameters with the least number of iterations in all data sets. This finding

was similar to that in the study by Forero et al. (2009). For set-1, ML method estimated parameters with the highest number of iterations, and for set-2 and set-3 DWLS method estimated parameters with the highest number of iterations. This finding shows that the estimation method which displayed optimum match with relevant data sets was ULS.

CONCLUSIONS and DISCUSSION

It is found that ULS and DWLS methods estimated parameters with less standard errors in ordinal data sets comparing to ML method. The ULS method yielded better results in lower sample sizes whereas DWLS method yielded better results in higher sample sizes. In estimation of parameters made by ULS and DWLS, fit values for the CFA model created were higher and values obtained were acceptable levels. Even in lower sample sizes, estimation made by ULS and DWLS methods achieved higher fit index values. The model crated with estimations made by ML method fell below an acceptable level. In ordinal data sets it is therefore revealed that fit index values for estimations made by ML method did not have adequate levels in univariate and multivariate Likert-type response patterns not normally distributed. As the size of the sample was increased, fit indexes were somewhat improved but fit indexes RMSE, CFI and GFI did not show the expected level. In estimations made by ULS and DWLS methods, as the size of sample was increased, fit index values were somewhat reduced, and all fit indexes had expected values except for AGFI obtained from estimations made by ULS for set-2.

In categorical data sets, it is found that the ULS and DWLS methods estimated parameters with less standard error as compared to the ML method; and the method which estimated parameters with the least error was DWLS. Since the fit index AGFI give fit values that are adjusted based on the degree of freedom for the model created by the number of variables when different models are applied to the same data set, it is more suitable value used to compare fit indexes obtained from all methods (Mîndrilă, 2010). The fit index AGFI was demonstrated to have a higher value in ULS and DWLS methods than that of ML method.

As provided in findings, the ULS method estimated parameters with the least number of iterations in all data sets, and it is the more accurate method to estimate parameters of relevant sample data.

In future research, the researches will examine the performance of WLS method with larger sample sizes (≥1000). And large number of categories can be used real data analyses for examine the ML method result. Because categorical methodology can outperform continuous methodology with more than five categories (Beauducel & Herzberg, 2006).

REFERENCES

Babakus, E. (1985). *The sensitivity of maximum likelihood factor analysis given violations of interval scale and multivariate normality* (Doctoral dissertation, The University of Alabama).

Babakus, E., Ferguson, C. E., & Joreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research, 37*, 72-141.

Bandalos, D. L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling: A Multidisciplinary Journal, 21,* 102 – 116.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13,* 186 – 203.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York, NY: John Wiley & Sons.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York, N. J.: Guiford Press.

Byrne, B. M. (1998). *Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming.* Lawrence Erlbaum Associates, Mahwah, N. J.

Cai, J. (2008). *Structural equation modeling analysis with correlated ordered and unordered categorical data* (Doctorial dissertation, The Chinese University of Hong Kong).

Deryakulu, D. & Büyüköztürk, Ş. (2002). Epistemolojik inanç ölçeğinin geçerlik ve güvenirlik çalışması. *Eğitim Araştırmaları, 8,* 111– 125.

Deryakulu, D., & Büyüköztürk, Ş. (2005). Epistemolojik inanç ölçeğinin faktör yapısının yeniden incelenmesi: cinsiyet ve öğrenim görülen program türüne göre epistemolojik inançların karşılaştırılması. *Eğitim Araştırmaları, 18,* 57 – 70.

Dinler-İçöz, S. (2014). *İşitme engelli çocuğa sahip olan ve olmayan annelerin umutsuzluk düzeylerinin incelenmesi* (Yüksek Lisans Tezi, Ankara Üniversitesi, Fen Bilimleri Fakültesi, Ankara).

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9,* 327 – 346.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9,* 466 – 491.

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A monte carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling, 16,* 625 – 641.

Green, S. B., Akey, T. M., Fleming, K. K., Hershberger, S. L., & Marquis, J. G. (1997). Effect of the number of scale points on chi-square fit indices in confirmatory factor analysis. *Structural Equation Modeling, 4,* 108-120.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research, 26,* 329 – 367.

Hoyle, R. H. (1995). *Structural equation modeling concepts, issues, and applications.* Thousand Oaksa London New Delhi: Sage Publications.

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112,* 351 – 362.

Hutchinson, S. R., & Olmos, A. (1998). Behavior of descriptive fit indexes in confirmatory factor analysis using ordered categorical data. *Structural Equation Modeling, 5,* 344 – 364.

Jöreskog, K. G. (1990). New developments in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity, 24,* 387-404.

Jöreskog, K. G., & Sörbom, D. (1996a). *LISREL 8: Structural equation modelling with the SIMPLIS command language.* Hove and London: Scientific Software ınternational.

Jöreskog, K.G., & Sörbom, D. (1996b). *LISREL 8: User's reference guide.* Chicago: Scientific Software International.

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., & Joreskog, K. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics and Data Analysis, 56*(12), 4243-4258.

Kline, R. B. (2005). *Principles and practice of structural equation modeling (Second Edition).* New York: The Guilford Publications.

Lee, S. Y., Poon, W. Y., & Bentler, P. M. (1995). A two-stage estimation of structural equation models with continuous and polytomous variables. *British Journal of Mathematical and Statistical Psychology, 48,* 339 – 358.

Lei, P. (2009). Evaluating estimation methods for ordinal data in structural equation modeling. *Quality & Quantity, 43,* 495 – 507.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57,* 519 – 530.

Marsh, H. W., & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 177−198). Thousand Oaks, CA: Sage.

Mîndrilă, D. (2010). Maximum likelihood (ML) and diagonally weighted least squares (DWLS) estimation procedures: A comparison of estimation bias with ordinal and multivariate nonnormal data. *International Journal of Digital Society, 1*(1), 60 – 66.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika, 49,* 115 – 132.

Muthén, B. O. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–243). Newbury Park, CA: Sage.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38,* 171–189.

Raykov, T., & Marcoulides, G. A. (2000). *A first course in structural equation modeling*. London: Lawrence Erlbaum Associates, Inc.

Rigdon, E. E., & Ferguson, C. E. (1991). The performance of the polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. *Journal of Marketing Research, 28,* 491 – 497.

Sammel, M. D., Ryan, L. M., & Legler, J. M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society, Series B, 59,* 667 – 678.

Savaşır, I., & Şahin, N. H. (1997). *Bilişsel-davranışçı terapilerde değerlendirme: sık kullanılan ölçekler*. Ankara. Türk Psikologlar Derneği Yayınları.

Schommer, M. (1990). Effects of beliefs about the nature of knowledge on comprehension. *Journal of Educational Psychology, 82*(3), 498 – 504.

Sümer, N. (2000). *Yapısal eşitlik modelleri: Temel kavramlar ve örnek uygulamalar.* Türk Psikoloji Yazıları, 3(6), 49-74.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics (Fifth Edition).* Boston: Allyn and Bacon.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with non-normal variables. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 56–75). Thousand Oaks, CA: Sage.

Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling, 17,* 392 – 423.

GENİŞ ÖZET

Giriş

Ölçek geliştirme ve uyarlama çalışmalarında oldukça sık kullanılan doğrulayıcı faktör analizinde (DFA), hangi kestirim yönteminin kullanılması gerektiğine doğru karar vermek, çalışmadan elde edilen sonuçları etkilemektedir. Çünkü kullanılan kestirim yöntemi, model parametreleri ve onların standart hataları ve uyum indeksi değerleri gibi sonuçlar üzerinde etkiye sahiptir. DFA yaparken en çok sürekli değişkenler için kullanılan ve gözlenen değişkenlerin çok değişkenli normal dağılım gösterdiğini varsayan maksimum olabilirlik (maximum likelihood – ML) yöntemi kullanılmaktadır. Ancak sosyal bilimlerde ve psikolojide kullanılan değişkenlerin birçoğu sürekli değil sınıflama/sıralama düzeyindedir (Yang-Wallentin vd., 2010). Sıralama düzeyindeki veriler için geliştirilen kestirim yöntemlerinden bazıları ağırlıklandırılmış en küçük kareler (weighted least square – WLS), ağırlıklandırılmamış en küçük kareler (unweighted least square – ULS) ve diyagonal en küçük kareler (diagonally weighted least squares – DWLS) kestirim yöntemleridir. Bu üç yöntemde de gözlenen kategorik değişkenlerden kestirilen polikorik korelasyon matrisinden elde edilen asimptotik kovaryans matrisi kullanılmaktadır (Katsikatsou, Moustaki, Yang-Wallentin, & Jöreskog, 2012).

Bu araştırmanın genel amacı, DFA'da kullanılan farklı kestirim yöntemlerinin performanslarını karşılaştırmaktır. Bu amaçla çalışmada, aşağıdaki problemlere cevap aranmıştır.

Sıralama düzeyindeki birinci (set-1), ikinci (set-2) ve sınıflama düzeyindeki (set-3) veri setinde;

1. ML, ULS ve DWLS kestirim yöntemlerinden elde edilen regresyon katsayıları ve bu katsayıların standart hataları nasıldır?

2. ML, ULS ve DWLS kestirim yöntemlerinden elde edilen t değerleri nasıldır?

3. ML, ULS ve DWLS kestirim yöntemlerinden elde edilen uyum indeksleri nasıldır?

4. ML, ULS ve DWLS kestirim yöntemlerinin yakınsayabilmek (parametre kestirimlerine ulaşmak) için yaptıkları tekrar (iteration) sayıları nasıldır?

Yöntem

Doğrulayıcı faktör analizinde farklı kestirim tekniklerinin karşılaştırıldığı bu araştırma temel bir araştırma niteliğindedir.

Bu çalışmada kullanılan sıralama düzeyindeki veri setleri, 2007-2008 eğitim öğretim yılında Türkiye'nin yedi bölgesindeki yedi farklı üniversitede, birinci ve dördüncü sınıfta öğrenim gören sınıf öğretmenliği öğrencisine Epistemolojik İnanç Ölçeği'nin uygulanmasıyla elde edilmiştir. Sıralama düzeyinde farklı örneklem büyüklüklerinden elde edilen bulguları karşılaştırabilmek amacıyla 548 kişilik örneklem büyüklüğünden 250 (set-1) ve 500 (set-2) kişilik seçkisiz olarak seçilen iki farklı veri seti oluşturulmuştur. Sınıflama düzeyindeki veri seti için Beck Umutsuzluk Ölçeği kullanılmıştır. Dinler-İçöz'ün (2014) yüksek lisans tez çalışması için 200 çocuktan elde ettiği veri seti (set-3), izin alınarak bu çalışmada da kullanılmıştır. Bütün veri setleri için tek değişkenli ve çok değişkenli normallik testleri yapılmış olup, veri setlerinde yer alan tüm maddelerin tek ve çok değişkenli normal dağılım göstermediği belirlenmiştir.

Sonuçlar ve Tartışma

Sınıflama ve sıralama düzeyindeki veri setlerinden asimptotik kovaryans matrisi elde edilmiştir. Bu matrisler ile regresyon değerleri ve bu değerlere ait standart hata değerleri, t değerleri ile uyum indeksleri elde edilmiş ve karşılaştırılmıştır.

Set-1 için farklı kestirim teknikleriyle hesaplanan regresyon katsayılarına ait bulgulara göre ULS ve DLWS'nin ML'ye göre daha stabil sonuçlar verdiği belirlenmiştir (ML: 0,40 - 74,38; ULS ve DWLS: 0,02 – 0,86). Standart hata değerleri incelendiğinde ise, ML ile yapılan tahminlerde 0,04 - 5,82; ULS'de 0,03 – 0,25 ve DWLS'de 0,03 – 0,27 arasında standart hata değerleri elde edilmiştir. Standart hata değerlerinin ortalaması ML için 0,677; ULS için 0,090 ve DWLS için 0,096'dır. Set-1 için en az hata ile parametre tahminini ULS tekniği yapmaktadır. Bu bulgu, Forero vd.'nin (2009) çalışması ile benzerlik göstermektedir.

Set-2 için farklı kestirim teknikleriyle hesaplanan regresyon katsayıları incelendiğinde yine ULS ve DLWS'nin ML'ye göre daha stabil sonuçlar verdiği belirlenmiştir (ML: 0,43 - 142,29; ULS ve DLWS: 0,02 – 0,86). Standart hata değerleri incelendiğinde, ML ile yapılan tahminlerde 0,04 - 7,88; ULS'de 0,02 – 0,33 ve DLWS'de 0,03 – 0,30 arasında standart hata değerleri elde edilmiştir. Standart hata değerlerinin ortalaması, ML için 0,806; ULS için 0,089 ve DWLS için 0,084'dır. Set-2 için en az hata ile parametre tahminini DWLS tekniği yapmaktadır. Bu bulgu, Mîndrilă'nın (2010) çalışması ile benzerlik göstermektedir.

Set-3 için regresyon katsayılarının ML ile yapılan tahminlerde 0,15 - 0,92; ULS tekniğinde 0,28 – 0,92 ve DWLS ile yapılan tahminlerde 0,31 – 0,94 arasında değerler aldığı belirlenmiştir. Standart hata değerleri incelendiğinde, ML ile yapılan tahminlerde 0,03 - 0,11; ULS tekniğinde 0,05 – 0,20 ve DWLS tekniğinde 0,04 – 0,13 arasında standart hata değerleri elde edilmiştir. Standart hata değerlerinin ortalaması ML için 0,094; ULS için 0,09 ve DWLS için 0,09'dur. Set-3 için en az hata ile parametre tahminini DWLS tekniği yapmaktadır. Tüm parametre tahmin tekniklerinde benzer regresyon katsayıları elde edilmekle birlikte, ULS ve DWLS tekniklerinin ML tekniğine göre daha az hata ile parametre tahminleri yapabildiği belirlenmiştir.

Set-1 ve set-2 için ML ile parametre tahminlerinde daha yüksek t değerlerinin elde edildiği belirlenmiştir. Standart hata değerlerinin ULS ve DWLS tahminlerinde daha düşük olmasına karşın, ML tahminlerinin daha yüksek t değerine sahip olmasının nedeni, ULS ve DWLS tahminlerinde bazı maddeler için oldukça düşük regresyon değerlerinin elde edilmesidir. Set-3 için t değerlerinin ML ile yapılan tahminlerde 2,36 - 10,22; ULS tekniğinde 2,33 – 23,00 ve DWLS ile yapılan tahminlerde 2,58 – 23,50 arasında değerler aldığı belirlenmiştir. Bu değerlerin ortalaması, ML için 6,48; ULS için 9,54 ve DWLS için 9,56'dır. Set-3 için ULS ve DWLS ile parametre tahminlerinde daha yüksek t değerlerinin elde edildiği belirlenmiştir.

Set-1 ve set-2 için ULS ve DWLS tahminleri ile elde edilen uyum indeksi değerleri birbirine oldukça benzer olduğu, ML tahminlerinin ise örneklem büyüklüğündeki artış ile birlikte uyum indekslerinde bir miktar iyileşme olsa da RMSE, CFI ve GFI uyum indekslerinin beklenen düzeyi göstermediği belirlenmiştir. Set-3 için de ULS ve DWLS daha iyi sonuç vermiştir.

ML, ULS ve DWLS kestirim yöntemlerinin yakınsayabilmek (parametre kestirimlerine ulaşmak) için yaptıkları tekrar (iteration) sayıları incelendiğinde ise ULS tekniğinin tüm veri setlerinde en az sayıda tekrar ile parametreleri tahmin ettiği belirlenmiştir. Bu bulgu, Forero vd.'nin (2009) çalışması ile benzerlik göstermektedir. Set-1 için ML, set-2 ve set-3 için ise DWLS tekniğinin en fazla sayıda tekrar ile parametreleri tahmin ettiği görülmektedir. Bu bulgu, ilgili veri setleri ile en iyi uyumu gösteren tahmin tekniğinin ULS olduğunu göstermektedir.

Sonuç olarak ULS tekniğinin tüm veri setlerinde en az sayıda tekrar ile parametreleri tahmin ettiği ve ilgili örneklem verisine ait parametreleri tahmin etmek için en uygun teknik olduğu belirlenmiştir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2015; 6(2); 365-384

İçerik Ağırlıklandırmasının Maddeler-Arası Boyutluluk Modeline Dayalı Çok Boyutlu Bilgisayar Ortamında Bireyselleştirilmiş Test Yöntemleri Üzerindeki Etkisinin İncelenmesi

Examining the Effect of Content Balancing on Multidimensional Computerized Adaptive Testing Based on Between-Item Dimensionality Model

Burhanettin ÖZDEMİR ** Selahattin GELBAL ***

Öz

Bu çalışmanın amacı, maddeler-arası boyutluluk modeline dayalı Çok Boyutlu Bilgisayar Ortamında Bireyselleştirilmiş (BOB) Test Yöntemlerinin performanslarının karşılaştırılması ve içerik ağırlıklandırmasının (content balancing) çok boyutlu BOB testi yöntemleri üzerindeki etkisinin incelenmesidir. Bu amaç doğrultusunda, 2009-2013 eğitim ve öğretim yıllarında Hacettepe Üniversitesi tarafından uygulanan İngilizce Yeterlik Sınavına (İYS) ait gerçek veri seti kullanılmıştır. Her bir testte yer alan dinleme, dilbilgisi ve okuduğunu anlamaya ilişkin maddeler ile üç boyutlu gerçek madde havuzu oluşturulmuştur. Maddeler-arası boyutluluk modeli ile kalibre edilerek oluşturulan madde havuzu toplamda 555 maddeden oluşmaktadır. En uygun çok boyutlu BOB testini belirlemek amacıyla; iki farklı yetenek kestirim yöntemi (Bayesyen MAP ve Fisher'in puanlama yöntemi), üç farklı madde seçim yöntemi (A-optimality, D-optimality ve seçkisiz) ve hata varyansı durdurma kuralına dayalı üç farklı ölçüt kullanılmıştır. Ayrıca içerik ağırlıklandırmasının çok boyutlu BOB testi yöntemleri üzerindeki etkisini incelemek amacıyla, içerik ağırlıklandırmasının yapıldığı ve içerik ağırlıklandırmasının yapılmadığı koşullara ilişkin bulgular karşılaştırılmıştır. Her bir koşula ilişkin çok boyutlu BOB testi bulguları, boyutlara ilişkin güvenirlik katsayıları, ölçmenin standart hatası ve RMSD değerlerine bakılarak karşılaştırılmıştır. Analiz sonuçlarına göre, A-Optimality madde seçim yöntemi kullanıldığında, hem Bayesyen MAP hem de Fisher'in Puanlama yöntemlerinin benzer sonuçlar verdiği bulgusuna ulaşılmıştır. Buna karşın, Fisher'in puanlama yönteminin hem madde seçim yöntemlerinden hem de içerik ağırlıklandırmasından etkilendiği söylenebilir. Ayrıca içerik ağırlıklandırması uygulandığında her bir koşul için testteki ortalama madde sayısı artarken, güvenirlik katsayılarının azaldığı, buna karşın RMSD ve standart hataların azaldığı bulgusuna ulaşılmıştır.

Anahtar Kelimeler: Çok boyutlu bireyselleştirilmiş testler, maddeler-arası boyutluluk modeli, içerik ağırlıklandırması,

Abstract

The purpose of this study is to compare the performance of Between-item dimensionality-based Multidimensional CAT designs and to examine the effect of content balancing on different MCAT designs. For this purpose, real data set obtained from English Proficiency Test (EPT), which was administered by Hacettepe University between 2009 and 2013 academic years, was used. The three dimensional item pool consisted of real items measuring students' listening, grammar and reading abilities. Item pool consisted of 555 items

* Bu çalışma, birinci yazarın Prof. Dr. Selahattin GELBAL danışmanlığında tamamlanan doktora tezinden türetilmiştir

** Arş. Grv. Dr., Siirt Üniversitesi, Eğitim Bilimleri Bölümü, Ankara-Türkiye, b.ozdemir025@gmail.com
*** Prof. Dr., Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara-Türkiye, s.gelbal@hacettepe.edu.tr

Geliş Tarihi: 07.12.2015
Kabul Tarihi: 25.12.2015

which was calibrated with 2PL between-item MIRT model. In this study, two different theta estimation (Fisher scoring and Bayesian MAP) methods, three different fisher information based item selection methods (A-optimality, D-optimality and Random) and three different precision based termination methods were used in order to determine the best MCAT design. In addition, results of MCAT algorithms with content distribution and without content distribution were compared so as to examine the effect of content balancing in the context of MCAT. The results of each MCAT condition were compared with respect to, reliability index, SEM, RMSD values associated with each dimension. According to results, both Bayesian MAP and Fisher's scoring methods yielded similar results when A-Optimality item selection method was used. However, Fisher's scoring method appeared to be affected from item selection methods and content balancing. Moreover, average number of items tended to increase and reliability coefficients tended to decrease somewhat, while standard error and RMSD values tended to decrease when content balancing was applied in MCAT.

Keywords: Multidimensional adaptive testing, between-item dimensionality model, content balancing

GİRİŞ

Bilgisayar teknolojilerindeki gelişmeler sosyal hayatımızı, çevremizi ve yaşam tarzımızı etkilediği gibi eğitimde kullanılan ölçme ve değerlendirme araçlarını, yöntem ve tekniklerini de aşamalı olarak etkilemektedir. Bu gelişmelere bağlı olarak günümüzde daha nitelikli ve verimli eğitim vermek için bilgisayar teknolojilerinden önemli ölçüde faydalanılmaya çalışılmaktadır. Bu amaç doğrultusunda bireylerin özelliklerini veya yeteneklerini bilgisayar teknolojilerini kullanarak ölçmeyi amaçlayan ölçme yöntemleri geliştirilmiştir. Bunlardan en temel olanı testlerin kâğıt-kalem yerine bilgisayar ortamında sorulmasıdır. Bu yöntem genellikle bilgisayar destekli testler (Computer-Based Tests-CBT) olarak adlandırılmaktadır.

Diğer bir alternatif ölçme yöntemi, bireyin yetenek düzeyi ile maddelerin özelliklerinin bilgisayar ortamında eşleştirildiği bilgisayar ortamında bireyselleştirilmiş test (Computerized Adaptive Testing-CAT) yöntemleridir (McBride ve Martin, 1983; Weiss ve Kingsbury, 1984). Bilgisayar ortamında bireyselleştirilmiş (BOB) testler, psikolojik ve eğitimsel testlerin daha etkili ve verimli bir şekilde uygulanması için yeniden düzenlenerek interaktif bilgisayarlar ile uygulanmasıdır (Van der Linden ve Glas, 2000; Wainer et al., 2000). Bu yöntemin temel amacı, her bireyin ölçülen özelliğini en etkili ve verimli bir şekilde ölçecek bireyin yetenek düzeyine uygun maddeleri seçmektir.

