Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN:1309-6575 *Yaz* 2018 *Summer* 2018 *Cilt: 9- Sayı: 2* *Volume: 9- 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 dört kez yayınlanan hakemli ulusal bir dergidir. Yayımlanan yazıların tüm sorumluğu ilgili yazarlara aittir.

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

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

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

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 Öğretim Üyesi)
Akif AVCU (Marmara Üni.)
Arif ÖZER (Hacettepe Üni.)
Asiye Şengül Avşar (Recep Tayyip Erdoğan Üni.)
Ayfer SAYIN (Gazi Üni.)
Aylin ALBAYRAK SARI (Hacettepe Üni.)
Ayşegül ALTUN (Ondokuz Mayıs Üni.)
Bahar Şahin Sarkın (İstanbul Okan Üni.)
Bayram BIÇAK (Akdeniz Üni.)
Bayram ÇETİN (Gazi Üni.)
Bilge BAŞUSTA UZUN (Mersin Üni.)
Bilge GÖK (Hacettepe Üni.)
Burak AYDIN (Recep Tayyip Erdoğan Üni.)
Burcu ATAR (Hacettepe Üni.)
Burhanettin ÖZDEMİR (Siirt Üni.)
Beyza AKSU DÜNYA (Bartın Üni.)
Cem Oktay GÜZELLER (Akdeniz Üni.)
Ceylan GÜNDEĞER (Hacettepe Üni.)
Cindy M. WALKER (Duquesne University)
Çiğdem AKIN ARIKAN (Hacettepe Üni.)
David KAPLAN (University of Wisconsin)
Deniz GÜLLEROĞLU (Ankara Üni.)
Derya ÇAKICI ESER (Kırıkkale Üni)
Derya ÇOBANOĞLU AKTAN (Hacettepe Üni.)
Didem Özdoğan (İstanbul Kültür 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 (Trakya Üni.)
Emine ÖNEN (Gazi Üni.)
Emrah GÜL (Hakkari Üni.)
Emre ÇETİN (Doğu Akdeniz Üni.)
Eren Halil Özberk (Trakya Üni.)
Ergül DEMİR (Ankara Üni.)
Esin TEZBAŞARAN (İstanbul Üni)
Esin YILMAZ KOĞAR (Niğde Ömer Halisdemir Üni.)
Esra Eminoğlu ÖZMERCAN (MEB)
Evrim ÇETİNKAYA YILDIZ (Erciyes Üni.)
Fatih KEZER (Kocaeli Üni.)
Fatih ORCAN (Karadeniz Teknik Üni)
Fatma BAYRAK (Hacettepe Üni.)
Fazilet TAŞDEMİR (Recep Tayyip Erdoğan Üni.)
Funda NALBANTOĞLU YILMAZ (Nevşehir Üni.)
Gonca Usta (Cumhuriyet Üni.)
Göksu GÖZEN (Mimar Sinan Güzel Sanatlar Üni.)
Gül GÜLER (İstanbul Aydın 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.)
İlhan KOYUNCU (Adıyaman Üni.)
İlker KALENDER (Bilkent Üni.)
İsmail KARAKAYA (Gazi Üni.)
Kaan Zülfikar DENİZ (Ankara Üni.)
Kübra ATALAY KABASAKAL (Hacettepe Üni.)
Levent YAKAR (Hacettepe. Üni.)
Mehmet KAPLAN (MEB)
Meltem ACAR GÜVENDİR (Trakya Üni.)
Mustafa ASİL (University of Otago)
Nagihan BOZTUNÇ ÖZTÜRK (Hacettepe Üni.)
Neşe GÜLER (İzmir Demokrasi Üni.)
Neşe ÖZTÜRK GÜBEŞ (Mehmet Akif Ersoy Üni.)
Nuri DOĞAN (Hacettepe Üni.)
Nükhet DEMİRTAŞLI (Emekli Öğretim Üyesi)
Okan BULUT (University of Alberta)
Onur ÖZMEN (TED Üniversitesi)
Ömer KUTLU (Ankara Üni.)
Ömür Kaya KALKAN (Pamukkale Üni.)
Önder SÜNBÜL (Mersin Üni.)
Özge BIKMAZ BİLGEN (Adnan Menderes Üni.)
Ragıp Terzi (Harran Üni.)
Recep Serkan ARIK (Dumlupınar Üni.)
Sakine GÖÇER ŞAHİN (University of Wisconsin Madison)
Seçil ÖMÜR SÜNBÜL (Mersin Üni.)
Sedat ŞEN (Harran Üni.)
Seher YALÇIN (Ankara Üni.)
Selahattin GELBAL (Hacettepe Üni.)
Sema SULAK (Bartın Üni.)
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)
Tuğba KARADAVUT AVCI (Kilis 7 Aralık Üni.)
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.