Bilgisayar ortamında bireyselleştirilmiş testler, Binet'in geliştirmiş olduğu uyarlanmış test yönteminin uygulayıcı birey yerinde bilgisayar programı kullanılarak uygulanmasıdır. BOB testi yönteminin uygulanma sürecinde test maddeleri kullanılacak bilgisayar programına yüklenir ve bilgisayar ekranında görünmesi sağlanır. Daha sonra birey klavyeyi veya fareyi kullanarak maddelere cevap verir.

Binet testini uygulayan araştırmacı gibi, bilgisayar programı bireyin teste nasıl başlayacağına karar verir; bireyin önceki maddelere vermiş olduğu cevaplara bağlı olarak maddeleri seçer ve bir ya da birkaç kural belirleyerek testi sonlandırır. İlk BOB uygulamaları Binet'in kullanmış olduğu yöntemin farklı versiyonları iken (Weiss, 1973), sonraki uygulamalar ise madde havuzunun oluşturulma biçimine göre farklılık göstermektedir (Lord, 1971a, 1971b, 1971c).

Madde tepki tepki kuramının geliştirilmesine paralel olarak, bilgisayar teknolojisindeki ilerlemeler, bilgisayar ortamında bireyselleştirilmiş testlerin uygulanabilirliğini arttırmıştır. Madde seçim ve yetenek kestirim yöntemleri için tek boyutlu madde tepki kuramının kullanıldığı bireyselleştirilmiş testler tek boyutlu bilgisayar ortamında bireyselleştirilmiş test (Tek boyutlu BOB testleri-Unidimensional CAT) olarak adlandırılmaktadır (Wang ve Chen, 2004).

BOB testlerinde tek boyutlu MTK yaygın olarak kullanılmasına rağmen, gerçek test uygulamaları için uygun olmayabilir. Özellikle bilişsel özelliklerin ölçülmesi ve değerlendirilmesinin gerektiği portfolyo değerlendirmeleri, performans görevleri, klinik yeteneklerinin ölçülmesi ve değerlendirilmesi, yazma ve konuşma becerilerinin ölçülmesi ve projelerin değerlendirilmesi söz konusu olduğunda bireylerin yeteneklerinin doğru bir şekilde ölçülmesi için çok boyutlu madde tepki

kuramına ihtiyaç duyulmaktadır (van der Linden ve Hambleton, 1997, s. 221). Çok boyutlu madde tepki kuramının uygulama alanlarının yaygınlaşması ve bilgisayar ortamında bireyselleştirilmiş testlerin kâğıt-kalem testlerine alternatif olarak görülmesi, her iki yöntemin birleşimi olan *çok boyutlu bireyselleştirilmiş testlerin* (Segall, 1996, 2001) geliştirilmesine olanak sağlamaktadır.

Bireyselleştirilmiş testlerin geleneksel kâğıt-kalem testlerine göre avantajlı yönleri olduğunu belirtilmesine karşın, bu durum her bireyselleştirilmiş testin geleneksel kağıt-kalem testlerinden üstün olduğu anlamına gelmeyebilir. Ayrıca en uygun BOB testine karar vermek için testin içeriğine ve ölçtüğü özelliğin yapısına uygun bir BOB testi algoritmasının geliştirilmesi gerekir. Dolayısıyla, her hangi bir BOB testi geliştirme aşamasında cevaplanması gereken bazı temel sorular vardır. Bu temel sorular; hangi modelin kullanılacağı, ilk maddenin nasıl seçileceği, test sürecinde bireyin yetenek parametresinin nasıl hesaplanacağı, bir sonraki maddenin nasıl seçileceği ve testin nasıl sonlandırılacağıdır (Diao ve Reckase, 2009). Bu sorular uygun bir şekilde cevaplandırıldığında geliştirilmesi amaçlanan en uygun BOB testi belirlenmiş olur. Ayrıca, bu temel sorular, BOB testi geliştirme sürecinin aşamalarını oluşturmaktadır.

Madde Seçim Yöntemleri

Çok boyutlu bilgisayar ortamında bireye uyarlanmış testlerde kullanılan madde seçim yöntemleri ile ilgili ilk çalışmalar Bloxom ve Vale (1987) tarafından yürütülmüştür. Bloxom ve Vale (1987)'in çok boyutlu BOB testleri ile ilgili yapmış olduğu çalışmayı Fan ve Hsu (1996), Luecht (1996), Segall (1996, 2000), van der Linden (1996, 1999), ve Veldkamp ve van der Linden'in (2002) yapmış olduğu çalışmalar takip etmiştir.

Çok boyutlu BOB testlerinde kullanılan madde seçim yöntemlerinden bazıları optimal desenlere bağlı madde seçim yöntemleridir. Optimal desenler bilgi matrisinin veya kovaryans matrisinin determinantı veya izinin istatistiksel çıkarımları optimize etmede kullanıldığı yöntemlerdir. Bu yöntemler sırasıyla D-optimality ve A-optimality olarak adlandırılmaktadır (Silvey, 1980, s. 10). Fisher'in bilgi matrisi bu yöntemlerde önemli bir rol oynamaktadır. Çünkü Fisher'in bilgi matrisi gözlenen değişkenleri açıklayan gizil değişkenlere ait bilgiyi ölçmekte kullanılır.

Optimal desenler kullanılarak çok boyutlu BOB testleri için geliştirilmiş birçok madde seçim yöntemleri vardır. Bunlar; D-optimality (determinant of the posterior information matrix) (Luecht, 1996; Segall, 1996), A-optimality (minimizing the error variance of the composite measure) (van der Linden, 1999), C-optimality ve önsel dağılımın maksimize edildiği Kullback–Leibler bilgi fonksiyonun (Kullback–Leibler information- KLI) (Veldkamp ve van der Linden, 2002) yöntemleridir. Aşağıda çok boyutlu BOB testlerinde kullanılan madde seçim yöntemleri hakkında bilgi verilmiştir.

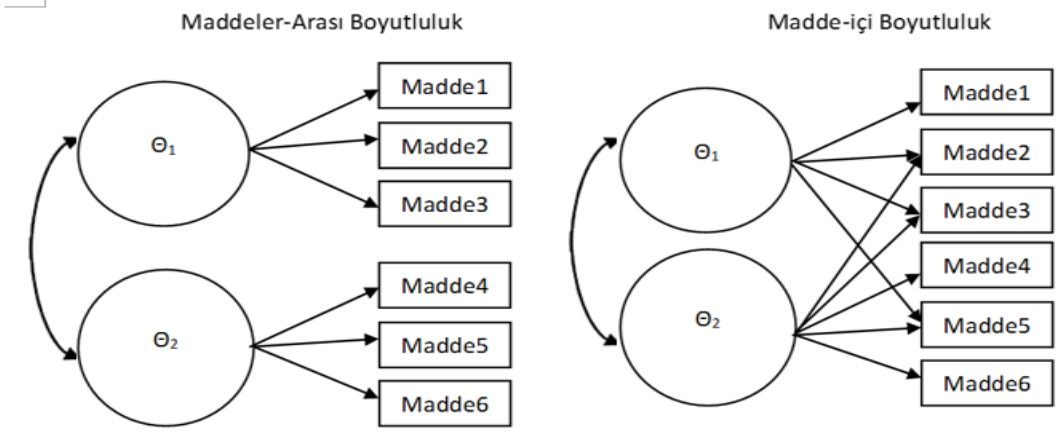
Geleneksel kâğıt-kalem testlerinde, test geliştirme sürecinde testin uygulanacağı alanın içeriğinin özellikleri göz önünde bulundurularak testi oluşturacak maddeler belirlenir. BOB testlerinde bireyin kestirilen yetenek parametresine ait en yüksek bilgiyi veren maddeler seçildiğinden test ile ölçülmek istenen farklı içeriklere ilişkin madde dağılımı farklılık gösterebilir. Dolayısıyla, bireyler ölçülen her bir içeriğe ya da konu alanına ait farklı sayıda maddelere cevap verir. Bu durum BOB testinin geçerliğini, puanların karşılaştırılabilirliğini tehlikeye sokmakla birlikte, testi alan ve testi uygulayan bireyler açısından sorun oluşturabilir. Bilgisayar ortamında bireyselleştirilmiş testlerde, yukarıda belirtilen problemleri ortadan kaldırmak için ilk olarak Green ve arkadaşları (1984) tarafından içerik ağırlıklandırması (content balancing) görüşü ortaya atılmış ve ilerleyen yıllarda farklı içerik ağırlıklandırması yöntemleri geliştirilmiştir.

Çok boyutlu MTK modelleri

Çok boyutlu BOB testlerinin temelini oluşturan çok boyutlu madde tepki kuramları Telafi-edici (compensatory) ve telafi-edici (noncompensatory) olmayan çok boyutlu modeller olarak ikiye ayrılır. Çok boyutlu modellerin dışında dikkat edilmesi gereken bir diğer nokta ise madde düzeyinde boyutluluktur. Madde düzeyindeki boyutluluk ise maddeler-arası boyutluluk (between-item dimensionality) ve madde-içi boyutluluk (within-item dimensionality) olmak üzere ikiye ayrılır

(Wang, Chen ve Cheng, 2004; Wang, Wilson ve Adams, 1997). Maddeler-arası boyutlulukta her bir madde sadece bir boyutta yük verir. Bu modelde maddelere ait ayırt edicilik parametreleri bir boyuttan sıfırdan farklı değer alırken diğer boyutlara ait ayırt edicilik parametresi sıfıra eşittir. Buna karşın, Madde-içi boyutluluk modelinde maddeler, birden fazla boyutta yük verir. Bu modelde maddelere ait madde ayırt edicilik parametresi ve madde yükleri diğer boyutlar içinde sıfırdan farklı değerler alabilir.

Şekil 1'de iki farklı boyutu ölçen bir test için telafi-edici çok boyutlu modellere ait madde düzeyinde boyutluluk modellerinden maddeler-arası ve madde-içi boyutluluk modelleri gösterilmiştir.



Şekil 1. Maddeler-arası ve madde-içi boyutluluk modeli (Wang ve Chen, 2004)

Bu çalışmada kullanılan telafi-edici çok boyutlu MTK modeli, maddelere ait şans başarının sıfıra eşit olduğu kabul edilen iki-parametreli çok boyutlu MTK modelidir (2PL-ÇBMTK/2PL MIRT). Bu modele ait formül aşağıda verilmiştir:

$$P(Uij = 1|\theta_j, a_i, d_i) = \frac{e^{(a_i \theta_j' + d_i)}}{1+e^{(a_i \theta_j' + d_i)}}$$

Burada θ_j, m yetenek düzeyinin ölçüldüğü 1xm şeklinde bir vektördür. Benzer şekilde a_i 1xm şeklinde ayrıt edicilik vektörü ve *d* kesişim parametresi ya da madde kolaylığı parametresi olarak adlandırılan değerdir.

Araştırmanın Amacı ve Önemi

Bu çalışmanın amacı, bireylerin yabancı dil yeteneklerinin maddeler-arası boyutluluk modeline dayalı Çok Boyutlu Bilgisayar Ortamında Bireyselleştirilmiş Test Yöntemleri ile ölçülmesi (Multidimensional Computerized Adaptive Testing-MCAT) ve içerik ağırlıklandırmasının (content balancing) çok boyutlu BOB testi yöntemlerinin performansları üzerindeki etkisinin incelenmesidir.

Alan yazınına bakıldığında, özellikle ülkemizde çok boyutlu BOB testi yöntemleri ile ilgili yapılan çalışmaların sınırlı sayıda olduğu görülmektedir. Bu araştırma çok boyutlu BOB testi yöntemlerinin performanslarının karşılaştırılmasına olanak sağladığından önemli görülmektedir. Ayrıca bu çalışmada testin yapısına bağlı olarak yapılan içerik ağırlıklandırmasının yapıldığı ve içerik ağırlıklandırmasının yapılmadığı duruma ilişkin analiz bulguları karşılaştırmaya olanak sağladığından önemli görülmektedir.

Bu çalışmayı önemli kılan bir diğer özelliği ise İYS sınavına ait gerçek verilerin kullanılması ve gerçek verilere dayalı simulasyon (post-hoc simulation) yönteminin kullanılmasıdır Monte Carlo simülasyon yönteminden temel farkı ise madde havuzunun gerçek maddelerden oluşmasıdır. Bu simülasyon yöntemi, genellikle, geleneksel kağıt kalem formatında kullanılan testin psikometrik özelliklerinde anlamlı bir değişim yapmadan BOB testi yöntemleri ile uygulandığında testteki

ortalama madde sayısında ne kadar azalma olacağını tespit etmeyi amaçlar (IACAT, 2015). Post-hoc simülasyon yönteminin diğer aşamaları, BOB testi yöntemi ile aynıdır.

Problem Cümlesi

İçerik ağırlıklandırmasının maddeler-arası boyutluluk modeline dayalı farklı yetenek kestirimi yöntemleri, madde seçim yöntemleri ve test sonlandırma kurallarının kullanıldığı çok boyutlu BOB testi yöntemlerinin performansları üzerindeki etkisi nasıldır?

Alt Problemler

1. A-optimality madde seçim yönteminin kullanıldığı çok-boyutlu BOB testi analizlerine ilişkin farklı yetenek kestirim yöntemleri ve durdurma kuralları altında her bir koşula ait güvenirlik katsayısı, standart hata, testin uzunluğu ve RMSD değerleri nasıldır?

2. D-optimality madde seçim yönteminin kullanıldığı çok-boyutlu BOB testi analizlerine ilişkin farklı yetenek kestirim yöntemleri ve durdurma kuralları altında her bir koşula ait güvenirlik katsayısı, standart hata, testin uzunluğu ve RMSD değerleri nasıldır?

3. Seçkisiz (random) madde seçim yönteminin kullanıldığı çok-boyutlu BOB test analizlerine ilişkin farklı yetenek kestirim yöntemleri ve durdurma kuralları altında her bir koşula ait güvenirlik katsayısı, standart hata, testin uzunluğu ve RMSD değerleri nasıldır?

4. İçerik ağırlıklandırmasının kullanılmadığı çok boyutlu BOB testleri ile kaşılaştırıldığında, İçerik ağırlıklandırmasının farklı madde seçim, yetenek kestirim yöntemleri ve durdurma kurallarının kullanıldığı çok-boyutlu BOB testi yöntemleri üzerindeki etkisi nasıldır?

Sayıltılar

Genellikle çok-boyutlu BOB testlerinde her bir boyuta ilişkin yetenek parametreleri arasındaki korelasyonun veya varyans-kovaryans matrisi önsel dağılımının bilindiği varsayılmaktadır (Yoo, 2011). Bu çalışmada bireylerin kestirilen yetenek parametreleri arasındaki korelasyona bakılarak varyans-kovaryans matrisine ait önsel dağılımının bilindiği varsayılmaktadır (vcv=c[0.9, 0.8, 0,8])

Ayrıca, 2009-2013 yıllarında uygulanan İYS testlerinin bireyin ölçülen özelliklerini doğru ve güvenilir bir şekilde ölçtüğü ve bireylerin test maddelerine verdiği cevapların gerçek yetenek düzeylerini yansıttığı varsayılmaktadır.

Sınırlılıklar

Bu çalışmada, gerçek verilere dayalı simülasyon yapılmasına olanak sağlayan ve çok-boyutlu BOB testleri için geliştirilmiş "MAT " (multidimensional adaptive testing, Choi ve King, 2011) R paket programı kullanılmıştır. Dolayısıyla, bu çalışmada kullanılan madde seçim yöntemleri, yetenek kestirim yöntemleri, test sonlandırma kuralı ve madde kullanım sıklığı kontrol yöntemleri paket programda tanımlı yöntemlerle sınırlıdır.

Telafi edici Çok boyutlu madde tepki kuramına dayalı madde düzeyindeki boyutluluk modelleri *madde-içi* ve *maddeler-arası boyutluluk* modelleri olmak üzere ikiye ayrılmaktadır. Bu çalışma maddeler-arası boyutluluk modeline dayalı çok boyutlu BOB testi yöntemleri ile sınırlı tutulmuştur.

YÖNTEM

Araştırma Yöntemi

Bu araştırmada, yabancı dil sınavına giren bireylerin yabancı dil yeteneklerinin maddeler-arası boyutluluk modeline dayalı Çok Boyutlu Bilgisayara ortamında bireyselleştirilmiş (Çok-Boyutlu BOB) test yöntemleri ile kestirilmesi ve çok boyutlu BOB testi yöntemlerinin performanslarının karşılaştırılması amaçlanmaktadır. Araştırmada var olan yöntem ve tekniklerin gerçek veri üzerinden performanslarının karşılaştırılması amaçlandığından araştırma nicel karşılaştırma araştırmasıdır.

Çalışma grubu

Araştırmanın çalışma grubunu Hacettepe Üniversitesinde İngilizce Yeterlik Sınavı (İYS)'na giren bireyler oluşturmaktadır. Her bir dönem sonunda ve dönem içerisinde İYS'ye giren öğrenci sayısı

1200 ile 2000 arasında değişmektedir. Araştırmanın örneklemini ise 2008-2013 eğitim-öğretim yıllarında ilkbahar ve sonbahar dönemlerinde Hacettepe Üniversitesinde İYS'ye giren bireyler oluşturmaktadır. Analiz sürecinde gerçek madde parametrelerine bağlı olarak bireylere ilişkin yetenek parametreleri türetilmiş ve her bir analiz için aynı yetenek parametreleri kullanılmıştır. Çok-boyutlu BOB testi sürecinde her bir koşul için yapılan analizlerde testi alan birey sayısı 500 ile sınırlandırılmıştır. Bireylerin her bir boyuta ilişkin yetenek parametreleri [-3, +3] aralığında çok değişkenli normal dağılım göstermektedir.

Araştırma Verileri

Araştırmanın verilerini, 2008-2013 eğitim-öğretim yıllarında ilkbahar ve sonbahar dönemlerinde Hacettepe Üniversitesinde İYS'ye ait testlerdeki maddeler ve sınava giren bireylere ait cevap örüntülerinin yer aldığı veri seti oluşturmaktadır. Sınav dinleme (listening), okuduğunu anlama (reading) ve dilbilgisi (grammar) olmak üzere üç temel bölümden oluşup her bir testteki ortalama madde sayısı ise toplam 65'dir.

Verilerin Analizi

Verilerin analizi 2 aşamadan oluşmaktadır. İlk aşamada çok boyutlu bilgisayar ortamında bireyselleştirilmiş testlerde kullanılacak maddelere ait madde parametrelerinin maddeler-arası boyutluluk modeline dayalı tanımlayıcı çok boyutlu MTK modelleri ile kestirilmiş ve her bir modele ait madde havuzu oluşturulmuştur. Bu aşamada, testlere ait maddeler telafi-edici çok boyutlu MTK modellerinden maddeler-arası boyutluluk modeli ile kalibre edilerek maddeler-arası boyutluluk modeline ait madde havuzu oluşturulmuştur. 2009-2013 yılları arasında uygulanan ve üç boyuttan oluşan İYS sınavına ait 10 testteki toplam 628 madde, telafi-edici çok boyutlu madde tepki kuramına (Compensatory 2PL-MIRT) dayalı maddeler-arası boyutluluk modeli ile kalibre edilmiştir. Boyutlara ait madde ayırt edicilik parametresi 0,5'in altında olan ve d-parametresi [-4,4] aralığının dışında olan toplam 73 madde havuzunda çıkartılmıştır. Sonuç olarak, telafi-edici çok boyutlu MTK'ya dayalı maddeler-arası boyutluluk modeline ait madde havuzu, dinleme boyutunda 115, dilbilgisi boyutunda 240 ve okuma boyutunda 200 madde olmak üzere toplam 555 maddeden oluşmuştur.

İkinci aşamada madde havuzu oluşturulduktan sonra post-hoc simülasyon yöntemi uygulanarak simülasyon veri seti yerine İYS'ye ait gerçek veri seti kullanılarak bireylere ait yetenek kestirilmiştir. Bireylere ait yetenek kestirimi yapılırken çok boyutlu BOB testi analizi için R-yazılımında tanımlı "MAT" paket programı (Choi ve King, 2011) kullanılmıştır. Çok boyutlu BOB testi yöntemleri ile analiz yapılırken, farklı yetenek kestirme yöntemleri, madde seçim yöntemleri ve durdurma kuralları kullanılarak her bir boyuta ilişkin yetenek kestirimi yapılmıştır.

Maddeler-arası boyutluluk modeline dayalı çok boyutlu BOB testi için en uygun madde seçim yöntemini belirlemek amacıyla Fisher'in bilgi matrisine dayalı *A-optimality*, *D-optimality* ve her hangi bir kuralın kullanılmadığı *seçkisiz (Random)* madde seçim yöntemleri kullanılmıştır. Çok-boyutlu BOB testi için en uygun yetenek kestirim yöntemini belirlemek için maksimum likelihood estimation (MLE) yetenek kestirim yöntemlerinden Fisher'in puanlama yetenek kestirim yöntemi ve Bayesyen maksimum a posteriori (MAP) yetenek kestirim yöntemi kullanılmıştır. Ayrıca, çok-boyutlu BOB testinde en uygun test sonlandırma kuralını belirlemek için ise farklı hata varyansı durdurma kuralı kullanılmıştır.

Bu çalışmada, içerik ağırlıklandırmasının çok boyutlu BOB testleri üzerindeki etkisi incelemek amacıyla farklı madde seçim yöntemleri, yetenek kestirim yöntemleri ve durdurma kurallarının kullanıldığı çok-boyutlu BOB testi algoritmalarına ait her bir boyuta ilişkin *güvenirlik katsayıları*, yetenek parametrelerine ilişkin *RMSD katsayıları* ve *ölçmenin standart hatası* değerleri hesaplanarak karşılaştırılmıştır. Her bir koşul için maddenin kullanım sıklığını (item exposure) kontrol etmeye olanak sağlayan randomesque yöntemi kullanılmıştır. Bu yöntem ile analizlerde madde havuzundan madde seçilirken test bilgisini maksimum yapan ilk on madde arasından birinin seçkisiz olarak atanması kuralı uygulanmıştır.