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Haziran 2018, Sayı: 9-2)

Journal of Measurement and Evaluation in Education and Psychology (June 2018, Issue: 9-2)

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(2);85-100

Öz-Yönetimli Öğrenme Becerileri Ölçeği: Geçerlik ve Güvenirlik Çalışması*

Self Directed Learning Skills Scale: Validity and Reliability Study

İlkay AŞKIN TEKKOL ** Melek DEMİREL ***

Öz

Bu araştırmanın amacı üniversite öğrencilerinin öz-yönetimli öğrenme becerilerini belirlemeye yönelik bir ölçme aracı geliştirmektir. Bu amaçla öncelikle, ilgili alanyazına dayalı olarak kavramsal çerçeve oluşturulmuş, bu alanda yurt içi ve yurt dışında yapılmış olan ölçek geliştirme çalışmaları incelenmiştir. Bu incelemenin ardından, ölçeğe ilişkin ölçütler ortaya konmuş ve öz-yönetimli öğrenme becerilerine sahip bireylerin özelliklerini ifade eden maddelerden oluşan 72 maddelik madde havuzu oluşturulmuştur. Uzman görüşlerinin ardından yapılan düzenlemeler sonucunda, ölçeğin 53 maddelik deneme formu oluşturulmuştur. Bu form, Hacettepe ve Başkent Üniversitelerinde öğrenim görmekte olan 753 üniversite öğrencisine uygulanarak, veriler açımlayıcı faktör analizine tabi tutulmuş ve yapılan analiz sonucunda ölçeğin 21 madde ve dört boyuttan (güdülenme, öz-izleme, öz-kontrol ve özgüven) oluştuğu ortaya konmuştur. Ayrıca, açımlayıcı faktör analizinin uygulandığı öğrencilerin dışında yer alan ve Hacettepe ve Başkent Üniversitelerinin ortak olan bölümlerinin birinci ve dördüncü sınıflarında öğrenim görmekte olan 2600 öğrenciye ölçek uygulaması yapılarak ölçeğin yapı geçerliğinin belirlenmesi amaçlanmıştır. Toplanan veriler doğrulayıcı faktör analizine tabi tutulmuştur. Analiz sonucunda ölçeğe ilişkin uyum indekslerinin iyi uyum gösteren ve kabul edilebilir değerler arasında olduğu görülmüş ve ölçeğin 4 faktörlü ve 21 maddeden oluşan yapısı bir model olarak doğrulanmıştır. Ölçeğin, Cronbach Alpha iç tutarlık katsayısı kullanılarak hesaplanan güvenirliği ise .895 bulunmuştur. Sonuç olarak geliştirilen "Öz-Yönetimli Öğrenme Becerileri Ölçeği"nin, geçerli bir ölçme aracı olduğu ve öz-yönetimli öğrenme becerilerini ölçmede güvenilir puanlar vereceği söylenebilir.