BULGULAR

Tablo 1'de maddeler-arası boyutluluk modeli için madde seçim yöntemlerinden A-optimality, durdurma kurallarından hata varyansı ve yetenek kestirim yöntemlerinden Fisher'in puanlama ve Bayesyen MAP yetenek kestirim yöntemlerine ait analiz bulgularına yer verilmiştir.
Her bir model için standart hata durdurma kuralı kullanılarak analizler bireylerin kestirilen yetenek parametrelerine ait hata varyansının sırasıyla 0,20, 0,25 ve 0,30'un altına düştüğünde testler sonlandırılmıştır. Ayrıca her bir koşul için içerik ağırlıklandırmasının kullanıldığı ve kullanılmadığı durumlara ait sonuçlar karşılaştırılarak, içerik ağırlıklandırmasının çok boyutlu BOB testleri üzerindeki etkisi incelenmiştir

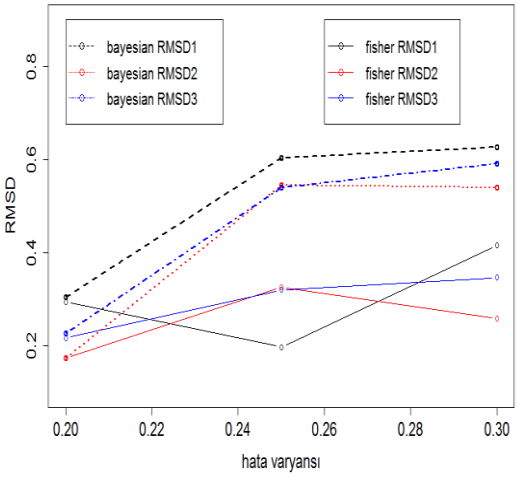
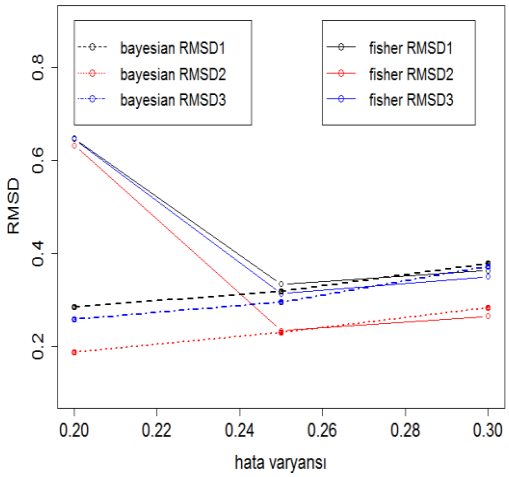
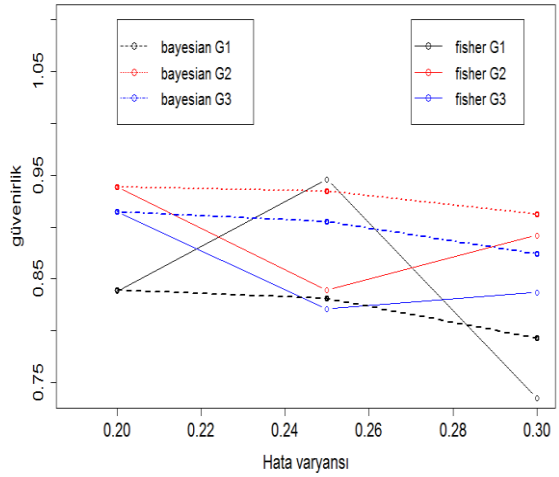
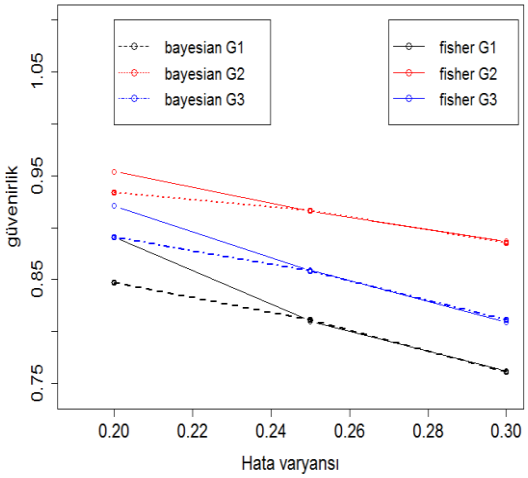
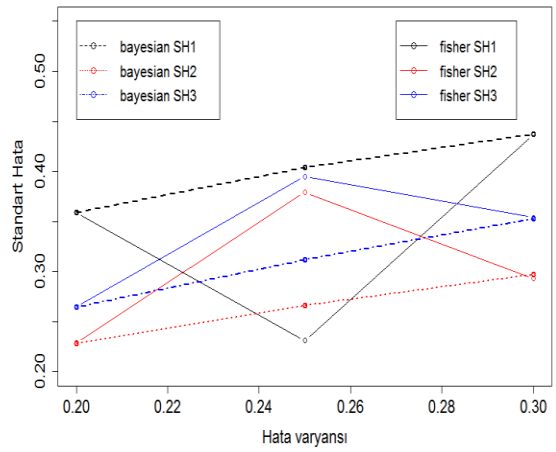
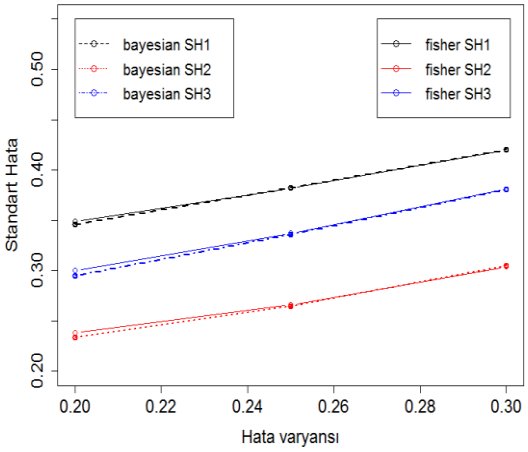
Tablo 1. A-Optimality madde seçim yöntemine ait Çok boyutlu BOB testi Bulguları

	Yetenek Kestirim yöntemi	Test sonlandırma kuralı	Test uzunluğu	güvenirlik			ÖSH*			RMSD		
		S. Hata	K	boy1	boy2	boy3	boy1	boy2	boy3	boy1	boy2	boy3
Eşit madde dağılımlı	Fisher	0,20	39,6	0,84	0,94	0,92	0,359	0,229	0,265	0,294	0,173	0,216
		0,25	31,4	0,95	0,84	0,82	0,231	0,379	0,395	0,197	0,326	0,320
		0,30	15,1	0,74	0,89	0,84	0,437	0,293	0,354	0,415	0,259	0,346
	Bayesian	0,20	39,7	0,84	0,94	0,92	0,359	0,228	0,265	0,305	0,174	0,226
		0,25	22,8	0,83	0,94	0,91	0,404	0,266	0,312	0,604	0,545	0,540
		0,30	15,2	0,79	0,91	0,87	0,437	0,297	0,353	0,626	0,540	0,591
İçerik ğırlıklandırmalı	Fisher	0,20	45,5	0,89	0,95	0,92	0,349	0,238	0,300	0,647	0,632	0,645
		0,25	29,7	0,81	0,92	0,86	0,382	0,266	0,337	0,334	0,234	0,314
		0,30	19,8	0,76	0,89	0,81	0,420	0,304	0,381	0,363	0,265	0,350
	Bayesian	0,20	44,9	0,85	0,93	0,89	0,346	0,234	0,295	0,285	0,187	0,259
		0,25	29,7	0,81	0,92	0,86	0,382	0,265	0,336	0,318	0,230	0,296
		0,30	20,1	0,76	0,89	0,81	0,420	0,305	0,381	0,379	0,283	0,372

*OSH= Ölçmenin standart Hatası

Tablo 1'deki çok-boyutlu BOB testi analiz bulgularına bakıldığında, genel olarak, madde seçim yöntemlerinden A-optimality, yetenek kestirim yöntemlerinden Fisher'in puanlama yöntemi ve standart hata durdurma kuralının 0,25 olduğu koşula ait testteki ortalama madde sayısının 31,4 olup daha güvenilir ve tutarlı sonuçlar verdiği söylenebilir. Hata varyansı durdurma kuralı 0,30 olarak belirlendiğinde, testteki ortalama madde sayısı 15,1 e düşmesine karşın boyutlara ilişkin güvenirlik katsayılarının düştüğü, standart hata ve RMSD değerlerinin arttığı görülmektedir. Diğer taraftan, yetenek kestirim yöntemlerinden Bayesyen MAP yetenek kestirim yöntemi kullanıldığında standart hatanın 0,25'e eşitlendiği durumda ortalama madde sayısının 22,8'e düştüğü ve diğer koşullara göre güvenilir ve tutarlı sonuçlar verdiği söylenebilir.

Tablo 1'deki içerik ağırlıklandırmasının yapıldığı duruma ilişkin çok boyutlu BOB testi sonuçlarına bakıldığında ise standart hata durdurma kuralının 0,25 olduğu koşula ait testteki ortalama madde sayısının 29,7 olup hem Fisher'in puanlama yöntemi hem de Bayesyen MAP yetenek kestirim yöntemlerinin benzer sonuçlar verdiği görülmektedir. Ancak içerik ağırlıklandırmasının yapılmadığı ve Fisher'in puanlama yetenek kestirim yöntemi kullanıldığı durumda birinci boyuta ilişkin güvenirlik katsayısı en yüksek değere sahipken diğer koşullarda ise, ikinci boyuta ilişkin güvenirlik katsayısı en yüksek değere sahiptir. Dolayısıyla, içerik ağırlıklandırması yapıldığında Fisher'in puanlama yönteminin daha güvenilir ve tutarlı sonuçlar verdiği söylenebilir.

Eşit madde dağılımlı	İçerik Ağırlıklandırmalı
a.1 RMSD- Hata varyansı ilişkisi	a.2 RMSD- Hata varyansı ilişkisi
	
b.1 Güvenirlik- Hata varyansı ilişkisi	b.2 Güvenirlik- Hata varyansı ilişkisi
	
c.1 Ölçmenin standart hatası- Hata varyansı	c.2 Ölçmenin standart hatası- Hata varyansı
	

Şekil 2. A-Optimality Madde Seçim Yöntemine İlişkin Grafikler

Şekil 2'de içerik ağırlıklandırmasının yapılmadığı eşit madde dağılımlı ve içerik ağırlıklandırmasının yapıldığı duruma ilişkin madde seçim yöntemlerinden A-optimality, testi durdurma kurallarından boyutlara ilişkin hata varyansı ve yetenek kestirim yöntemlerinden Fisher'in puanlama ve Bayesyen MAP yetenek kestirim yöntemlerinin kullanıldığı çok-boyutlu BOB testi analizlerine ait grafikler verilmiştir.

Şekil 2'de yer alan a.1 ve a.2 grafikleri sırasıyla içerik ağırlıklandırmasının yapılmadığı ve içerik ağırlıklandırmasının yapıldığı duruma ilişkin Fisher'in puanlama ve Bayesyen MAP yetenek kestirim yöntemlerine ait her bir boyuta ilişkin RMSD değerleri ile yetenek parametrelerine ilişkin hata varyansları arasındaki ilişkiyi vermektedir. Şekil 2 a.1'e bakıldığında eşit madde dağılımlı durumda, yetenek parametrelerine ilişkin hata varyansı arttıkça Bayesyen MAP yetenek kestirimine ait her bir boyuta ilişkin RMSD değerlerinin arttığı görülmektedir. Ancak, Fisher'in puanlama yönteminin kullanıldığı durumda hata varyansı arttıkça RMSD değerlerinin önce artıp sonra azaldığı görülmektedir. Dolayısıyla Bayesyen yetenek kestirimi yönteminin daha güvenilir ve tutarlı sonuçlar verdiği söylenebilir.

İçerik ağırlıklandırmasının yapıldığı duruma ilişkin RMSD değerleri ile yetenek parametrelerine ait hata varyansı arasındaki ilişkiyi veren Şekil 2 a.2'ye bakıldığında, kestirilen yetenek parametrelerine ilişkin hata varyansı arttıkça Fisher'in puanlama yetenek kestirimine ait her bir boyuta ilişkin RMSD değerlerinin önce azaldığı, daha sonra ise artma eğilimi gösterdiği görülmektedir. Buna karşın, Bayesyen MAP yetenek kestirim yönteminde hata varyansı arttıkça boyutlara ilişkin RMSD değerlerinin düzenli olarak artma eğilimi gösterdiği görülmektedir. Ayrıca içerik ağırlıklandırması uygulandığında, Bayesyen yetenek kestirim yöntemine ait RMSD değerlerinin daha düşük olduğu görülmektedir.

Şekil 2'de yer alan b.1 ve b.2 grafikleri içerik ağırlıklandırmasının yapılmadığı ve içerik ağırlıklandırmasının yapıldığı duruma ilişkin yetenek kestirim yöntemlerine ait her bir boyuta ilişkin güvenirlik katsayıları ile yetenek parametrelerine ait hata varyansları arasındaki ilişkiyi vermektedir. Şekil 2 b.1'e bakıldığında yetenek parametrelerine ait hata varyansları arttıkça, Bayesyen MAP yetenek kestirimine yöntemine ait her bir boyuta ilişkin güvenirlik katsayılarının düzenli bir şekilde azaldığı görülmektedir. Ancak, Fisher'in puanlama yönteminin kullanıldığı durumda hata varyansı arttıkça boyutlara ilişkin güvenirlik katsayılarının artıp azaldığı görülmektedir.

İçerik ağırlıklandırmasının yapıldığı duruma ilişkin güvenirlik katsayıları ile yetenek parametrelerine ait hata varyansı arasındaki ilişkiyi veren Şekil 2 b.2'ye bakıldığında, her bir boyuta ilişkin hata varyansı arttıkça hem Bayesyen MAP hem de Fisher'in puanlama yetenek kestirim yöntemine ait güvenirlik katsayılarının azaldığı görülmektedir. Hata varyansı durdurma kuralı arttıkça testteki ortalama madde sayısı arttığından her iki yetenek kestirim yöntemine ait güvenirlik katsayıları birbirine yakın değerler aldığı görülmektedir. Dolayısıyla içerik ağırlıklandırması yapıldığı durumda hem Bayesyen hem de Fisher'in puanlama yöntemi benzer sonuçlar verdiği söylenebilir.

Şekil 2'de yer alan c.1 ve c.2 grafikleri sırasıyla içerik ağırlıklandırmasının yapılmadığı ve içerik ağırlıklandırmasının yapıldığı duruma ilişkin Fisher'in puanlama ve Bayesyen MAP yetenek kestirim yöntemlerine ait her bir boyuta ilişkin ölçmenin standart hatası ile hata varyansı arasındaki ilişkiyi vermektedir. Şekil 2 c.1'e bakıldığında, Bayesyen MAP yetenek kestirimi yöntemi kullanıldığında, yetenek parametrelerine ilişkin hata varyansı arttıkça, her bir boyuta ilişkin ölçmenin standart hatasının da arttığı görülmektedir. Buna karşın, Bayesyen MAP yetenek kestirimi yöntemi yerine Fisher'in puanlama yöntemi kullanıldığında, boyutlara ilişkin hata varyansı arttıkça boyutlara ilişkin standart hatanın düzenli bir artış göstermediği görülmektedir.

Şekil 2 c.2'ye bakıldığında, boyutlara ilişkin hata varyansı arttıkça hem Fisher'in puanlama hem de Bayesyen MAP yetenek kestirimine ait her bir boyuta ilişkin ölçmenin standart hatasının da arttığı görülmektedir. Ayrıca içerik ağırlıklandırmasının yapıldığı durumda, Fisher'in puanlama ve Bayesyen MAP yetenek kestirimine ait her bir boyuta ilişkin ölçmenin standart hatasının birbirine yakın değerler aldığı görülmektedir.

Genel olarak hata varyansı durdurma kuralının kullanıldığı maddeler-arası çok boyutlu BOB testi analiz yöntemlerine ilişkin bulgular karşılaştırıldığında, hem Bayesyen MAP hem de Fisher'in puanlama yöntemlerine ait en uygun hata varyansı durdurma kuralının 0,25 olduğu görülmektedir. Ayrıca Fisher'in puanlama yöntemine ait çok boyutlu BOB testi sonuçlarının içerik ağırlıklandırmasından etkilendiği ve içerik ağırlıklandırmasının uygulandığı çok-boyutlu BOB testi uygulamalarında daha tutarlı sonuçlar verdiği görülmektedir. Diğer taraftan, Bayesyen yetenek kestirim yönteminin içerik ağırlıklandırmasından çok az etkilendiği ve özellikle içerik ağırlıklandırması uygulandığı durumda testteki ortalama madde sayısında az bir artış olmasına rağmen daha düşük standart hata ve RMSD değerlerine sahip olduğu yorumu yapılabilir.

Tablo 3'te maddeler-arası boyutluluk modeli için madde seçim yöntemlerinden D-optimality, durdurma kurallarından hata varyansı, yetenek kestirim yöntemlerinden Fisher'in puanlama ve Bayesyen MAP yetenek kestirim yöntemlerine ait analiz bulgularına yer verilmiştir. Ayrıca her bir koşul için içerik ağırlıklandırmasının kullanıldığı ve kullanılmadığı durumlara ait sonuçlar karşılaştırılarak, içerik ağırlıklandırmasının çok boyutlu BOB testleri üzerindeki etkisi incelenmiştir.

Tablo 2. D-Optimality madde seçim yöntemine ait Çok boyutlu BOB testi Bulguları

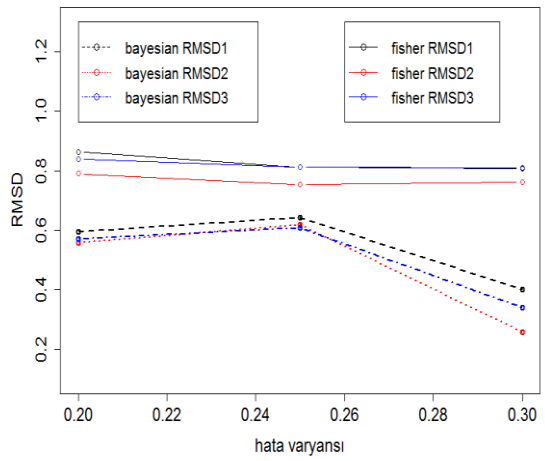
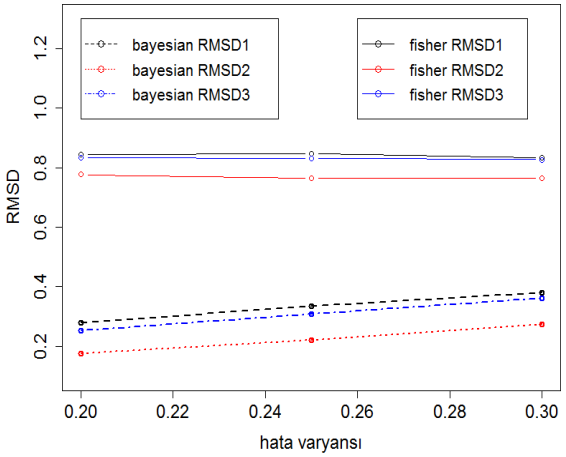
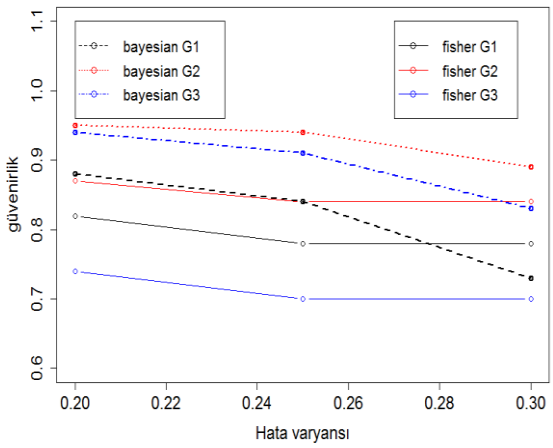
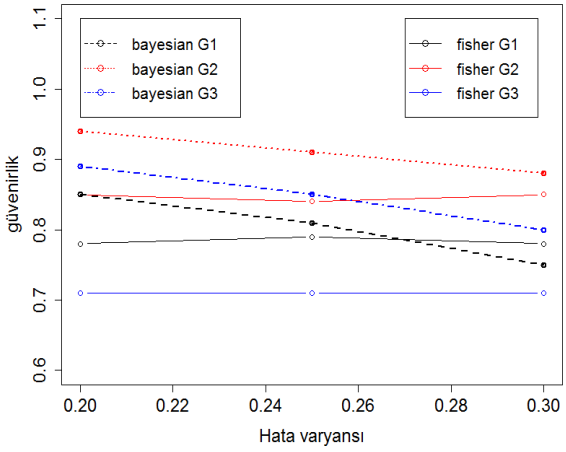
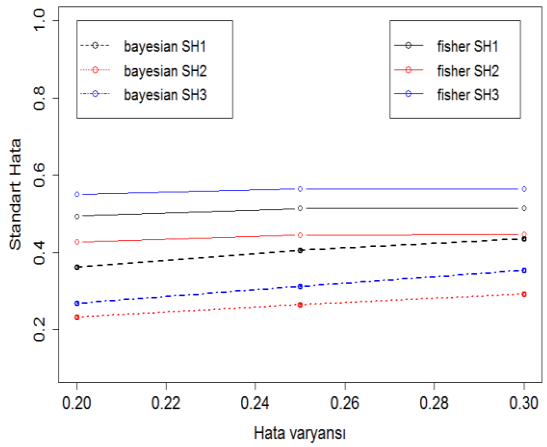
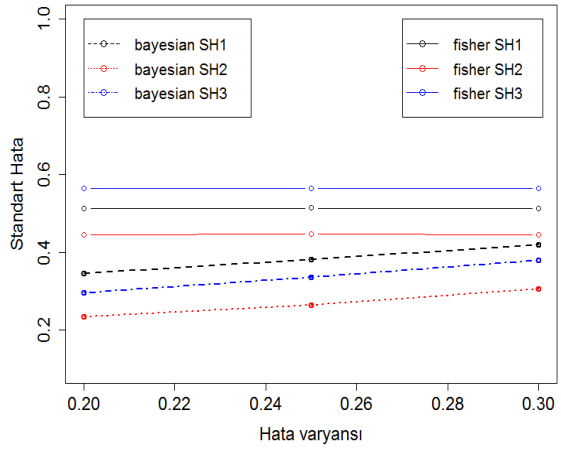
	Yetenek Kestirim yöntemi	Test Sonlandırma kuralı	Test uzunluğu	güvenirlik			ÖSH*			RMSD		
		S. Hata	K	boy1	boy2	boy3	boy1	boy2	boy3	boy1	boy2	boy3
madde		0,20	50,00	0,82	0,87	0,74	0,494	0,427	0,551	0,863	0,790	0,840
		0,25	50,00	0,78	0,84	0,70	0,514	0,446	0,565	0,812	0,753	0,812
	Fisher	0,30	49,43	0,78	0,84	0,70	0,514	0,447	0,565	0,806	0,763	0,811
Eşit dağılımlı		0,20	39,49	0,88	0,95	0,94	0,361	0,232	0,269	0,596	0,558	0,571
		0,25	22,83	0,84	0,94	0,91	0,406	0,265	0,311	0,644	0,620	0,608
	Bayesian	0,30	15,06	0,73	0,89	0,83	0,436	0,293	0,353	0,400	0,257	0,339
İçerik Ağırlıklandırmalı		0,20	50,00	0,78	0,85	0,71	0,513	0,445	0,564	0,844	0,777	0,833
		0,25	50,00	0,79	0,84	0,71	0,514	0,448	0,565	0,847	0,765	0,831
	Fisher	0,30	49,34	0,78	0,85	0,71	0,512	0,445	0,564	0,834	0,765	0,827
		0,20	45,31	0,85	0,94	0,89	0,346	0,235	0,297	0,278	0,175	0,253
		0,25	29,61	0,81	0,91	0,85	0,382	0,265	0,336	0,336	0,222	0,307
	Bayesian	0,30	19,59	0,75	0,88	0,80	0,420	0,306	0,380	0,381	0,275	0,362

*OSH= Ölçmenin standart Hatası

Tablo 2'deki çok-boyutlu BOB testi analiz bulgularına bakıldığında, yetenek kestirim yöntemlerinden Fisher'in puanlama yöntemi kullanıldığında, hata durdurma kuralının 0,30 olduğu koşulda bile testteki ortalama madde sayısının 49,43 olduğu görülmektedir. İngilizce yeterlik sınavının kâğıt-kalem testi formatındaki ortalama madde sayısı göz önünde bulundurularak çok-boyutlu BOB testi analizlerinde testteki maksimum madde sayısı 50 ile sınırlandırılmıştır. Böylece, standart hata durdurma kuralının kullanıldığı çok-boyutlu BOB testi analizlerinde çok uzun testler ile yetenek kestirimi önlenmiştir. Testteki ortalama madde sayısının diğer koşullarla karşılaştırıldığında oldukça yüksek olmasına karşın boyutlara ilişkin standart hata ve RMDS değerlerinin de yüksek olduğu görülmektedir. Diğer taraftan, yetenek kestirim yöntemlerinden Bayesyen MAP yetenek kestirim yöntemi kullanıldığında standart hatanın 0,25'e eşitlendiği durumda ortalama madde sayısının 22,8'e düştüğü ve A-Optimality madde seçim yönteminin kullanıldığı durum ile benzer sonuçlar verdiği söylenebilir.

Tablo 2'deki içerik ağırlıklandırmasının yapıldığı duruma ilişkin çok boyutlu BOB testi sonuçlarına bakıldığında ise Fisher'in puanlama yöntemi için testteki ortalama madde sayısında değişme olmazken güvenirlik katsayılarında az da olsa azalma olduğu görülmektedir. Genel olarak içerik ağırlıklandırmasının uygulandığı, her bir koşul için Bayesyen MAP yetenek kestirim yönteminin

daha az madde ile daha güvenilir ve tutarlı sonuçlar verdiği söylenebilir. Fisher'in puanlama yönteminin kullanıldığı durumda boyutlara ilişkin hata varyanslarının azalması testteki ortalama madde sayısını ve her boyuta ilişkin güvenirlik katsayılarını etkilemediği görülmektedir.

Eşit Madde Dağılımlı	*İçerik Ağırlıklandırmalı*
a.1 RMSD- Hata varyansı ilişkisi	a.2 RMSD- Hata varyansı ilişkisi
	
b.1 Güvenirlik- Hata varyansı ilişkisi	b.2 Güvenirlik- Hata varyansı ilişkisi
	
c.1 Ölçmenin standart hatası- Hata varyansı	c.2 Ölçmenin standart hatası- Hata varyansı
	

Şekil 3. D-optimality Madde Seçim Yöntemine İlişkin Grafikler

Şekil 3'de eşit madde dağılımlı ve içerik ağırlıklandırmasının yapıldığı duruma ilişkin madde seçim yöntemlerinden D-optimality, testi durdurma kurallarından boyutlara ilişkin hata varyansı ve yetenek kestirim yöntemlerinden Fisher'in puanlama ve Bayesyen MAP yetenek kestirim yöntemlerinin kullanıldığı çok-boyutlu BOB testi analizlerine ait grafikler verilmiştir.

Şekil 3'de yer alan a.1 ve a.2 grafikleri sırasıyla içerik ağırlıklandırmasının yapılmadı ve içerik ağırlıklandırmasının yapıldığı duruma ilişkin Fisher'in puanlama ve Bayesyen MAP yetenek kestirim yöntemlerine ait her bir boyuta ilişkin RMSD değerleri ile yetenek parametrelerine ilişkin hata varyansları arasındaki ilişkiyi vermektedir. Şekil 3 a.1'e bakıldığında eşit madde dağılımlı durumda, yetenek parametrelerine ilişkin hata varyansı arttıkça Bayesyen MAP yetenek kestirimine ait her bir boyuta ilişkin RMSD değerlerinin düzenli bir dağılım göstermediği görülmektedir. Ancak, Fisher'in puanlama yönteminin kullanıldığı durumda hata varyansı arttıkça RMSD değerlerinin artması beklenirken boyutla ilişkin RMSD değerlerinin azaldığı görülmektedir. Dolayısıyla hem Bayesyen hem de Fisher'in puanlama yetenek kestirimi yöntemlerinin tutarlı sonuçlar vermediği söylenebilir.

İçerik ağırlıklandırmasının yapıldığı duruma ilişkin RMSD değerleri ile yetenek parametrelerine ait hata varyansı arasındaki ilişkiyi veren Şekil 2 a.2'ye bakıldığında, kestirilen yetenek parametrelerine ilişkin hata varyansı arttıkça Fisher'in puanlama yetenek kestirimine ait her bir boyuta ilişkin RMSD değerlerinde artma ya da azalma olmadığı görülmektedir. Buna karşın, Bayesyen MAP yetenek kestirim yönteminde hata varyansı arttıkça boyutlara ilişkin RMSD değerlerinin düzenli olarak artma eğilimi gösterdiği görülmektedir. Ayrıca içerik ağırlıklandırması uygulandığında, Bayesyen yetenek kestirim yöntemine ait RMSD değerlerinin daha düşük olduğu görülmektedir. Dolayısıyla, çok boyutlu BOB testlerinde içerik ağırlıklandırması yapıldığında hem Bayesyen MAP hem de Fisher'in puanlama yönteminin daha güvenilir ve tutarlı sonuçlar verdiği yorumu yapılabilir.

Şekil 3'te yer alan b.1 ve b.2 grafikleri içerik ağırlıklandırmasının yapılmadığı ve içerik ağırlıklandırmasının yapıldığı duruma ait her bir boyuta ilişkin güvenirlik katsayıları ile yetenek parametrelerine ait hata varyansları arasındaki ilişkiyi vermektedir. Şekil 3 b.1'e bakıldığında yetenek parametrelerine ait hata varyansları arttıkça, Bayesyen MAP yetenek kestirimine yöntemine ait her bir boyuta ilişkin güvenirlik katsayılarının düzenli bir şekilde azaldığı görülmektedir. Ancak, Fisher'in puanlama yönteminin kullanıldığı durumda hata varyansı arttıkça boyutlara ilişkin güvenirlik katsayılarında anlamlı bir değişim gözlenmemektedir

İçerik ağırlıklandırmasının yapıldığı duruma ilişkin güvenirlik katsayıları ile yetenek parametrelerine ait hata varyansı arasındaki ilişkiyi veren Şekil 3 b.2'ye bakıldığında, her bir boyuta ilişkin hata varyansı arttıkça Bayesyen MAP yetenek kestirim yöntemine ait güvenirlik katsayılarının azaldığı, Fisher'in puanlama yöntemine ait güvenirlik katsayılarının ise değişmediği görülmektedir. Bunun temel sebebi, Fisher'in puanlama yöntemi için hata varyansı durdurma kuralı 0,20'den 0,30 çıkmasına rağmen testteki ortalama madde sayısı değişmemesidir. Dolayısıyla, D-Optimality madde seçim yöntemi için içerik ağırlıklandırması uygulandığında Bayesyen MAP yetenek kestirim yöntemi daha az madde ile daha güvenilir ve tutarlı sonuçlar verdiği söylenebilir.

Şekil 3'de yer alan c.1 ve c.2 grafikleri sırasıyla içerik ağırlıklandırmasının yapılmadığı ve içerik ağırlıklandırmasının yapıldığı duruma ilişkin her bir boyuta ait ölçmenin standart hatası ile hata varyansı arasındaki ilişkiyi vermektedir. Şekil 3 c.1'e bakıldığında, Bayesian MAP yetenek kestirimi yöntemi kullanıldığında, yetenek parametrelerine ilişkin hata varyansı arttıkça, her bir boyuta ilişkin ölçmenin standart hatasının da arttığı görülmektedir. Buna karşın, Bayesyen MAP yetenek kestirimi yöntemi yerine Fisher'in puanlama yöntemi kullanıldığında, boyutlara ilişkin hata varyansı arttıkça boyutlara ilişkin standart hatanın değişmediği görülmektedir.

Şekil 3 c.2'ye bakıldığında, içerik ağırlıklandırmasının yapılmadığı durum ile benzer sonuçlar verdiği görülmektedir. Ayrıca içerik ağırlıklandırmasının yapıldığı durumda, Bayesyen MAP yetenek kestirimine ait her bir boyuta ilişkin standart hatanın Fisher'in puanlama yöntemine ait hata varyansı değerlerinden daha düşük olduğu görülmektedir.

Genel olarak hata varyansı durdurma kuralının kullanıldığı maddeler-arası çok boyutlu BOB testi analiz yöntemlerine ilişkin bulgular karşılaştırıldığında, Bayesyen MAP yetenek kestirim yöntemine

ait en uygun hata varyansı durdurma kuralının 0,25 olduğu görülmektedir. Ayrıca, D-Optimality madde seçim yöntemi ve içerik ağırlıklandırmasının uygulandığı çok-boyutlu BOB testi uygulamalarında Bayesyen MAP yetenek kestirim yönteminin daha tutarlı sonuçlar verdiği görülmektedir. Diğer taraftan, Fisher'in puanlama madde seçim yöntemlerinden etkilendiği ve D-Optimality madde seçim yöntemi kullanıldığında testteki ortalama madde sayısı 50 olduğu durumda bile güvenilir ve turtarlı sonuçlar vermediği söylenebilir.

Tablo 3'te madde seçim yöntemlerinden seçkisiz (Random) madde seçim yöntemi, durdurma kurallarından hata varyansı, yetenek kestirim yöntemlerinden Fisher'in puanlama ve Bayesyen MAP yetenek kestirim yöntemlerine ait analiz bulgularına yer verilmiştir. Ayrıca her bir koşul için içerik ağırlıklandırmasının kullanıldığı ve kullanılmadığı durumlara ait sonuçlar karşılaştırılarak, içerik ağırlıklandırmasının çok boyutlu BOB testleri üzerindeki etkisi incelenmiştir.

Tablo 3. Seçkisiz (Random) madde seçim yöntemine ait Çok boyutlu BOB testi Bulguları

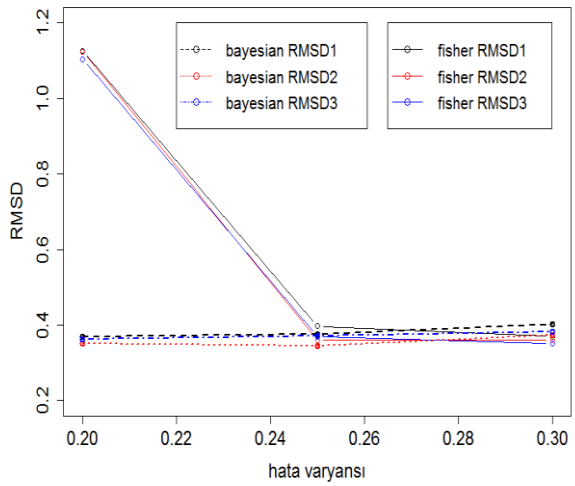
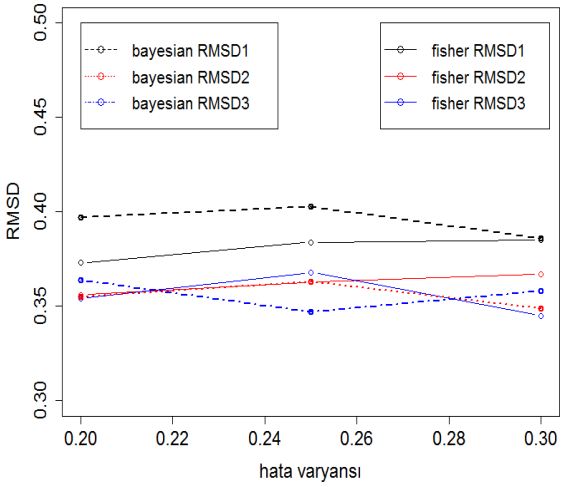
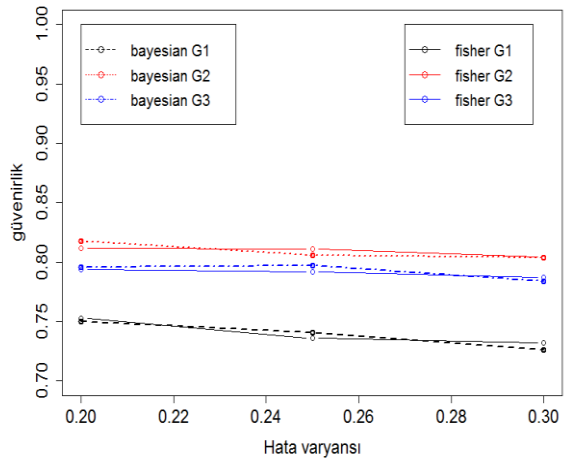
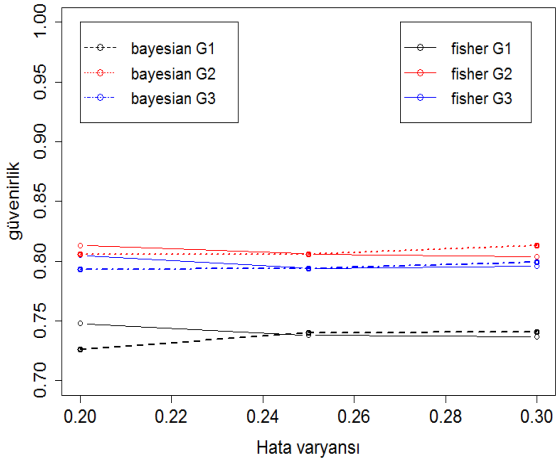
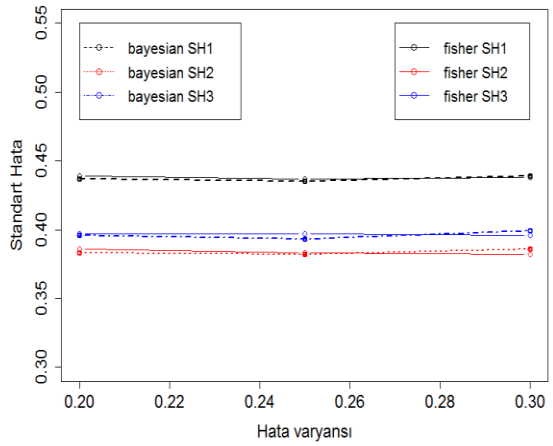
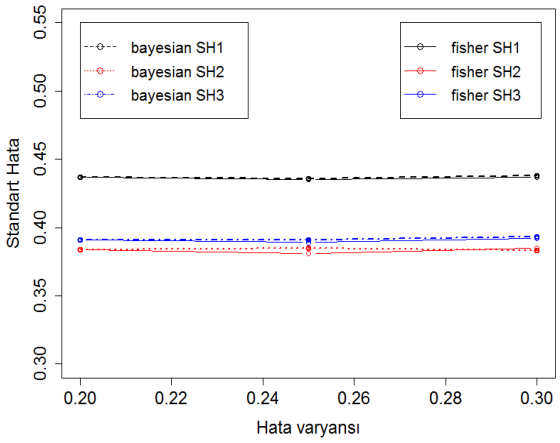
	Yetenek Kestirim yöntemi	Test Sonlandırma kuralı	Test uzunluğu	Güvenirlik			ÖSH			RMSD		
		S. Hata	K	boy1	boy2	boy3	boy1	boy2	boy3	boy1	boy2	boy3
Eşit madde dağılımlı	Fisher	0,20	50,00	0,75	0,81	0,79	0,439	0,386	0,397	1,126	1,124	1,104
		0,25	49,99	0,74	0,81	0,79	0,437	0,383	0,397	0,397	0,360	0,369
		0,30	48,69	0,73	0,80	0,79	0,438	0,382	0,396	0,370	0,361	0,352
	Bayesian	0,20	50,00	0,75	0,82	0,80	0,437	0,383	0,396	0,370	0,352	0,362
		0,25	50,00	0,74	0,81	0,80	0,435	0,382	0,393	0,376	0,344	0,371
		0,30	48,79	0,73	0,80	0,78	0,439	0,386	0,399	0,403	0,373	0,382
İçerik Ağırlıklandırmalı	Fisher	**0,20**	50,00	0,75	0,813	0,805	0,437	0,384	0,391	0,373	0,356	0,354
		0,25	49,97	0,74	0,806	0,794	0,435	0,381	0,389	0,384	0,363	0,368
		0,30	48,83	0,74	0,804	0,796	0,437	0,385	0,392	0,385	0,367	0,345
	Bayesian	0,20	50,00	0,73	0,806	0,793	0,437	0,384	0,391	0,397	0,355	0,364
		0,25	49,98	0,74	0,806	0,794	0,436	0,385	0,391	0,403	0,363	0,347
		0,30	48,30	0,74	0,813	0,799	0,438	0,383	0,393	0,386	0,349	0,358

*OSH= Ölçmenin standart Hatası

Tablo 3'teki analiz bulgularına bakıldığında, çok boyutlu BOB testi sürecinde maddeler her hangi bir kurala bağlı kalmadan seçkisiz (random) olarak seçildiğinde her bir yetenek kestirim yöntemi için hata varyansı durdurma kuralı 0,20'den 0,30'a çıkarıldığında testteki ortalama madde sayısı 48 ile 50 arasında değişkenlik gösterdiği görülmektedir. Ayrıca, diğer madde seçim yöntemleri ile karşılaştırıldığında her bir koşul için güvenirlik katsayılarının düşük, standart hata ve RMSD değerlerinin ise yüksek olduğu görülmektedir. Diğer taraftan, içerik ağırlıklandırmasının yapılmadığı ve yapıldığı duruma ait her bir yetenek kestirimin yönteminin benzer sonuçlar verdiği görülmektedir.

Şekil 4'te içerik ağırlıklandırmasının yapılmadığı eşit madde dağılımlı ve içerik ağırlıklandırmasının yapıldığı duruma ilişkin madde seçim yöntemlerinden seçkisiz madde seçim yöntemi, testi durdurma kurallarından boyutlara ilişkin hata varyansı ve yetenek kestirim yöntemlerinden Fisher'in puanlama ve Bayesyen MAP yetenek kestirim yöntemlerinin kullanıldığı çok-boyutlu BOB testi analizlerine ait grafikler verilmiştir

Şekil 4'te yer alan seçkisiz madde seçim yönteminin kullanıldığı çok boyutlu BOB testlerine ait RMSD değerlerini veren a.1 ve a.2 grafiklerine bakıldığında, Bayesyen MAP yetenek kestirim yöntemi için hata varyansı artmasına rağmen RMSD değerlerinde önemli bir artış gözlenmemektedir. Diğer taraftan, Fisher'in puanlama yetenek kestirim yöntemi kullanıldığında ise boyutlara ilişkin RMSD değerlerinin önce azaldığı daha sonra ise değişmediği görülmektedir.

Eşit madde dağılımlı	*İçerik Ağırlıklandırmalı*
a.1 RMSD- Hata varyansı ilişkisi	a.2 RMSD- Hata varyansı ilişkisi
	
b.1 Güvenirlik- Hata varyansı ilişkisi	b.2 Güvenirlik- Hata varyansı ilişkisi
	
c.1 Ölçmenin standart hatası- Hata varyansı	c.2 Ölçmenin standart hatası- Hata varyansı
	

Şekil 4. Seçkisiz (Random) Madde Seçim Yöntemine İlişkin Grafikler

İçerik ağrılıklandırması uygulandığında ise Fisher'in puanlama yetenek kestirim yöntemlerine ait RMSD değerlerinin daha düşük olduğu görülmektedir. Bayesyen yöntemi için ise hata varyansı arttıkça RMSD değerlerinin azalıp arttığı görülmektedir. Genel olarak, hem Bayesyen MAP hem de Fisher'in puanlama yöntemine ait her bir boyuta ilişkin RMSD değerlerini yüksek olduğu görülmektedir.

Şekil 4'te yer alan seçkisiz madde seçim yönteminin kullanıldığı çok boyutlu BOB testlerine ait her bir boyuta ilişkin güvenirlik katsayıları ile hata varyansı arasındaki ilişkiyi veren b.1 ve b.2 grafiklerine bakıldığında, hata varyansı arttıkça her bir yetenek kestirim yöntemine ait güvenirlik katsayılarının değişmediği ve düşük olduğu görülmektedir. İçerik ağırlıklandırmasının yapıldığı durumda da benzer sonuçlar elde edilmiştir. Bunun temel sebebi, boyutlara ilişkin hata varyansı artmasına karşın testteki ortalama madde sayısının değişmemesidir.

Son olarak, Şekil 4'te çok boyutlu BOB testlerine ait her bir boyuta ilişkin standart hata ile hata varyansı arasındaki ilişkiyi veren c.1 ve c.2 grafiklerine bakıldığında, hem Bayesyen MAP hem de Fisher'in puanlama yönteminin benzer sonuçlar verdiği görülmektedir. Ayrıca, her bir boyuta ilişkin hata varyansı artmasına rağmen standart hata değerlerinin değişmediği görülmektedir. Diğer taraftan, içerik ağırlıklandırması uygulandığında boyutlara ilişkin standart hata değerlerinde her hangi bir iyileşme olmadığı ve eşit madde dağılımı ile benzer sonuçlar verdiği görülmektedir. Bu bulgular doğrultusunda, bireyin yeteneğinin çok boyutlu BOB testleri ile ölçüldüğü durumda, maddelerin seçkisiz olarak seçilmesi yerine, Fisher'in bilgi matrisine dayalı A-Optimality ve D-Optimality madde seçim yönteminin kullanılması daha az madde ile daha yüksek güvenilirlikte ölçümler yapılmasına olanak sağlar.

SONUÇLAR ve TARTIŞMA

Bu çalışmada içerik ağırlıklandırmasının maddeler-arası boyutluluk modeline dayalı çok boyutlu BOB testleri üzerindeki etkisini incelemek amacıyla, farklı madde seçim yöntemleri, yetenek kestirim yöntemleri ve durdurma kuralının kullanıldığı çok boyutlu MTK modellerinden maddeler-arası boyutluluk modellerine dayalı çok-boyutlu bilgisayar ortamında bireyselleştirilmiş (BOB) test (MCAT) yöntemlerinin performansları karşılaştırılmıştır.

Genel olarak, maddeler-arası boyutluluk modeline dayalı çok-boyutlu BOB testi analiz sonuçları karşılaştırıldığında, Fisher'in puanlama yetenek kestirim yöntemi için A-Optimality madde seçim yöntemlerine ait güvenirlik katsayılarının daha yüksek, RMSD ve standart hata değerlerinin ise daha düşük olduğu görülmektedir. Diğer taraftan, Bayesyen MAP yetenek kestirim yönteminin hem A-Optimality hem de D-Optimality madde seçim yöntemi için benzer sonuçlar verdiği görülmektedir. Dolayısıyla, Fisher'in puanlama yönteminin madde seçim yöntemlerinden etkilendiği sonucu çıkartılabilir.

İçerik ağırlıklandırması uygulandığında ise hem A-Optimality hem de D-Optimality madde seçim yöntemi için Bayesyen MAP yetenek kestirim yöntemine ait testteki ortalama madde sayısı az da olsa artarken, boyutlara ilişkin güvenirlik katsayılarının ise değişmediği görülmektedir. Buna karşın boyutlara ilişkin RMSD ve standart hata değerinin düştüğü görülmektedir. Bunun temel sebebi, kâğıt-kalem formatındaki testte her bir boyuta ait maddelerin oranlarının içerik ağırlıklandırması yöntemiyle korunarak, her bir boyut için seçilecek maddelerin ve madde sayılarının sınırlandırılmasıdır. Diğer taraftan, A-Optimality madde seçim yöntemi kullanıldığı çok boyutlu BOB testi yöntemlerinde içerik ağırlıklandırması uygulanmadığında Bayesyen MAP ve Fisher'in puanlama yöntemi farklı sonuçlar verirken, içerik ağırlıklandırması uygulandığında ise benzer ve tutarlı sonuçlar verdiği söylenebilir.

Çok boyutlu BOB testlerinde madde seçim yöntemlerinden hangisinin kullanılması gerektiğini testin amacı belirler. Eğer testin ölçtüğü bütün boyutlar ölçülmek istendiğinde, A-Optimality ve D-Optimality madde seçim yöntemleri en iyi sonucu verir (Mulder ve van der Linden, 2009; Lin, 2012). Dolayısıyla, yetenek kestirim yöntemlerinden Bayesyen MAP yöntemi kullanıldığında, madde seçim yöntemi olarak hem A-Optimality hem de D-Optimality madde seçim yöntemi benzer

sonuçlar vereceğinden tercih edilebilir. Eğer yetenek kestirim yöntemlerinden Fisher'in puanlama yöntemi kullanılacaksa, madde seçim yöntemlerinden A-Optimality yönteminin içerik ağırlıklandırması uygulanarak kullanılması önerilmektedir.

Diao (2009) yapmış olduğu çalışmada, yetenek kestirim yöntemi olarak MLE yönteminin ve madde seçim yöntemlerinden A-Optimality ve D-Optimality yöntemlerinin kullandığında testteki madde sayısını 50 olduğunda A-Optimality ve D-Optimality madde seçim yöntemlerine ait RMSE ve ortalama yanlılık değerlerinin birbirine çok yakın olduğunu belirtmektedir. Ayrıca testteki madde sayısı 50 olduğunda her iki yöntemin benzer sonuçlar verdiğini belirtmektedir. Nitekim bu çalışmada da testteki madde sayısı arttıkça her iki madde seçim yönteminin benzer sonuçlar verdiği bulgusuna ulaşılmıştır. Ayrıca bireyin yeteneğinin çok boyutlu BOB testleri ile ölçüldüğü durumda, maddelerin seçkisiz olarak seçilmesi yerine, Fisher'in bilgi matrisine dayalı A-Optimality ve D-Optimality madde seçim yönteminin kullanılması daha az madde ile daha yüksek güvenilirlikte ölçümler yapılmasına olanak sağlar.

Çok-boyutlu BOB testlerine ilişkin yapılan çalışmalara bakıldığında, A-optimality ve D-optimality madde seçim yöntemlerinin karşılaştırıldığı birçok çalışma yapılmıştır (Segall, (1996); Luecht, (1996); van der Linden,1999; Mulder ve van der Linden, 2009; Diao, 2009; Diao ve Reckase; 2009; Yoo, 2011; Lin, 2012). Genel olarak D-optimality yönteminin daha avantajlı olduğu ve daha yaygın olarak kullanıldığı belirtmektedir (Berger ve Veerkamp, 1996; Passo, 2007). Bunun temel sebebi olarak D-optimality madde seçim yönteminin daha güvenilir ve daha kararlı sonuçlar verdiği belirtilmektedir. Ancak bu yapılan çalışmalara bakıldığında sadece madde-içi boyutluluk modelinin kullanıldığı görülmektedir. Bu çalışmada ise maddeler-arası boyutluluk modeline dayalı çok-boyutlu BOB testlerinde A-optimality ve D-optimality madde seçim yöntemlerinin Bayesyen MAP yetenek kestirim yöntemi kullanıldığında benzer sonuçlar verdiği bulgusuna ulaşılmıştır.

Genel olarak, maddeler-arası boyutluluk modeline dayalı çok-boyutlu BOB testi analiz sonuçları karşılaştırıldığında, aynı madde seçim yöntemi ve durdurma kuralı koşulları altında Bayesyen MAP yöntemine ait güvenirlik ve kestirilen yetenek parametrelerine ilişkin korelasyon değerlerinin daha yüksek, RMSD ve standart hata değerlerinin ise daha düşük olduğu görülmektedir. Dolayısıyla her bir madde seçim ve durdurma kuralı için maddeler-arası boyutluluk modelinde Bayesyen MAP yetenek kestirim yönteminin daha az madde ile daha güvenilir sonuçlar verdiği yorumu yapılabilir. Bu bulgular doğrultusunda, maddeler-arası boyutluluk modeline dayalı çok-boyutlu BOB testi yöntemlerinde her bir madde seçim yöntemi için en uygun yetenek kestirim yönteminin Bayesyen MAP yetenek kestirim yöntemi olduğu söylenebilir

Bayesyen yetenek kestirim yöntemlerini MLE yetenek kestirim yöntemlerine göre üstün kılan en önemli özelliği, Segall (1996)'in de belirttiği gibi, Bayesyen yöntemlerin boyutlar arasındaki korelasyon ve yetenek parametrelerine ait önsel dağılım bilgisini kullanarak kestirim yapmasıdır (Diao ve Reckase, 2009). Bundan dolayı Bayesyen yöntemleri gerçek θ değerine daha çabuk yakınsar ve daha az madde ile daha güvenirlik ve kararlı kestirimler yapar. Buna karşın, kestirilen yetenek parametreleri hakkında yeterli bilgi olmadığı veya belirlenen önsel parametreler zayıf olduğu durumda, Bayesyen yöntemler ile kestirilen yetenek parametreleri yanlılık gösterir.

Hata varyansı durdurma kuralına ilişkin maddeler-arası boyutluluk modeline dayalı A-optimality madde seçim yöntemi ve Bayesyen MAP yetenek kestirim yönteminin kullanıldığı çok-boyutlu BOB testlerinde en uygun hata varyansı durdurma kuralının 0,25 olduğu görülmektedir. Bu durumda maddeler-arası boyutluluk modeli için testi alan her bir bireyin test sürecinde cevaplamış olduğu ortalama madde sayısı 22,8'e eşit olurken, içerik ağırlıklandırması yapıldığında testteki ortalama madde sayısı 29,6'ya yükseldiği görülmektedir.

Hata varyansı durdurma kuralı daha güvenilir ve etkili olmasına karşın bireylerin test sürecinde farklı sayıda maddelere cevap vermesi testin adil olmadığı algısını oluşturabilir (Gershon, 2005). Bu yüzden eğitimdeki gerçek BOB testi uygulamalarında genellikle sabit madde sayısı durdurma kuralı kullanılır (Yoo, 2011). Sabit madde sayısı durdurma kuralının hata varyansı durdurma kuralına tercih edilmesinin bir diğer nedeni ise test süresi ve testten sıkılma gibi koşulların her bir birey için standartlaştırılmasına olanak sağlamasıdır (Segall, 2004). Buna karşın, Rizavi ve Swaminathan

(2001) sabit madde sayısı durdurma kuralının kullanıldığı durumlarda testteki madde sayısı bireyin yeteneğini güvenilir bir şeklide ölçmek için yeterli olmadığında problemlere neden olacağını ve testin güvenirlik ve geçerliğini düşüreceğini savunmuştur.

Maddeler-arası boyutluluk modeline dayalı çok-boyutlu BOB testi yöntemlerinin kullanıldığı durumda madde seçim yöntemlerinden ve madde düzeyindeki boyutluluk modellerinden daha az etkilendiği için Bayesyen MAP yetenek kestirim yönteminin kullanılması önerilmektedir. Ayrıca, çok-boyutlu BOB testi yöntemleri için testteki madde sayısının az olduğu ve boyutlar arasındaki korelasyonun yüksek olduğu durumlarda Bayesyen MAP yetenek kestirim yönteminin kullanılması önerilmektedir.

Sonuç olarak, kâğıt-kalem testleri ile karşılaştırıldığında hem maddeler-arası modeline dayalı çok-boyutlu BOB testlerinin daha az madde ile daha yüksek güvenirlikte ölçümler yaptığından benzer formattaki testlerin çok-boyutlu bilgisayar ortamında bireyselleştirilmiş testler ile uygulanması önerilmektedir.

Bu çalışmada, gerçek veriye dayalı simülasyon yöntemi (post-hoc simulation method) kullanılarak, testin üç boyutlu olduğu durumda farklı yetenek kestirimi, madde seçim ve durdurma kurallarına ilişkin çok-boyutlu BOB testi sonuçları karşılaştırılmıştır. Farklı simülasyon çalışmaları yapılarak, testin ölçtüğü boyut sayısının, boyutlara ilişkin madde havuzu büyüklüğünün ve farklı çok boyutlu modellerin çok-boyutlu BOB testi yöntemleri üzerindeki etkisi incelenebilir.

Bu çalışmada, çok-boyutlu BOB testi sürecinde her bir boyut için sorulacak madde sayısını ve testin formatını kontrol altında tutmak için kâğıt-kalem testi formatındaki her bir boyuta ait madde oranına bağlı olarak içerik ağırlıklandırması (content balancing) yapılmıştır. Gelecekte yapılacak çalışmalarda, daha önce geliştirilmiş olan içerik ağırlıklandırması yöntemleri (örn. Düzgünleştirilmiş-alfa deseni, Chang, Qian ve Ying, 2001; Sympson ve Hetter's yöntemi, Sympson ve Hetter 1985) kullanılarak BOB testi süreci daha gerçekçi hale getirilebilir. Ayrıca, çok-boyutlu BOB testlerinde kullanılan madde kullanım sıklığı ve içerik ağırlıklandırması yöntemlerinin birbirini nasıl etkilediği üzerinde çalışmalar yapılabilir.

Ülkemizde tek boyutlu BOB testi uygulamalarına yönelik çalışmalar olmasına karşın, çok-boyutlu BOB testlerinin gerçek hayatta uygulamalarına ilişkin bir çalışma henüz yapılmamıştır. Bu çalışma sonuçları ışığında gerçek çok-boyutlu BOB testi uygulamaları yapılması önerilmektedir. Ayrıca bu çalışmada bireylerin sadece İngilizce dil becerilerinin çok-boyutlu BOB testleri ile ölçülmesi amaçlanmıştır. Matematik ve fen bilgisi gibi alanlarda da benzer çalışmalar yapılabilir.

KAYNAKÇA

Berger, M.P. F., & Veerkamp, W. J. J. (1996). A review of selection methods for optimal tests design. In M. Wilson & G. Engelhard (Eds.), Objective measurement: Theory into practice (Vol. 3, pp. 437-455). Norwood, NJ: Ablex

Bloxom, B., & Vale, C.D. (1987).Multidimensional adaptive testing: An approximate procedure for updating. In *Meeting of the psychometric society*. Montreal, Canada, June.

Chang, H.-H., Qian, J. and Ying, Z. (2001). A-stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement,* 25, 333-341.

Choi, S. W. & King D. R. (2011). *MAT: Multidimensional adaptive testing.* [Çevirim içi: https://cran.r-project.org/web/packages/MAT/MAT.pdf], Erişim tarihi: 15 Temmuz 2015.

Diao, Q. (2009). *Comparison of ability estimation and item selection methods in multidimensional computerized adaptive testing*. Unpublished Doctoral Dissertation. Michigan State University.

Diao, Q. & Reckase, M. (2009). Comparison of ability estimation and item selection methods in multidimensional computerized adaptive testing. In: Weiss DJ (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. pp. 1-13.

Fan, M., & Hsu, Y. (1996). Multidimensional computer adaptive testing. *In Annual meeting of the American educational research association*. New York City, NY, April.

Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement* 6:109-27.

Green, B. G., Bock, R.D., Humphries, L. G., Linn, R.L., & Reckase, M.D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21,* 347-360

IACAT Official Web Site. [Çevirim-içi: http://iacat.org/content/research-strategies-cat]. Erişim tarihi: 24 Aralık 2015.

Lin, H. (2012). *Item selection methods in multidimensional computerized adaptive testing adopting polytomously-scored items under multidimensional generalized partial credit model*. Unpublisjed Doctoral Dissertation. University of Illinois at Urbana-Champaign.

Lord, F. M. (1971a). Tailored testing, an approximation of stochastic approximation. Journal of the American Statistical Association, 66, 707–711.

Lord, F. M. (1971b). A theoretical study of the measurement effectiveness of flexilevel tests. *Educational and Psychological Measurement, 31,* 805–813.

Lord, F. M. (1971c). A theoretical study of two-stage testing. *Psychometrika, 36,* 227–242.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement, 20* (4), 389-404.

McBride, J.R. & Martin, J.T. (1983). Reliability and Validity of Adaptive Ability Tests in a military setting. in Weiss D.J. (Ed.) *"New Horizons in Testing"* New York: Academic Press.

Mulder, J., & van der Linden, W. J. (2009). Muldimensional adaptive testing with optimal design criteria for item selection. *Psychometrika, 74* (2), 273-296.

Passos, V. L., Berger, M. P. F., & Tan, F. E. (2007). Test design optimization in CAT early stage with the nominal response model. Applied Psychological Measurement, 31, 213-232.

Rizavi, S. & Swaminathan, H. (2001). The effect of test and examinee characteristics on the occurrence of aberrant response patterns in a computerized adaptive test. *Paper presented at the annual meeting of the American Educational Research Association*, Seattle WA. (2001)

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*(2), 331-354.

Segall, D.O. (2000). Principles of multidimensional adaptive testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 53–73). Boston: Kluwer Academic.

Segall, D. O. (2001). General ability measurement: An application of multidimensional item response theory. *Psychometrika*, 66, 79-97.

Silvey, S.D. (1980). *Optimal design*. London: Chapman & Hall.

Sympson, J.B. and Hetter, R.D. (1985, october). Controlling item exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association (pp. 973-977).* San Diego, CA: Navy Personnel Research and Development Center.

van der Linden, W.J. (1996). Assembling tests for the measurement of multiple traits. *Applied Psychological Measurement, 20,* 373–388.

van der Linden, W.J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics, 24*, 398–412.

van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice.* Boston: Kluwer.

van der Linden, W. J. & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer.

Veldkamp, B. P. , & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika, 67*(4), 575-588.

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B., Mislevy, R., et al. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Hillsdale, NJ: Erlbaum.

Wang, W. C. & Chen, P.H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement* 2004 28: 295. DOI: 10.1177/0146621604265938.

Wang, W., Chen, P., & Cheng, Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, 9, 116–136.

Wang, W.-C., Wilson, M., and Adams, R. (1997). Rasch models for multidimensionality between items and within items. In G. Englehard, Wilson, Mark (Ed.), *Objective Measurement* (Vol. 4,): Greenwich, CN: Ablex Publishing.

Weiss, D. J., & Betz, N. E. (1973). *Ability measurement: Conventional or adaptive? (Research Report 73-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.*

Weiss, D.J., & Kingsbury, G.G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21:4 361-375.

Yoo, H. (2011). *Evaluating several multidimensional adaptive testing procedures for diagnostic assessment*. Unpublished Doctoral Dissertation. University of Massachusetts Amherst)

EXTENDED ABSTRACT

Developments in computer technologies not only affect our social life, environment and our life styles but also affect our education systems, measurement and evaluation tools and learning methods. These developments also provide new methods to measure students' different abilities. Along with these developments, stakeholders in education benefit from computer technologies so as to provide education in higher standards. As a result of these developments, new measurement and evaluation methods that utilize computer technologies has been developed to measure students' abilities or skills. An example of this is computer-based tests (CBT) that uses computer environment instead of paper-pencil tests. Another alternative measurement method is computerized adaptive testing (CAT) methods in which students abilities are tailored to items' features by means of a computer program (McBride and Martin, 1983; Weiss and Kingsbury, 1984).

There are a lot of advantages of measuring students' abilities with CAT methods compared to paper&pencil tests. One of the most important advantage of CAT is that it enable us to measure students' abilities with shorter tests and higher reliabilities (Wainer, 1993). İn addition, it provides flexible testing time and provides result of test to examinees as soon as test is terminated (Lin, 2012).

Purpose of study

The purpose of this study is to compare the performance of Between-item dimensionality model-based Multidimensional CAT designs and to examine the effect of content balancing on different MCAT designs.

Methods

For this purpose, real data set from English Proficiency Test (EPT) administered by Hacettepe University was used to create multidimensional item pool, in which each test consist of three dimensions listening, reading and grammar respectively. In this study, 10 EPT data sets, which consist of 628 items in total, administered between 2009 and 2013 were used to construct item pool. Number of items in each test ranged from 59 to 64 and administered to students ranged between 1200 and 2000. Item parameters were estimated with compensatory multidimensional 2 parameter logistic model (CM-2PLM) based on between-item dimensionality model. After item calibration, 73 items with low discrimination parameters and items which had difficulty parameters out of the [-4, 4] interval were excluded from the item pool. Finally, multidimensional item pool consist of 555 items of which 240 items are related to grammar, 115 items are related to listening and 200 items are related to reading dimension. For each MCAT algorithm, examinee ability parameters (θ) were drawn from the multivariate standard normal distribution with the population variance-covariance matrix and the number of examinee was restricted to 500.

In order to determine the best MCAT algorithm for the EPT, two different theta estimation methods (MLE based Fisher scoring and Bayesian MAP), three different item selection methods (fisher information based A-optimality and D-optimality, Random item selection) and two different termination methods (fixed number of item, precision) were used. In addition, results of MCAT algorithms with content distribution and without content distribution were compared. Results of these conditions were compared with respect to, reliability index, RMSE, averaged number of items administered and RMSD values between full bank theta and estimated MCAT theta.

Results and Discussion

Results indicated that using different theta estimation and item selection methods affected RMSE, averaged number of items administered and RMSD values for each MCAT algorithm. When Bayesian MAP ability estimation method was used both A-optimality and D-optimality yielded similar results with respect to reliability coefficients, SEM and RMSD values. On the other hand, A-

optimality item selection method outperformed both D-optimality and non-adaptive random item selection methods when MLE based Fisher's scoring ability estimation method was used. In multidimensional case, there are several studies investigating A- and D-optimality for dichotomous MIRT models. Segall (1996), Luecht (1996), and Mulder and van der Linden (2009) used the D-optimality. van der Linden (1999) studied A-optimality. As Mulder and van der Linden (2009) stated, A-optimality and D-optimality yield the most accurate estimates in which all measured abilities were intentional.

Using A-optimality rather than D-optimality to select items both decreased average number of items administered and RMSD values between true theta and estimated theta, and increased test reliability. Overall, MCAT design with A-optimality and Bayesian theta estimation method outperformed other MCAT designs.

The comparison of MCAT with content balancing and without content balancing showed that content balancing yielded more accurate and consistent results. In addition, using content balancing yielded higher reliability coefficients with shorter test. For each MCAT condition SEM and RMSD statistics associated with each dimension tended to decrease when content distribution was taken into account. Therefore, there was a trade-off between reliability coefficients and error statistics associated with each dimension. All in all, post-hoc simulation based on MCAT with content balancing for EPT provided ability estimations with higher reliability with fewer items compared to paper and pencil format. Results of this study would also provide an important guideline for live MCAT application of EPT.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2015; 6(2); 385-399

Maslach Tükenmişlik Envanteri-Eğitimci Formu'nu Türkçe'ye Uyarlama Çalışması*

The Adaptation Study of Maslach Burnout Inventory-Educators Survey to Turkish

Nuri Barış İNCE ** Ali E. ŞAHİN ***

Öz
Bu çalışmada Maslach ve Jackson tarafından geliştirilen Maslach Tükenmişlik Envanteri'nin bazı değişikliklerle Maslach, Jackson ve Schwab tarafından eğitimcilere uyarlanması sonucunda elde edilen Maslach Tükenmişlik Envanteri-Eğitimci Formu'nu Türkçe'ye uyarlama çalışmasının yapılması amaçlanmıştır. Araştırmanın çalışma grubunu Ankara ilinin merkez ve taşra ilçelerinde çalışan 760 sınıf öğretmeni oluşturmaktadır. Özgün formu 7'li Likert tipinde 22 maddeden ve 3 boyuttan (duygusal tükenme, duyarsızlaşma ve kişisel başarı) oluşan envanterin Türkçe formunun uygulandığı çalışma grubunda yapılan geçerlik ve güvenirlik çalışmaları envanterin özgün yapısının korunduğunu göstermektedir. Her bir alt boyut için ayrı ayrı hesaplanan Cronbach Alpha katsayısı duygusal tükenme boyutu için 0.88, duyarsızlaşma boyutu için 0.78, kişisel başarı boyutu için 0.74 olarak bulunmuştur. Madde-toplam korelasyonları dikkate alındığında envanterde yer alan maddelerin iyi bir ayırt ediciliğe sahip olduğu anlaşılmaktadır. Yapı geçerliği için uygulanan doğrulayıcı faktör analizi sonucunda elde edilen AGFI, GFI, RMSEA, CFI ve NFI gibi uyum indeksleri doğrulamak için kurulan modelin kabul edilebilir uyum düzeyine sahip olduğunu göstermektedir.

Anahtar Kelimeler: tükenmişlik, sınıf öğretmeni, ölçek uyarlama.

Abstract
In this study, it is aimed to make Turkish adaptation study of Maslach Burnout Inventory-Educators Survey, which was obtained as a result of adaptation of Maslach Burnout Inventory, developed by Maslach and Jackson, to the educators with some modifications by Maslach, Jackson and Schwab. The study group of the research consisted of 760 classroom teachers working in central district and provinces of Ankara. Validity and reliability studies of the scale, of which the original form consists of 22 items in 7 point Likert type and 3 dimensions, (emotional exhaustion, depersonalization and personal accomplishment), applied to the study group in Turkish form indicated that the original structure of the scale has been preserved. Cronbach's alpha coefficient calculated separately for each sub-dimension was determined as 0.88 for emotional exhaustion, 0.78 for depersonalization and 0.74 for personal accomplishment. Considering the item-total correlations, the items in the scale are understood to have a good discrimination indice. Model-fit indices such as AGFI, GFI, RMSEA, CFI and NFI obtained as a result of confirmatory factor analysis applied for the construct validity show that the model established for compliance have acceptable level of model-fit.

Keywords: burnout, classroom teacher, scale adaptation.

GİRİŞ

Son yıllarda popüler bir kavram olarak birçok araştırmaya konu olan tükenmişlik, 1970'li yıllarda Freudenberger tarafından akademik bir kavram olarak kullanılmıştır. Araştırmacı "ilk olarak gönüllü sağlık çalışanları arasında görülen yorgunluk, enerji kaybı, hayal kırıklığı, güdülenme eksikliği ve işi

* Bu çalışma, ilk yazarın ikinci yazar yönetiminde hazırladığı ve 09.07.2014 tarihinde tamamlanan "Birleştirilmiş ve Bağımsız Sınıf Öğretmenlerinin Mesleki Doyum ve Tükenmişlik Düzeylerinin Karşılaştırılması" isimli yüksek lisans tezinin bir kısmıdır.

** Arş. Gör., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, e-posta: nbaris.ince@hacettepe.edu.tr
*** Doç. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, e-posta: alisahin@hacettepe.edu.tr

Geliş Tarihi: 05.10.2015
Kabul Tarihi: 24.12.2015

bırakmayla ilgili bir durumu tanımlamak için" tükenmişlik kavramını ortaya atmıştır (Avcı ve Seferoğlu, 2011, s. 13). Aynı yıllarda Maslach ve meslektaşları Freudenberger'den bağımsız olarak Kaliforniya'da insanlarla yüz yüze olmayı gerektiren mesleklerde çalışanların bilişsel stratejiler kullanarak duygusal sorunları ile nasıl başa çıktıklarını araştırmıştır (Schaufeli, Leiter ve Maslach, 2009). Yetmişli yıllardan günümüze gelene kadar tükenmişlik konusunda birçok araştırma gerçekleştirilirken sadece 1993 ile 2004 yılları arasında 1784 makale, kitap ve tez yazılmıştır (Hallesleben ve Buckley, 2004).

Tükenmişlik kavramı Freudenberger (1974, s. 159) tarafından "enerji, güç ve kaynaklar üzerindeki aşırı taleplerden dolayı kişinin başarısız olması, yıpranması ve tükenmiş hale gelmesi" şeklinde tanımlanmış olsa da Maslach ve arkadaşlarının tükenmişlik kavramına ilişkin yaptıkları tanım ve geliştirdikleri model oldukça yaygın olarak kabul edilmektedir. Maslach'a (2003) göre tükenmişlik, iş yerindeki stres yapıcı unsurlara karşı bir tepki olarak uzun sürede ortaya çıkan psikolojik bir sendrom, çalışan ile işi arasındaki uyumsuzluğun sonucu olan kronik bir gerginlik durumudur. Maslach, tükenmişlik kavramını duygusal tükenme, duyarsızlaşma ve kişisel başarı olmak üzere üç boyutu olan bir süreç şeklinde değerlendirmektedir. Maslach modeline göre insanların psikolojik taleplerine daha fazla cevap veremeyecek konuma gelen çalışanlar ilk olarak duygusal tükenme yaşamaktadır (Maslach ve Jackson, 1981). İş ortamındaki aşırı taleplerle başa çıkamayan çalışanlar hizmet verdikleri insanlara karşı mesafe alarak duyarsızlaşırken; içinde bulundukları topluma ve çalıştıkları kuruma yapmayı bekledikleri katkı ile mevcut davranışları arasında uyumsuzluk olduğunu fark etmekte ve kişisel başarılarının yetersiz olduğunu düşünmektedir (Cordes ve Dougherty, 1993).

Duygusal, zihinsel ve fiziksel sonuçları olan tükenmişlik kavramı mumun eriyerek sönmesi metaforunda olduğu gibi insanın enerjisinin kalmamasıyla açıklanmaktadır (Schaufeli ve diğerleri, 2009). Çalışanların hizmet verdikleri insanlarla ilişkileri bağlamında yaptıkları içsel değerlendirmeler sonucu ortaya çıkan tükenmişliğin olumsuz sonuçları hem çalışanlara hem de hizmet verdikleri insanlara yansımaktadır. Hayes ve Weathington (2007) bireylerin işlerine ilişkin konularda nadiren hissedilen yorgun ve sinirli olma veya bunalımda hissetme gibi duygusal durumların seyrek olmaktan çıkarak sürekli hale gelmesi durumunu tükenmişlik olarak açıklamaktadır. Cherniss de benzer biçimde tükenmişliğin geçici bir yorgunluk veya zorlanma durumu değil, bireyin işinden soğuması ile sonuçlanan kalıcı bir durum olduğunu düşünmektedir (aktaran, Avşaroğlu, Deniz ve Kahraman, 2005).

İnsanların fiziksel ihtiyaçları, tutkuları, arkadaşları ve aileleri ile ilişkileri üzerinde etkili bir kavram olan tükenmişlik sürece bağlı bir gelişim izlemektedir. Çalışanlar, iş ve aile ortamında girdikleri etkileşimler sonucunda tükenmişliğe yönelik bazı fiziksel, zihinsel, duygusal ve davranışsal belirtiler göstermektedir. Tükenmişlik yaşayan çalışanlarda; enerji düşüklüğü, kronik yorgunluk, güçsüzlük, bitkinlik, uyku problemleri, sıklıkla tekrarlanan soğuk algınlığı ve baş ağrısı gibi fiziksel belirtiler gözlenmektedir (Weisberg ve Sagie, 1999; Lambie, 2007). Zihinsel düzeyde ise kendisine, işine ve yaşamına karşı olumsuz tutumlar geliştiren, mesleğinde yetersiz ve başarısız olduğuna inanan insanlar çalışma ortamındaki değişikliklere uyum sağlamakta zorlanmaktadır. Tükenmişlik belirtileri gösteren çalışanlar zihinsel değişimlerin yanı sıra duygusal düzensizlikler de yaşamaktadır. Ani öfke patlamaları ve sinir krizleri geçiren, kolaylıkla ağlayan, depresyon, kaygı ve huzursuzluk yaşayan çalışanlar tükenmeye başladıklarına ilişkin duygusal belirtiler göstermektedir (Freudenberger, 1974; Friedman, 1991). Bireylerin çalışma ortamında saldırganca davranışlarda bulunması, iş değiştirmesi veya devamsızlık yapması, madde kullanması ise tükenmişliğin davranışsal belirtileri olarak öne çıkmaktadır (Guglielmi ve Tatrow, 1998; Lambie, 2007).

Yaşamın önemli bir bölümünü çalışarak geçiren bireylerin gösterdikleri tükenmişlik belirtilerini ortaya çıkaran bireysel ve örgütsel pek çok faktör bulunmaktadır. Cinsiyet, yaş, medeni durum, hizmet yılı gibi demografik özellikler ile çalışanların kişilik özellikleri gibi bireye özgü olan pek çok faktör tükenmişliği etkileyen bireysel faktörlerdendir. Uyumluluk, çalışkanlık, dışadönüklük, nevrotizm (duygusal dengesizlik) ve açıklık olmak üzere bireylere ait beş kişilik özelliği ile tükenmişlik arasındaki ilişkileri inceleyen araştırmalardan elde edilen bulgulara göre dışadönük,

çalışkan ve uyumlu kişilik özellikleri tükenmişlik riskini azaltırken, nevrotik kişilik özellikleri ise tükenmişliğe yakalanma riskini arttırmaktadır (Wallin, 2010).

Tükenmişlik düzeyini etkileyen bireysel faktörlerden birisi olan hizmet yılı göz önüne alınarak sınıf öğretmenlerinin tükenmişlik düzeylerindeki gelişim seyrini belirlemeyi amaçlayan bir çalışmada öğretmenlerin duygusal tükenme, duyarsızlaşma ve kişisel başarı boyutlarındaki tükenmişlik düzeylerinin dördüncü yıla doğru artış gösterdiği, sekizinci yıla doğru ise düştüğü gözlenmiştir (Gökçakan ve Murat, 2007). Kuzey Kıbrıs'ta görev yapan ilkokul öğretmenlerinin tükenmişlik düzeylerinin çeşitli değişkenler ile ilişkisinin belirlenmesinin amaçlandığı araştırmada ise öğretmenlik mesleğini gerçekten isteyerek seçmeyen öğretmenlere göre öğretmenlik yapmayı gerçekten isteyenlerin duygusal tükenme ve kişisel başarı boyutlarındaki tükenmişlik düzeylerinin anlamlı şekilde daha düşük olduğu tespit edilmiştir (Ozan, 2009). Sınıf öğretmenlerinin sınıf içi disiplin anlayışları ile tükenmişlik düzeyleri arasındaki ilişkinin araştırıldığı bir başka araştırmada katı disiplin anlayışına sahip olan öğretmenlerin tükenmişlik ölçeğinin her bir alt boyutunda demokratik disiplin anlayışına sahip olan öğretmenlere göre daha yüksek düzeyde tükenmişlik gösterdiği anlaşılmıştır (Tümkaya, 2005). Sınıf ve branş öğretmenlerinin öz yeterlilikleri ile tükenmişlik düzeyleri arasındaki ilişkilerin araştırıldığı bir çalışmanın bulguları öz yeterlilik değişkenleri olan öğrenci sorumluluğu, öğretim stratejileri ve sınıf yönetimi ile tükenmişliğin duygusal tükenme ve duyarsızlaşma boyutu arasında düşük, kişisel başarı boyutu arasında orta düzeyde ters yönlü anlamlı bir korelasyon olduğunu göstermiştir (Bümen, 2010).

Tükenmişliğe neden olan bireysel faktörlerin yanı sıra iş yükü, kontrol, ödüller, çalışma arkadaşları, adalet ve değerler olarak öne çıkan örgütsel faktörler de çalışanların tükenmişlik düzeylerinin farklılaşmasına neden olmaktadır (Leiter ve Maslach, 2005). Tükenmişliği etkileyen örgütsel faktörlerden biri olan iş yükü, çalışanların sahip olduğu kaynakları ve zamanı aşan iş talepleri olarak tanımlanmaktadır. Duygusal tükenme boyutunu doğrudan etkileyen iş yükünde önemli olan noktanın aşırı talepler nedeniyle bireyin enerjisinin tükenerek kendisini yenilemesinin imkânsız hale gelmesi olduğu ifade edilmektedir (Leiter ve Maslach, 2003). Çalışanların işlerini etkili bir şekilde yapmaları için gereken kaynaklar ve alınan kararlar üzerinde etkili olması anlamına gelen kontrol faktörü tükenmişlik düzeyini etkileyen örgütsel faktörlerden bir diğeridir. Lee ve Ashforth (1993) çalışanların karar alma süreçlerine aktif katılımları ile yüksek düzeyde kişisel başarı ve düşük düzeyde duygusal tükenme arasında tutarlı ilişkiler olduğunu tespit etmiştir.

Örgütlerin çalışanlarına yönelik ödüllendirme faaliyetlerinde bulunması tükenmişlik düzeylerinin belirlenmesi açısından önem taşımaktadır. Birçok çalışan için ödül veya ceza almak işleri nasıl yaptıklarını bilmenin ve kurumdaki diğer insanların kendileri hakkında ne düşündüğünü öğrenmenin bir yolu olarak görülmektedir (Schwab, Jackson ve Schuler, 1986). Çalışanların kurumlarından beklediği ödüllendirmeler sadece mali konularla sınırlı kalmamakta örgütsel ve sosyal boyutları da bulunmaktadır. Önemli bir işi başaran çalışanın kendisi ile gurur duyması içsel; maaş ve ek ödemeler ise dışsal ödüllendirmeler olarak nitelendirilmektedir (Maslach ve Goldberg, 1998). Tükenmişlik düzeyini etkileyen örgütsel faktörlerden biri olan bireylerin çalışma arkadaşları ile ilişkileri çalışanlar arasındaki sosyal iletişim ve etkileşimin kalitesine vurgu yapmaktadır. Bireyler arasında korku ve düşmanlığın hüküm sürdüğü, kurum politikalarının ve hedeflerinin şeffaf olmadığı, çalışanlar arasında karşılıklı desteğin sağlanmadığı örgütlerde insanların tükenmişlik yaşama olasılıkları yüksektir. Çalışanların iş yükü üzerinde etkisi olan yöneticilerden sağlanan destek duygusal tükenme, çalışma arkadaşlarından sağlanan destek ise kişisel başarı boyutu ile ilişkili gözükmektedir (Maslach ve Leiter, 2008).

Bireylerin çalışma ortamında karşılanmasını beklediği adalet, tükenmişlik düzeyine etki eden örgütsel faktörlerden bir diğeridir. Örgütsel adalet, çalışanların işte alınan kararların adil ve eşit olduğuna inanmasıdır (Leiter ve Maslach, 2003). Çalışanların örgütsel adalet algısında yöneticilerin önemli bir rolü bulunmaktadır. Çalışanlarını destekleyen ve adil davranan yöneticilerin bulunduğu örgütlerde görev yapan bireyler daha az tükenmişlik yaşamakta ve büyük örgütsel değişimlere fazla direnç göstermemektedir (Leiter ve Harvie, 1998). Toplumsal yaşamda insan ilişkilerinin merkezinde yer alan değerler de çalışma ortamındaki insanlar için oldukça önemlidir. Örgütsel yaşamda çalışanlar bazen kendi değerleri ile kurumun beklentileri arasında kalmaktadır. Çalışanlar

kendi değerlerine uymayan ve etik olmayan işler yapmak durumunda kaldıklarında iş güvencesine sahip olmak ile dürüst davranmak arasında gidip gelmektedir (Maslach ve Goldberg, 1998). Bireylerin sahip olduğu değerlerin çalışma ortamında paylaşılmaması örgütsel bağlılığı zayıflatmakta ve tükenmişlik düzeylerini etkilemektedir.

Sağlık, emniyet ve hukuk gibi çalışma alanlarında olduğu üzere insanlarla yüz yüze çalışan öğretmenlerin tükenmişliğinin nedenlerini bireylerin psikolojik özellikleri ve çalıştıkları örgütlerin çevresinde yer alan değişkenlerle açıklamaya çalışan araştırmalar bulunmaktadır. Argon ve Ateş (2007) sınıf öğretmenlerinin etkilendikleri stres faktörlerini ortaya çıkartarak, sosyal çevreden kaynaklanan faktörlerin ilk sırayı aldığını bunu fiziksel çevre, kendini yorumlama ve iş çevresi ile ilgili faktörlerin izlediğini bulmuştur. Kokkinos'un (2007) Kıbrıs'ta çalışan ilkokul öğretmenlerinin tükenmişliğinin işe ilişkin stres yapıcılar ve kişilik özellikleri ile bağlantısını araştırdığı çalışmada elde edilen bulgular duygusal tükenme ve duyarsızlaşmanın daha çok çevresel stres yapıcılar ile ilişkili olduğunu; kişisel başarı boyutunun ise kişilik özellikleri ile açıklandığını göstermiştir. Yapılan analizlere göre sınıf düzenini yönetme ve disiplini sağlama etkenlerinden kaynaklanan stresin artması duygusal tükenme ve duyarsızlaşmayı anlamlı düzeyde yordamıştır.

Friesen ve Sarros (1989) "Eğitimcilerdeki Tükenmişliğin Kaynakları" isimli çalışmada öğretmenlerdeki tükenmişliği yordayan iş doyumu değişkenlerini değerlendirmiştir. Araştırmanın bulguları genel iş stresinin öğretmenlerin duygusal tükenmişliklerini yordayan en önemli değişken olduğunu göstermiştir. İşin zorluğu öğretmenlerin duyarsızlaşma düzeylerini, statü ve tanınmadan elde edilen doyum ise kişisel başarı boyutunu yordayan değişkenler olarak öne çıkmıştır. Cano-Garcia, Padilla-Munoz ve Carrasco-Ortiz (2005) ise öğretmenlerin tükenmişliğini yordayan değişkenleri belirlemek amacıyla gerçekleştirdikleri araştırmanın analizlerine göre öğretmenlerin duygusal tükenmişlik düzeylerine ait toplam varyansın %72'sini açıklayan değişkenler olarak yöneticilerle ilişkilerin yetersizliği, yüksek düzeyde duygusal dengesizlik düzeyinde bulunmak, terfi imkânlarının eksik olması, mesleki saygınlığın az olduğunun farkında olmak, uzun bir süre işteki aynı pozisyonda kalmak ve öğrenci sayısı olduğunu tespit etmişlerdir.

Çalışanların tükenmişlik düzeylerini ölçmek amacıyla uluslararası alanda geliştirilen bazı araçların Türkçe uyarlama çalışmalarının yapıldığı bilinmektedir. Bu araçlardan birisi olan "Tükenmişlik Ölçeği Kısa Versiyonu" (The Burnout Measure Short Version) Pines tarafından 2005 yılında geliştirilmiş Tümkaya ve Çavuşoğlu (2010) tarafından sınıf öğretmeni adaylarının tükenmişlik düzeylerini belirlemek amacıyla kullanılmıştır. Bu ölçeğin yanı sıra "Kopenhag Tükenmişlik Ölçeği" (Copenhagen Burnout Inventory) Kristensen, Borritz, Villadsen ve Christensen tarafından 2005 yılında geliştirilmiş ve çeviri çalışmaları Bakoğlu Deliorman, Taştan Boz, Yiğit ve Yıldız (2009) tarafından yapılarak akademik personelin tükenmişlik düzeylerinin ölçülmesinde kullanılmıştır. Uluslararası alanyazında sıklıkla kullanılan veri toplama araçlarından biri olan "Maslach Tükenmişlik Envanteri" (MTE) (Maslach Burnout Inventory) Maslach ve Jackson tarafından 1981 yılında insana hizmet veren mesleklerde çalışan bireylerin tükenmişlik düzeylerini belirlemek için geliştirilmiştir. Ergin (1993) doktorların ve hemşirelerin tükenmişlik düzeylerini incelemek üzere MTE'yi Türkçe'ye uyarlamıştır.

1986 yılına gelindiğinde ise Maslach ve arkadaşları tükenmişlik envanterinin eğitimci formunu geliştirmiştir. Maslach Tükenmişlik Envanteri-Eğitimci Formu ile ulusal literatürde yapılan çalışmalara bakıldığında veri toplama aracının Türkçe'ye uyarlama çalışmalarının Günseli Girgin (1995) ve Asuman Baysal (1995) tarafından gerçekleştirildiği ifade edilmektedir (Girgin ve Baysal, 2005). Her iki araştırmada da benzer şekilde birlikte geçerlik tekniğinin kullanıldığı ifade edilmiş ve öğretmenlerin kendilerinin tükenmişlik formuna verdikleri yanıtlarla, yakın iş arkadaşlarının araştırmaya katılan öğretmenleri değerlendirdikleri tükenmişlik formuna verdikleri yanıtlar karşılaştırılmıştır. Öğretmenlerin tükenmişlik düzeylerini belirlemek için sonrasında yapılan ulusal çalışmalarda da bu araştırmalara atıfta bulunularak herhangi bir geçerlik çalışması yapılmamış sadece güvenirlik sonuçlarının sunulmasıyla yetinilmiştir (Gündüz, 2005; Girgin, 2010; Ertürk ve Keçecioğlu, 2012). Dolayısıyla Uluslararası Test Komisyonunun ölçeklerin adaptasyonuna yönelik yayınlamış olduğu bağlam, test geliştirme ve adaptasyonu, uygulama ve sonuçların yorumlanmasına ilişkin rehber ilkelerin (Internatioal Test Commission, 2005) dikkate alınması ve çok değişkenli

istatistik tekniklerinin kullanılmasıyla Maslach Tükenmişlik Envanteri-Eğitimci Formu'nu Türkçe'ye uyarlama çalışmasının yapılması amaçlanmıştır.

YÖNTEM

Çalışma Grubu

Nicel yaklaşımlardan betimsel yöntemin benimsendiği bu araştırmada Ankara ilindeki devlet ilkokullarında görev yapan sınıf öğretmenleri çalışmanın ulaşılabilir evrenini meydana getirmektedir. Ankara İl Milli Eğitim Müdürlüğü istatistiklerine göre 2012-2013 eğitim-öğretim yılında Ankara'da 15544 sınıf öğretmeni görev yaparken araştırmanın çalışma grubu bu öğretmenlerden seçilmiştir. 5'i merkez 9'u taşra olmak üzere 14 ilçede bulunan okullar basit seçkisiz yöntemle seçilmiş ve bu okullarda görev yapan sınıf öğretmenleri çalışma grubuna alınmıştır. Bu şekilde ulaşılan 220 öğretmenden toplanan verilerle araştırmanın pilot uygulaması gerçekleştirilmiştir. Ardından 540 öğretmenden elde edilen veriler ise asıl uygulama olarak değerlendirilmiştir. Araştırmanın geçerlik ve güvenirlik analizleri her iki uygulama için de yapılmıştır.

Asıl uygulama grubunda yer alan öğretmenlerin demografik bulguları incelendiğinde öğretmenlerin 402'si kadın (%74.4), 138'i (%25.6) erkektir. Öğretmenler yaş gruplarına göre incelendiğinde 30 yaş ve altında olan 138 (%25.6) öğretmen, 31-40 yaş aralığında olan 213 (%39.4) öğretmen, 41-50 yaş aralığında olan 152 (%28.1) öğretmen ile 51 yaş ve üstünde 37 (%6.9) öğretmen bulunmaktadır. Medeni duruma göre 475 (%88) öğretmen evli, 65 (%12) öğretmen ise bekârdır. Hizmet yılı değişkeninde ise 10 yıl ve altı tecrübeye sahip 210 (%38.9) öğretmen, 11 ile 20 yıl arasında 246 (%45.6) öğretmen, 21 ile 30 yıl arasında 63 (%11.7) öğretmen, 31 yıl ve üzerinde deneyime sahip 21 (%3.9) öğretmen asıl uygulama grubuna ait örneklemde yer almaktadır. Öğretmenlerin 509'u (%94.3) kadrolu olarak görevlerini yerine getirirken 31 öğretmen (%5.7) ücretli statüsünde çalışmaktadır. Ayrıca 87 (%16.1) öğretmenin uzman öğretmen unvanı ile çalıştığı anlaşılmaktadır. Kendilerine mezun oldukları program sorulan öğretmenlerin 346'sı (%64.1) sınıf öğretmenliği programından, 194'ü ise sınıf öğretmenliğinden farklı bir programdan (%35.9) mezun olduklarını ifade etmiştir.

Veri Toplama Aracı

Maslach ve Jackson, polisler, öğretmenler, hemşireler ve psikiyatristler gibi insanlarla yüz yüze iletişim gerektiren sağlık ve hizmet sektöründe görev yapan 1025 kişinin yer aldığı bir örneklem üzerinden tükenmişlik envanterini geliştirme çalışmalarında bulunmuştur. Yapılan açımlayıcı ve doğrulayıcı faktör analizleri sonucunda duygusal tükenme, duyarsızlaşma ve kişisel başarı olmak üzere üç boyutun bulunduğu 22 maddeden oluşan 7'li Likert tipi MTE'yi geliştirmişlerdir. Bireyin işindeki duygusal taleplere cevap veremeyecek duruma gelmesi ölçeğin duygusal tükenme boyutunu; çalışanın hizmet verdiği kişiler ile arasına belli bir mesafe koyması ve onları görmezden gelmesi duyarsızlaşma boyutunu; bireyin kendini işinde başarılı bulması ise kişisel başarı boyutunu meydana getirmektedir (Maslach, Schaufeli ve Leiter, 2001). Bu envanterde duygusal tükenme boyutu 9 madde, duyarsızlaşma boyutu 5 madde ve kişisel başarı boyutu ise 8 maddeden oluşmaktadır.

Maslach, Jackson ve Schwab, 1981 yılında Maslach ve Jackson tarafından geliştirilen MTE'yi küçük değişikliklerle eğitimcilere uyarlayarak MTE-EF'yi elde etmiştir. Hizmet sektöründeki çalışanların tükenmişlik düzeylerini ölçmek amacıyla geliştirilen özgün formdaki soru maddelerinde bulunan ve hizmeti alan kimse anlamına gelen *"recipient"* kelimesi yerine eğitimcilerin tükenmişlik düzeylerini belirlemek amacıyla uyarlanan formda öğrenci anlamına gelen *"student"* kelimesi kullanılmaktadır (Maslach, Jackson ve Leiter, 2010). Envanterde yer alan duygusal tükenme, duyarsızlaşma ve kişisel başarı boyutları ayrı ayrı puanlanmaktadır. Elde edilen yüksek duygusal tükenme ve duyarsızlaşma puanları bireyin yüksek düzeyde tükenmişlik yaşadığını göstermektedir. Kişisel başarı boyutunda alınan puanların düşük olması ise kişinin işinde karşılaştığı aşırı talepler nedeniyle kendisini yetersiz hissettiğini ve yüksek düzeyde tükenmişlik yaşadığını göstermektedir.

Envanterin puanlama anahtarında duygusal tükenme boyutundan alınabilecek en yüksek puan 54, kişisel başarı boyutundan 48, duyarsızlaşma boyutundan ise 30 puandır. Kişisel başarı boyutunda alınan yüksek puanların karşılığı düşük düzeyde yaşanan tükenmişliktir. Bir başka ifade ile bu boyutta düşük puan alan çalışanlar yüksek düzeyde tükenmişlik yaşamaktadır. Kişisel başarı boyutunun puanlanması ve yorumlanması arasındaki ters yönlü ilişkiden dolayı bu boyut kişisel başarısızlık hissi veya azalan kişisel başarı olarak da nitelendirilmektedir. Envanterin her bir maddesinden alınabilecek en düşük puan "0", en yüksek puan ise "6" olarak belirlenmiştir. Bu derecelendirmeye göre "0-Hiçbir zaman", "1-Yılda birkaç kez", "2-Ayda bir kez", "3-Ayda birkaç kez", "4-Haftada bir kez", "5-Haftada birkaç kez", "6-Her gün" olarak düzenlenen puanlama seçenekleri araştırmanın veri toplama sürecinde de aynı şekilde kullanılmıştır. Envanterin üç boyutundan alınabilecek puan aralıkları ve bu aralıkların karşılık geldiği düzeyler Tablo 1'de yer almaktadır.

Tablo 1. MTE-EF Puanlama Anahtarı

		Tükenmişlik	
	Düşük Düzey	Orta Düzey	Yüksek Düzey
Duygusal Tükenme	0-16	17-26	27 ve üzeri
Duyarsızlaşma	0-8	9-13	14 ve üzeri
Kişisel Başarı	37 ve üzeri	31-36	0-30

Kaynak: Maslach, Jackson ve Leiter (2010).

Maslach ve Jackson (1981) MTE'nin geçerliğini belirlemeye yönelik birkaç yöntem kullanmıştır. İlk olarak çalışanların iş arkadaşları veya eşlerinden bireylerin davranışlarını değerlendirmeleri istenmiştir. İş arkadaşları ve eşleri tarafından davranışları değerlendirilen çalışanların elde ettikleri puanlar ile tükenmişlik düzeyleri arasında anlamlı bir ilişki olduğu tespit edilmiştir. Yapılan değerlendirmelere göre duygusal tükenmişlik ve duyarsızlaşma düzeyi yüksek çalışanların işlerinde mutsuz ve hizmet verdikleri kişilerden şikâyetçi oldukları anlaşılmaktadır.

Maslach ve diğerlerine (2010) göre MTE'nin geçerlik çalışması için kullanılan başka bir yöntem Hackman ve Oldham'ın İş Betimleme Ölçeğinin bazı boyutları ile MTE'nin boyutları arasında ilişkinin olup olmadığının incelenmesine yöneliktir. İşten elde edilen geri bildirim ve işin anlamlılığı ile tükenmişlik envanterinin boyutları arasında anlamlı bir ilişki bulunmuştur. Üçüncü bir yöntem olarak iş'teki yetiştirme ve geliştirme fırsatları, çalışanların işini anlamlı bulması, iş'ten ayrılma niyeti, iş'te daha az zaman geçirme isteği, iş'te ve iş dışında insan ilişkilerinin bozulması, aile ve arkadaşlarla yaşanan sorunlar ile uykusuzluk, alkol ve madde kullanımının artması ile çalışanların tükenmişlik düzeyleri arasında anlamlı ilişkiler tespit edilmiştir. Ayrıca MTE'nin ayırt edici geçerliğini ortaya koymak için envanterin boyutları ile Crowne-Marlowe'un Sosyal İstenirlik Ölçeği arasında anlamlı bir ilişki olup olmadığına bakılmış ve 0.05 düzeyinde anlamlı bir ilişki olmadığı bulunmuştur.

Maslach ve Jackson (1981) MTE'nin güvenirliğini Cronbach alfa katsayısının hesaplanması ve test-tekrar test yönteminin kullanılması ile belirlemiştir. Envanterin iç tutarlılığını tahmin etmek için hesaplanan alfa katsayısının duygusal tükenme boyutunda 0.89, duyarsızlaşma boyutunda 0.77 ve kişisel başarı boyutunda ise 0.74 olduğu tespit edilmiştir. Sağlık kuruluşlarında çalışan yöneticiler ve sosyal hizmetlerde okuyan öğrencilere uygulanan test-tekrar test yöntemi sonucunda duygusal tükenme boyutu için 0.82, duyarsızlaşma boyutu için 0.60 ve kişisel başarı boyutu için 0.80 güvenirlik katsayısı elde edilmiştir. Öğretmenlerin tükenmişlik düzeylerini belirlemek için Iwanicki ve Schwab'ın 465 öğretmen ile gerçekleştirdiği çalışmada Cronbach alfa güvenirlik katsayıları duygusal tükenme boyutu için 0.90, duyarsızlaşma ve kişisel başarı boyutu için 0.76 olarak belirlenmiştir (Iwanicki ve Schwab; aktaran, Maslach ve diğerleri, 2010).

MTE-EF'nin geliştirildiği kültürde elde edilen bulgular dikkate alınarak geçerli ve güvenilir bir ölçme aracı olduğuna karar verilmiştir. Uyarlama çalışmasına geçmeden önce envanterin yazarlarından olan Christina Maslach ile iletişime geçilmiştir. Kendisinin yönlendirmesi üzerine

"mindgarden.com" adresiyle gerekli yazışmalar yapılarak ve telif ücreti ödenerek (Ek-1) tükenmişlik envanterinin uyarlama çalışmasına başlanmıştır. Buna göre envanterin İngilizce olan özgün formunda yer alan maddeler araştırmacılar tarafından Türkçe'ye çevrilerek bir form hazırlanmış ve uzman görüşüne başvurulmuştur. Çeviri formuna envanterin özgün dilindeki maddeler ile Türkçe çevirileri yazılarak ifadelerin uygun olup olmadığı sorulmuştur. Çeviri ifadelerinin uygun olmadığını düşünen uzmanların düzeltme önerileri verebilecekleri bir alanın bırakılmasına özen gösterilmiştir. Hazırlanan çeviri formu Eğitim Yönetimi, Sınıf Öğretmenliği ve Yabancı Diller Eğitimi bölümünde görev yapan 10 akademisyene elektronik posta ya da yüz yüze yapılan görüşmeler yoluyla ulaştırılmıştır. Yabancı Diller Eğitimi alanından uzmanların dönütleri Türkçe'ye çevrilen ifadelerin İngilizce'ye geri çevrilmesi yoluyla alınmıştır. Uzmanların tamamından elde edilen geri bildirimler sonucunda uzlaşı sağlanamadığı tespit edilen maddeler üzerinde düzeltmeler yapılmıştır. Her bir madde için uzmanlardan alınan dönütlerin toplamı çevirinin uygun olmadığı yönünde belirgin bir eğilim ortaya koyduğunda veya çevirinin uygun olduğu ya da olmadığı yönünde dengeli bir durum ortaya çıkardığında gerekçeli kararlarını ifade eden uzmanların önerileri dikkate alınarak düzeltme işlemi gerçekleştirilmiştir.

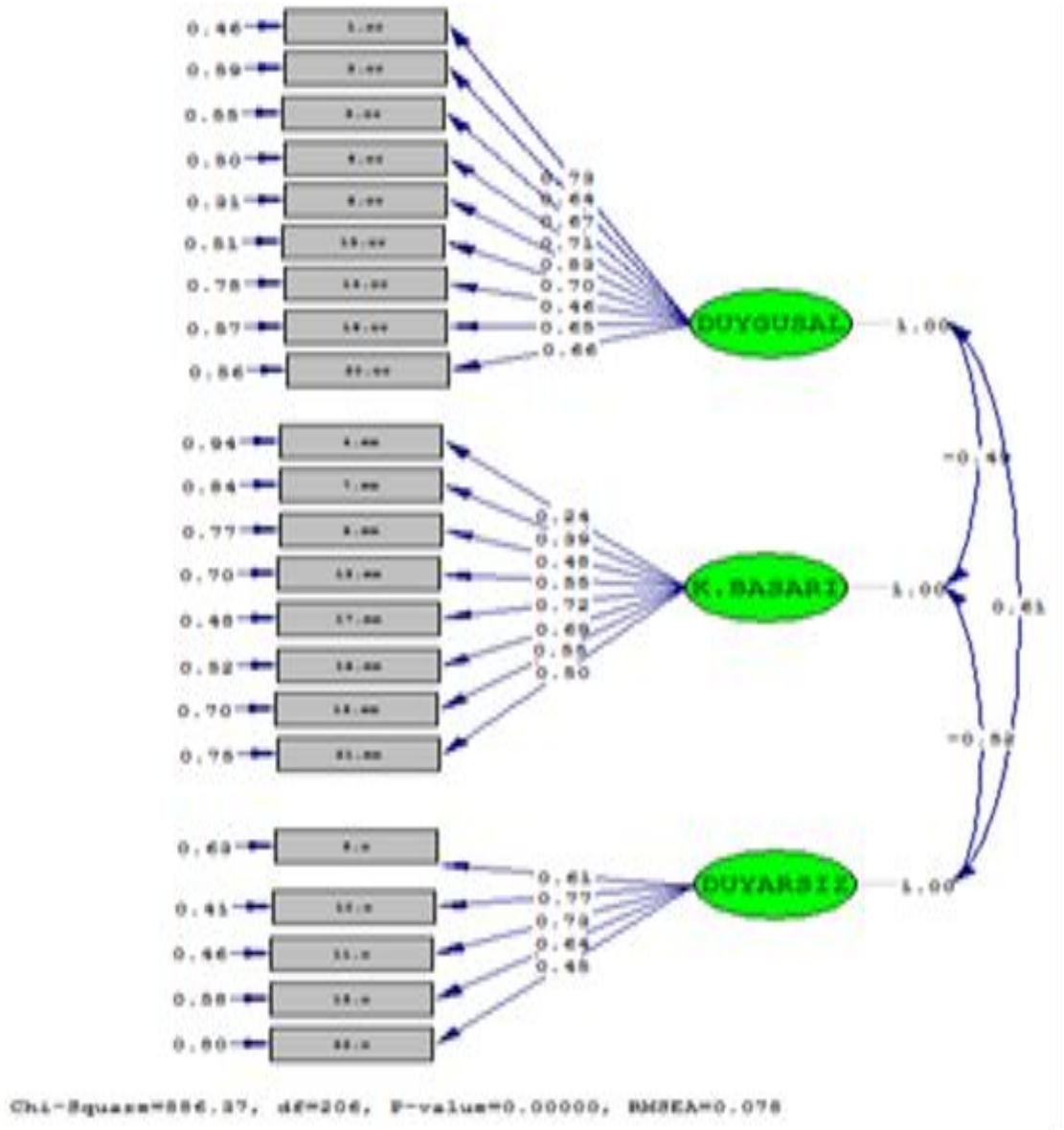
İşlem

Veri toplama aracını uygulamaya başlamadan önce 2 Aralık 2013-23 Ocak 2014 tarihleri arasında hafta içi mesai saatlerini kapsayan bir çalışma takvimi oluşturulmuştur. Mesleki tükenmişlik zamana bağlı ortaya çıkan bir kavram olarak değerlendirilmektedir. Bu nedenle öğretmenlerin dönem içerisinde bir süre mesleklerini icra etmeleri beklenmiş ve veri toplama aracının uygulama zamanı olarak dönem sonuna denk gelen aylar tercih edilmiştir. Uygulamalara önce merkez ilçelerden başlanmış daha sonra taşra ilçelerde çalışan sınıf öğretmenlerinden veri toplanarak çalışma tamamlanmıştır. Öncelikle okul müdürleri ile görüşme gerçekleştirilmiş, araştırmanın amaç ve içeriğinden bahsedilerek uygulama için alınan resmi izin yazıları sunulmuştur. Okul müdürlerinden alınan onayın ardından öğretmenlerin ders araları beklenmiştir. Öğretmenlere yapılan gerekli açıklamalardan sonra gönüllülük esasına dayalı olarak öğretmenler odasında ya da dersliklerde veri toplama aracının uygulanmasına geçilmiştir. Uygulama formunda yer alan maddelere yönelik katılımcıların sorularına uygun bir şekilde cevap verilmesine özen gösterilmiştir. Araştırmanın veri toplama aşaması belirtilen tarihler arasında planlandığı şekilde tamamlanmıştır.

BULGULAR

Geçerlik

MTE-EF'nin yapı geçerliğini belirlemek için Linear Structural Relations (LISREL) 8.80 programı kullanılarak doğrulayıcı faktör analizine başvurulmuştur. MTE-EF için oluşturulan yol analizi diyagramı incelendiğinde envanterin p değerinin 0.01 düzeyinde manidar olduğu görülmektedir. Anlamlı farklılığın olmaması gereken "p" değerinin örneklem büyüklüğü nedeniyle manidar çıkmasının hoşgörü ile karşılanabileceği düşünülmektedir (Çokluk, Şekercioğlu ve Büyüköztürk, 2012). Kişisel başarı boyutunda yer alan 4. maddenin 0.94 ile yüksek düzeyde bir hata varyansına sahip olduğu anlaşılmaktadır. Ancak söz konusu maddenin "t" değeri ise manidar çıkmıştır. Tükenmişlik envanterinin üç boyutu arasındaki korelasyon değerlerinin orta düzeyde yer alması nedeniyle farklı özellikleri ölçmekte oldukları yorumu yapılabilir. Envanterin özgün formundaki üç boyutlu faktör yapısına sadık kalınarak asıl uygulama verileri üzerinden yapılan faktör analizine ilişkin yol analizi diyagramına Şekil 1'de yer verilmektedir.



Şekil 1. MTE-EF Yol Analizi Diyagramı

MTE-EF'ye uygulanan doğrulayıcı faktör analizi sonucunda envanterin uyum düzeylerini gösteren değerler hesaplanmıştır. Envanterin uyum indekslerinin karşılaştırıldığı iyi uyum ve kabul edilebilir uyum ölçütleri Schumacker ve Lomax (2010) ile Schermelleh-Engel, Moosbrugger ve Müller'in (2003) çalışmalarından derlenerek pilot ve asıl uygulamadan elde edilen verilerle birlikte Tablo 2'de sunulmaktadır.

Tablo 2. MTE-EF Uyum İndeksleri

Uyum indeksi	İyi Uyum	Kabul Edilebilir Uyum	Pilot Uygulama	Asıl Uygulama
x^2/sd	$0 \leq x^2/sd \leq 2$	$2 \leq x^2/sd \leq 3$	1.5	4.3
AGFI	$0.90 \leq AGFI \leq 1.00$	$0.85 \leq AGFI \leq 0.90$	0.86	0.84
GFI	$0\ .95 \leq GFI \leq 1.00$	$0.90 \leq GFI \leq 0.95$	0.88	0.87
RMSEA	$0.00 \leq RMSEA \leq 0.05$	$0.05 \leq RMSEA \leq 0.08$	0.05	0.07
CFI	$0.97 \leq CFI \leq 1.00$	$0.95 \leq CFI \leq 0.97$	0.96	0.94
NFI	$0.95 \leq NFI \leq 1.00$	$0.90 \leq NFI \leq 0.95$	0.90	0.93

Tablo 2 incelendiğinde pilot ve asıl uygulamadan elde edilen ki-kare ile serbestlik derecesi oranının farklılaştığı dikkat çekmektedir. Ki-kare ile serbestlik derecesi oranının iki uygulama grubu arasında farklı olması örneklem büyüklüğünden kaynaklanmaktadır. Çokluk ve diğerleri (2012) ki-kare değerinin örneklem büyüklüğüne karşı duyarlı olduğunu örneklem büyüdükçe ki-kare değerinin de artacağını belirtmektedir. Araştırmanın asıl uygulama grubunda yer alan öğretmenlerin sayısı pilot uygulama olarak değerlendirilen diğer gruptaki öğretmenlerin sayısına göre yaklaşık iki buçuk kat daha fazladır. Tabloda yer alan uyum indekslerinin tamamı (x2/sd=4.3, RMSEA=0.07, CFI=0.94, NFI=0.93, GFI=0.87, AGFI=0.84) dikkate alındığında doğrulamak için kurulan faktör modeli genel olarak kabul edilebilir bir uyum düzeyi göstermektedir.

Güvenirlik

MTE-EF'nin güvenirlik analizleri madde-toplam korelasyonlarının belirlenmesi ve Cronbach alfa iç tutarlılık katsayılarının hesaplanması yöntemi ile gerçekleştirilmiştir. Envanterde yer alan maddelerin ölçülmek istenen özelliklere sahip olan ve olmayan bireyleri ayırt edip etmediğini belirlemek amacıyla elde edilen madde-toplam korelasyonlarına ilişkin verilere Tablo 3'te yer verilmektedir.

Tablo 3. MTE-EF Madde-Toplam Korelasyonlarının Dağılımları

	Madde-Toplam Korelasyonları	
Maddeler	Pilot Uygulama	Asıl Uygulama
I. Duygusal Tükenme		
1. Öğretmenlikten duygusal olarak soğuduğumu hissediyorum.	0.62	0.66
2. Okulda günü bitirdiğimde kendimi bitkin hissediyorum.	0.58	0.63
3. Sabah kalkıp yeni bir iş gününe başlamam gerektiğinde kendimi yorgun hissediyorum.	0.61	0.63
6. Bütün gün öğrencilerle çalışmak beni gerçekten zorluyor.	0.61	0.66
8. Öğretmenliğin beni tükettiğini hissediyorum.	0.79	0.76
13. Öğretmenlik mesleğinin beni hayal kırıklığına uğrattığını düşünüyorum.	0.64	0.63
14. Öğretmenlikte iş yükümün çok fazla olduğunu hissediyorum.	0.51	0.45
16. Öğrencilerle çalışıyor olmak beni oldukça strese sokuyor.	0.58	0.61
20. Öğretmenliğe daha fazla dayanamayacakmışım gibi hissediyorum.	0.61	0.58
II. Duyarsızlaşma		
5. Bazı öğrencilere sanki nesnelermiş gibi davrandığımı hissediyorum.	0.55	0.54
10. Öğretmenliğe başladığımdan beri öğrencilere karşı daha çok duyarsızlaştım.	0.46	0.65
11. Öğretmenliğin beni duygusal olarak katılaştırdığını düşünüyorum.	0.43	0.59
15. Bazı öğrencilere ne olduğunu gerçekten umursamıyorum.	0.48	0.55
22. Öğrencilerin bazı sorunlarından dolayı beni suçladıklarını hissediyorum.	0.39	0.40
III. Kişisel Başarı		
4. Öğrencilerimin bir konu hakkında ne hissettiğini kolayca anlayabiliyorum.	0.23	0.22
7. Öğrencilerimin sorunlarıyla çok etkin bir şekilde ilgileniyorum.	0.33	0.37
9. Bir öğretmen olarak öğrencilerin yaşamlarını olumlu bir şekilde etkilediğimi hissediyorum.	0.38	0.46
12. Kendimi çok zinde hissediyorum.	0.34	0.40
17. Rahat bir çalışma ortamını öğrencilerimle birlikte kolayca yaratabiliyorum.	0.44	0.59
18. Öğrencilerimle iç içe gerçekleştirdiğim bir çalışmadan sonra içimin coşkuyla dolduğunu hissediyorum.	0.44	0.57
19. Öğretmenlikte kayda değer pek çok şey başardım.	0.42	0.48
21. İşimde karşılaştığım duygusal problemlerle oldukça sakin bir şekilde baş ediyorum.	0.27	0.41

Madde-toplam korelasyonlarına ilişkin Tablo 3 incelendiğinde envanterde yer alan maddelerin iyi bir ayırt edicilik düzeyine sahip olduğu görülmektedir. Büyüköztürk (2010, s.171) madde-toplam korelasyonu 0.30 ve daha yüksek olan maddelerin bireyleri iyi derecede ayırt ettiğini, 0.20 ile 0.30 arasında kalan maddelerin zorunlu görülmesi durumunda teste alınabileceğini belirtmektedir. Elde edilen bulgular envanterin iyi derecede ayırt edici maddelere sahip olduğunu göstermektedir.

Cronbach alfa iç tutarlılık katsayılarının hesaplanması sonucunda elde edilen güvenirlik bulgularına ise Tablo 4'te yer verilmektedir.

Tablo 4. MTE-EF Güvenirlik Katsayıları

Boyutlar	Pilot Uygulama	Asıl Uygulama
Duygusal Tükenme	0.88	0.88
Duyarsızlaşma	0.71	0.78
Kişisel Başarı	0.67	0.74

MTE-EF'nin güvenirlik katsayılarının verildiği Tablo 4 incelendiğinde pilot ve asıl uygulamanın her ikisinde de envanterin duygusal tükenme boyutunun 0.88 ile aynı güvenirlik düzeyinde bulunduğu tespit edilmiştir. Duyarsızlaşma ve kişisel başarı boyutları ele alındığında ise asıl uygulama için hesaplanan güvenirlik katsayıları pilot uygulama verilerine göre daha yüksek çıkmıştır.

SONUÇLAR ve TARTIŞMA

Bu çalışmada MTE-EF'nin faktör yapısının Türkiye'deki sınıf öğretmenlerinden elde edilen verilerle ne derece uyumlu olduğunu incelemek amaçlanmıştır. Çalışma grubundaki öğretmenlerden elde edilen verilere uygulanan doğrulayıcı faktör analizi sonuçları MTE-EF'nin özgün yapısının korunduğunu ortaya koymuştur. Elde edilen uyum indeks değerleri (x2/sd=4.3, RMSEA=0.07, CFI=0.94, NFI=0.93, GFI=0.87, AGFI=0.84) kabul edilebilir uyum ölçütlerini karşılamıştır. Envanterde yer alan maddelerin faktör yük değerlerinin 0.24 ile 0.83 arasında değişkenlik gösterdiği tespit edilmiştir. Çalışmada elde edilen güvenirlik katsayıları asıl uygulama için duygusal tükenme boyutunda 0.88, duyarsızlaşma boyutunda 0.78, kişisel başarı boyutunda ise 0.74'dür. Ergin (1993) benzer şekilde MTE için yapmış olduğu faktör analizlerinde envanterin üç faktörde yığıldığını tespit etmiştir. Envanterin güvenirlik düzeyleri Cronbach alfa iç tutarlılık katsayısı ile duygusal tükenme boyutu için 0.83, duyarsızlaşma boyutu için 0.65, kişisel başarı boyutu için 0.72 olarak hesaplanmıştır. Test-tekrar test yöntemi ile hesaplanan katsayıların ise duygusal tükenme boyutunda 0.83, duyarsızlaşma boyutunda 0.72, kişisel başarı boyutunda ise 0.67 düzeyinde olduğu belirlenmiştir. Ergin'in çalışmasından elde edilen bulgular ile MTE-EF'nin güvenirlik düzeylerine ilişkin bu araştırmadan elde edilen bulgular arasında tutarlılık gözükmektedir.

Envanterde yer alan maddelerin madde-toplam korelasyon değerlerinin de 0.37 ile 0.76 arasında değiştiği sadece 4. maddenin 0.30'un altında kaldığı anlaşılmıştır. Kişisel başarı boyutunda yer alan 4. madde pilot ve asıl uygulama grubundan elde edilen verilere göre kabul edilebilir düzeyde bir korelasyon katsayısına sahiptir. "Öğrencilerimin bir konu hakkında ne hissettiğini kolayca anlayabiliyorum." maddesinin hata varyansının da yüksek çıktığı göz önüne alındığında ilgili maddenin formu yanıtlayan öğretmenler tarafından yeterince açık bir şekilde anlaşılmamış olabileceğini düşündürmektedir. Dördüncü maddede geçen *"bir konu"* ifadesi ile öğretmenlerin anlayabilecekleri konunun öğrencilerin akademik meselelerine ilişkin mi yoksa günlük yaşam sorunlarına yönelik mi olduğu konusunda kararsız kalmış olabilecekleri gündeme gelmektedir. Burada bir ayrıma gitmek yerine bu maddenin öğrencilerin karşılaşabilecekleri her durumu kapsayıcı bir yapıda olması amaçlanmaktadır. Dolayısıyla kişisel başarı boyutunda yer alan 4. madde, *"Öğrencilerimin herhangi bir konu hakkında ne hissettiğini kolayca anlayabiliyorum."* şeklinde yeniden düzenlenerek envanterde yer alması sağlanabilir.

Uluslararası alanyazında Maslach Tükenmişlik Envanteri-Eğitimci Formu'nun çeşitli kültürlerde yapılan uyarlama çalışmalarının bulguları ile bu araştırmadan elde edilen bulguların tutarlılık gösterdiği söylenebilir. Chen ve meslektaşları (2014) Maslach Tükenmişlik Envanteri-Eğitimci Formu'nun Malezya sürümündeki uyarlama çalışmasında Cronbach alpha güvenirlik katsayılarını duygusal tükenme boyutu için 0.91, duyarsızlaşma ve kişisel başarı boyutu için 0.78 olarak bulmuştur. Ayrıca envantere uyguladıkları temel bileşenler analizi sonucunda maddelerin faktör yüklerinin 0.24 ile 0.86 arasında değiştiğini tespit etmişler ve envanterin özgün formundaki gibi üç faktörde yığıldığını görmüşlerdir. Başka bir uyarlama çalışmasında ise Schwarzer, Schmitz ve Tang (2000) Almanya ve Hong-Kong'daki öğretmenler ile Maslach Tükenmişlik Envanteri-Eğitimci Formu'nun geçerlik çalışmalarını yapmıştır. Yapılan faktör analizi, envanterin Hong Kong sürümünün faktör yük değerlerinin 0.36 ile 0.87 arasında; Almanya sürümünün ise 0.36 ile 0.82 arasında olduğunu ve her iki kültürde de envanterin maddelerinin üç faktörde yığıldığını göstermiştir. Araştırmanın güvenirlik analizlerinde envanterin Hong Kong sürümü için duygusal tükenme boyutunda 0.88, duyarsızlaşma boyutunda 0.79 ve kişisel başarı boyutunda 0.83 katsayıları elde edilmiştir. Almanya sürümünde ise duygusal tükenme, duyarsızlaşma ve kişisel başarı boyutu olmak üzere sırasıyla 0.88, 0.69 ve 0.82 katsayıları bulunmuştur. Kokkinos (2006) ise Kıbrıs Rum Kesiminde görev yapan 771 öğretmenden topladığı verilerle Maslach Tükenmişlik Envanteri-Eğitimci Formu'nun psikometrik özelliklerini ve faktör yapılarını ortaya koyan bir çalışma gerçekleştirmiştir. Araştırmanın geçerliği için açıklayıcı ve doğrulayıcı faktör analizi yapmış ve envanterin özgün formundaki gibi üç boyutlu olduğunu ortaya koymuştur. Açıklayıcı faktör analizinde gözlenen faktör yüklerinin 0.32 ile 0.84 arasında değiştiğini tespit eden araştırmacı, çalışmanın doğrulayıcı faktör analizinde ise envanterin üç faktörlü modeline ilişkin uyum indekslerini (x2/sd=4.7, CFI=0.83, RMSEA=0.08, SRMR=0.08) elde etmiştir. Envanterin güvenirlik analizlerinde ise Cronbach alpha katsayılarının duygusal tükenme boyutunda 0.85, duyarsızlaşma boyutunda 0.63 ve kişisel başarı boyutunda 0.79 olduğunu bulmuştur.

Araştırmadan elde edilen bulgular genel olarak değerlendirildiğinde Maslach Tükenmişlik Envanteri-Eğitimci Formu'nun öğretmenlerin tükenmişlik düzeylerinin belirlenmesinde kullanılabileceği söylenebilir. Çalışmanın verileri Ankara ilinde çalışan sınıf öğretmenlerinden toplandığı için envanterin geçerlik ve güvenirlik çalışmaları farklı coğrafi bölgelerden veya branşlardan öğretmenlerin katılımlarıyla genişletilebilir. Ayrıca farklı analiz tekniklerinin kullanılmasıyla gerçekleşecek araştırmalarla envanterin geçerlik ve güvenirliğine ilişkin daha fazla kanıt elde edilebilir.

KAYNAKÇA

Argon, T., ve Ateş, H. (2007). İlköğretim okulu birinci kademe öğretmenlerini etkileyen stres faktörleri. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 7*(2), 51-60.

Avcı, Ü., ve Seferoğlu, S. S. (2011). Bilgi toplumunda öğretmenin tükenmişliği: teknoloji kullanımı ve tükenmişliği önlemeye yönelik alınabilecek önlemler. *Akdeniz Eğitim Araştırmaları Dergisi*, *9*, 13-26.

Avşaroğlu, S., Deniz, M. E., ve Kahraman, A. (2005). Teknik öğretmenlerde yaşam doyumu iş doyumu ve mesleki tükenmişlik düzeylerinin incelenmesi. *Selçuk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 14*, 115-129.

Bakoğlu Deliorman, R., Taştan Boz, İ., Yiğit, İ., ve Yıldız, S. (2010). Tükenmişliği ölçmede alternatif bir araç: Kopenhag Tükenmişlik Envanterinin Marmara Üniversitesi akademik personeli üzerine uyarlaması. *Yönetim*, *20*(63), 77-98.

Bümen, N.T. (2010). The relationship between demographics, self efficacy, and burnout among teachers. *Eurasian Journal of Educational Research, 40*, 16-35.

Büyüköztürk, Ş. (2010). *Sosyal bilimler için veri analizi el kitabı.* Ankara: Pegem Akademi.

Cano-García, F. J., Padilla-Muñoz, E. M., & Carrasco-Ortiz, M. Á. (2005). Personality and contextual variables in teacher burnout. *Personality and Individual Differences, 38*(4), 929-940.

Chen, W. S., Haniff, J., Siau, C. S., Seet, W., Loh, S. F., Jamil, M. H. A., ... & Baharum, N. (2014). Translation, cross-cultural adaptation and validation of the Malay version of the Maslach Burnout Inventory (MBI) in Malaysia. *International Journal of Social Science Studies*, *2*(2), 66-74.

Cordes, C. L., & Dougherty, T. W. (1993). A review and an integration of research on job burnout. *Academy of Management Review*, *18*(4), 621-656.

Çokluk, Ö., Şekercioğlu, G., ve Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik SPSS ve LISREL uygulamaları.* Ankara: Pegem Akademi.

Ergin, C. (1993). Doktor ve hemşirelerde tükenmişlik ve Maslach Tükenmişlik Ölçeğinin uyarlanması. Rüveyde Bayraktar ve İhsan Dağ (Ed.), *VII. Ulusal Psikoloji Kongresi Bilimsel Çalışmaları içinde (143-154).* Ankara: VII. Ulusal Psikoloji Kongresi Düzenleme Kurulu ve Türk Psikologlar Derneği.

Ertürk, E., ve Keçecioğlu, T. (2012). Çalışanların iş doyumları ile mesleki tükenmişlik düzeyleri arasındaki ilişkiler: öğretmenler üzerine örnek bir uygulama. *Ege Akademik Bakış, 12*(1), 41-54.

Freudenberger, H. J. (1974). Staff burn-out. *Journal of Social Issues, 30*(1), 159-165.

Friedman, I. A. (1991). High and low-burnout schools: school culture aspects of teacher burnout. *The Journal of Educational Research, 84*(6), 325-333.

Friesen, D., & Sarros, J. C. (1989). Sources of burnout among educators. *Journal of Organizational Behavior, 10*(2), 179-188.

Girgin, G. (2010). Öğretmenlerde tükenmişliğe etki eden faktörlerin araştırılması. *Elektronik Sosyal Bilimler Dergisi, 9*(32), 32-48.

Girgin, G., ve Baysal, A. (2005). Zihinsel engelli öğrencilere eğitim veren öğretmenlerin mesleki tükenmişlik düzeyi ve bazı değişkenler (İzmir örneği). *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 18*(18), 1-10.

Gökçakan, S., ve Murat, M. (2007). Sınıf öğretmenlerinde on yıllık hizmet sürecinde tükenmişliğin gelişimine yönelik bir haritalama çalışması. [Çevrim-içi: http://web.firat.edu.tr/daum/docs/53/35%20S%C4%B1n%C4%B1f%20%C3%96%C4%9Fr.de%2010%20y%C4%B1ll%C4%B1k%20hizmet--Mehmet%20Murat%20-%C3%B6dendi--9%20syf--177-185.doc], Erişim Tarihi: 10Ekim 2013.

Guglielmi, R. S., & Tatrow, K. (1998). Occupational stress, burnout, and health in teachers: a methodological and theoretical analysis. *Review of Educational Research, 68*(1), 61-99.

Gündüz, B. (2005). İlköğretim öğretmenlerinde tükenmişlik. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 1*(1), 152-166.

Hallesleben, J. R., & Buckley, M. R. (2004). Burnout in organizational life. *Journal of Management, 30*(6), 859-879.

Hayes, C. T., & Weathington, B. L. (2007). Optimism, stress, life satisfaction, and job burnout in restaurant managers. *The Journal of Psychology, 141*(6), 565-579.

International Test Commission (2005). International guidelines on test adaptation. [www.intestcom.org]

Kokkinos, C. M. (2006). Factor structure and psychometric properties of the Maslach Burnout Inventory-Educators Survey among elementary and secondary school teachers in Cyprus. *Stress and Health, 22*(1), 25-33.

Kokkinos, C. M. (2007). Job stressors, personality and burnout in primary school teachers. *British Journal of Educational Psychology, 77*(1), 229-243.

Lambie, G. W. (2007). The contribution of ego development level to burnout in school counselors: implications for professional school counseling. *Journal of Counseling & Development, 85*(1), 82-88.

Lee, R. T., & Ashforth, B. E. (1993). A further examination of managerial burnout: toward an integrated model. *Journal of Organizational Behavior, 14*(1), 3-20.

Leiter, M. P., & Harvie, P. (1998). Conditions for staff acceptance of organizational change: burnout as a mediating construct. *Anxiety, Stress & Coping, 11*(1), 1-25.

Leiter, M. P., & Maslach, C. (2003). Areas of worklife: A structured approach to organizational predictors of job burnout. In P. L. Perrewé and D. C. Ganster (Eds.). *Emotional and physiological processes and positive intervention strategies (Research in occupational stress and well-being, Volume 3)* 91-134. Bingley: Emerald Group Publishing Limited.

Leiter, M. P., & Maslach, C. (2005). *Banishing burnout.* San Francisco: Josey-Bass.

Maslach, C. (2003). Job burnout new directions in research and intervention. *Current Directions in Psychological Science, 12*(5), 189-192.

Maslach, C., & Goldberg, J. (1998). Prevention of burnout: new perspectives. *Applied and Preventive Psychology, 7*(1), 63-74.

Maslach, C., & Jackson, S. E. (1981). The measurement of experienced burnout. *Journal of Organizational Behavior, 2*(2), 99-113.

Maslach, C., Jackson, S. E., & Leiter, M. P. (2010). *Maslach Burnout Inventory manual.* (3rd ed.). mindgarden.com.

Maslach, C., & Leiter, M. P. (2008). Early predictors of job burnout and engagement. *Journal of Applied Psychology, 93*(3), 498-512.

Maslach, C., Schaufeli, W. B., & Leiter, M. P. (2001). Job burnout. *Annual Review of Psychology, 52*(1), 397-422.

Ozan, M. B. (2009). A study on primary school teacher burnout levels: The Northern Cyprus case. *Education, 129*(4), 692-703.

Schaufeli, W. B., Leiter, M. P., & Maslach, C. (2009). Burnout: 35 years of research and practice. *Career Development International, 14*(3), 204-220.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23-74.

Schumacker, R. E., & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling.* New York: Routledge.

Schwab, R. L., Jackson, S. E., & Schuler, R. S. (1986). Educator burnout: sources and consequences. *Educational Research Quarterly, 10*(3), 14-30.

Schwarzer, R., Schmitz, G. S., & Tang, C. (2000). Teacher burnout in Hong Kong and Germany: a cross-cultural validation of the Maslach Burnout Inventory. *Anxiety, Stress & Coping, 13*(3), 309-326.

Tümkaya, S. (2005). Öğretmenlerin sınıf içi disiplin anlayışları ve tükenmişlikle ilişkisi. *Kuram ve Uygulamada Eğitim Yönetimi Dergisi*, *11*(4), 549-568.

Tümkaya, S., ve Çavuşoğlu, İ. (2010). Sınıf öğretmenliği son sınıf öğretmen adaylarının tükenmişlik düzeylerinin incelenmesi. *Ç.Ü. Sosyal Bilimler Enstitüsü Dergisi*, *19*(2), 468-481.

Wallin, M. I. (2010). Personality and burnout, [Çevrim-içi: http://www.diva-portal.org/smash/get/diva2:325820/FULLTEXT01.pdf], Erişim tarihi: 14 Nisan 2014.

Weisberg, J., & Sagie, A. (1999). Teachers' physical, mental, and emotional burnout: impact on intention to quit. *The Journal of Psychology, 133*(3), 333-339.

EXTENDED ABSTRACT

Introduction

Maslach considers the burnout concept as a process with three dimensions including emotional exhaustion, depersonalization and personal accomplishment. According to Maslach model, the employees who cannot fulfill people's psychological demands any more first live emotional exhaustion (Maslach & Jackson, 1981). The staff who cannot cope with the excess demands in the business environment keep distance from the clients and become desensitized, realize discrepancy between current behavior and the contribution they expect to do to their society and their work, and consider their personal accomplishment insufficient (Cordes & Dougherty, 1993). It is aimed to carry out the adaptation study of Maslach Burnout Inventory-Educators Survey (MBI-ES) to Turkish in order to determine the reflection of the problems encountered in the teaching profession on the burnout levels of educators.

Method

Classroom teachers working in state elementary schools in Ankara constitute accessible population of this study in which descriptive method, one of the quantitative methods, has been adopted. Validity and reliability analysis have been made with the data obtained from burnout scale applied to 760 classroom teachers from the study group. While the data obtained from randomly selected 220 teachers was considered as a pilot application of the study, the data obtained from 540 teachers was used as the main application data.

Maslach & Jackson (1981) used several methods to determine the validity of the MBI. First, colleagues or spouses of the employees were asked to evaluate the behaviors of individuals. Significant correlation has been identified between the scores of the employees, whose behaviors were assessed by colleagues and spouses, and burnout levels. In another method for the validity study of MBI, significant correlation has been identified between some dimensions of Hackman & Oldham's Job Descriptive Scale and some dimensions of MBI (Maslach, Jackson & Leiter, 2010). In the third method, it has been examined whether there are significant correlation between training and development opportunities in business, the employees' considering their job significant, the intention to leave the job, the desire to spend less time at work, disruption of human relations at and outside

the work, the problems experienced with family and friends, insomnia, the increase in alcohol and drug use, and burnout levels of the employees.

Christina Maslach, one of the authors of the scale, was contacted before the adaptation study. Upon her guidance, adaptation study of the burnout scale was started after making necessary correspondence with "mindgarden.com" and paying royalty fee. According to this, a form was prepared after translating the items in the original English form of the scale into Turkish and was consulted to experts. The items in the original language of the scale and Turkish translations were written on the translation form and were asked if the expressions were appropriate. The translation form prepared was sent to 10 scholars in the departments of Educational Administration, Elementary Teacher Education and Foreign Language Education either by e-mail or face-to-face interviews. Corrections were made on the items with no agreement as a result of the feedback and were taken to the data collection tool.

Results and Discussion

Confirmatory factor analysis was applied using the Linear Structural Relations (LISREL) 8.80 program to determine the construct validity of the MBI-ES. p-value of the scale appears to be significant at the level of 0.01 according to the path diagram related to the factor analysis made on the basis of main application data by staying with the three-dimensional factor structure at the original form of the scale. Significance of "p" value, in which there should be no significant difference, due to sample size is expected to be met with tolerance (Çokluk, Şekercioğlu & Büyüköztürk, 2012). It is possible to interpret that correlation values between the three dimensions of the burnout scale measure different properties because of taking place in mid-level. Considering the values (x2/sd=4.3, RMSEA=0.07, CFI=0.94, NFI=0.93, GFI=0.87, AGFI=0.84) indicating the level of compliance of the scale as a result of the confirmatory factor analysis applied to MBI-ES, the factor model established for verification generally shows a good level of alignment.

Reliability analysis of MBI-ES were performed by the determination of item-total correlations and Cronbach alpha coefficient calculation method. According to the findings obtained as a result of calculating Cronbach alpha reliability coefficient, it is understood that the values was found 0.88 for emotional exhaustion, 0.78 for depersonalization and 0.74 for personal accomplishment. The items included in the scale have a good discriminant level according to the data related to the item-total correlations used in order to determine whether the items in the scale discriminate the individuals with and without the features required to be measured.

MBI-ES's findings regarding the confirmatory factor analysis shows that the original structure of the scale consisting of 22 items and three dimensions has been protected in the study group of this research. The reliability levels for the subscales are understood to be sufficient. Ergin (1993) has determined in factor analysis made for MBI likewise that the scale piled on three factors. Reliability levels of the scale with Cronbach alpha coefficient were calculated as 0.83 for emotional exhaustion dimension, 0.65 for depersonalization, and 0.72 for personal accomplishment dimension. The coefficients calculated by the test-retest method were determined to be 0.83 for emotional exhaustion dimension, 0.72 for depersonalization dimension, and 0.67 for personal accomplishment dimension. The findings obtained from this study regarding the MBI-ES reliability levels seem to be consistent with the findings from the study of Ergin. Therefore, the Turkish version of the scale is thought to be used to determine the vocational burnout levels of primary school teachers.

Ekler

Ek-1

mind garden

www.mindgarden.com

To whom it may concern,

This letter is to grant permission for the above named person to use the following copyright material for his/her thesis or dissertation research:

Instrument: ***Maslach Burnout Inventory, Forms: General Survey, Human Services Survey & Educators Survey***

Copyrights:

MBI-General Survey (MBI-GS): Copyright ©1996 Wilmar B. Schaufeli, Michael P. Leiter, Christina Maslach & Susan E. Jackson. All rights reserved in all media. Published by Mind Garden, Inc., www.mindgarden.com

MBI-Human Services Survey (MBI-HSS): Copyright ©1981 Christina Maslach & Susan E. Jackson. All rights reserved in all media. Published by Mind Garden, Inc., www.mindgarden.com

MBI-Educators Survey (MBI-ES): Copyright ©1986 Christina Maslach, Susan E. Jackson & Richard L. Schwab. All rights reserved in all media. Published by Mind Garden, Inc., www.mindgarden.com

Three sample items from a single form of this instrument may be reproduced for inclusion in a proposal, thesis, or dissertation.

The entire instrument may not be included or reproduced at any time in any published material.

Sincerely,

Robert Most
Mind Garden, Inc.
www.mindgarden.com