Anahtar Kelimeler: Öz-yönetimli öğrenme, öz-yönetimli öğrenme becerileri, ölçek geliştirme, üniversite öğrencileri

Abstract

The aim of this study was to develop a psychometric scale that aims to measure self-directed learning skills of undergraduate students. The literature was first reviewed, and a conceptual framework was established to identify the basic characteristics of the scale. In addition, similar scale development studies in and outside Turkish literature were examined. After then, criteria were execute which related scale and a 72-item pool was formed, and it was revised in line with expert views, and a version with remaining 53item was implemented on 753 students attending Hacettepe and Başkent Universities. The data were implemented exploratory factor analysis. It was found that the scale consisted of four sub-scales (motivation, self-control, self-monitoring and self-confidence), and twenty one items associated with the sub-scales. Also, in order to determine the construct validity of the scale, a confirmatory factor analysis was carried out with 2600 students attending Hacettepe and Başkent Universities. This second group of students were the first and last year of students in the common programs in both universities. The fit indices of the scale showed good or acceptable values. According to this, the 21-item and four-factor structure of the 'Self-Directed Learning Skills Scale' was confirmed as a model. The Cronbach Alpha internal consistency coefficient was ,895. As a result, it was concluded that the "Self-

* Bu çalışma birinci yazarın doktora tezinden üretilmiştir. (Üniversite Öğrencilerinin Öz-Yönetimli Öğrenme Becerilerinin İncelenmesi, Danışman: Doç. Dr. Melek DEMİREL, 2015)

** Dr. Öğr. Üyesi, Kastamonu Üniversitesi, Eğitim Fakültesi, Kastamonu-Türkiye, e-posta:ilkayaskin@hotmail.com, ORCID ID: orcid.org/0000-0003-0964-1528

*** Prof. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, e-posta:melekdemirel@gmail.com, ORCID ID: orcid.org/0000-0002-2449-5824

Geliş Tarihi :02.02.2018
DOI: 10.21031/epod.389208 *Kabul Tarihi: 25.03.2018*

directed Learning Skills Scale" was a valid measurement tool and produced reliable scores for measuring self-directed learning skills.

Keywords: self-directed learning, self-directed learning skills, scale development, undergraduate students

GİRİŞ

Bilginin, bilgiyi sağlayan otoriteler tarafından değişmez parçalar halinde bireylere aktarılması anlayışı, günümüzde pek çok farklı kaynaktan bilgiye ulaşılabiliyor olması nedeniyle değişikliğe uğramıştır. Bu sayede, bireylerin bilgileri ezber yolu ile doğrudan almaları düşüncesi yerine, bireylerin bilgileri kavrama ve sorgulama yoluyla ele almaları görüşü giderek artan bir önem kazanmıştır. Bu durumda, herhangi bir disipline ilişkin bilgiyi ezberlemek yerine, bilginin doğasını kazanmak, bilgiye ulaşma yolunu öğrenmek; diğer bir deyişle öğrenmeyi öğrenmek önemli bir hale gelmiştir.

Öğrenmeyi öğrenme, önceleri öğrencinin daha çok öğrenmesi anlamında kullanılırken; artan bilgi miktarı ve öğrencilerin öğrenmeye harcadıkları zamanın artması nedeniyle değişikliğe uğramıştır (Csapo, 2007). Wingate (2007) öğrenmeyi öğrenmenin iki temel öğesi olduğunu ifade etmektedir. Bu öğeler: (1) "öğrenme"yi anlama ve bağımsız öğrenen olma ve (2) "bilgi"yi anlama ve bilgiyi yapılandırma konusunda yeterli olmadır. Wingate'e (2007) göre bu öğeler öğrenmeyi öğrenen bireylerin yetişmesi konusunda büyük bir öneme sahiptir ve her iki öğe de belirli özellikleri içermektedir. Bu özellikler şu şekilde açıklanabilir (Wingate, 2007):

1. Öğrenmeyi anlama ve bağımsız öğrenen olma
    a. Öğrenme anlayışları konusunda farkındalık kazanma
    b. Öğrenen olarak yeteneklerini değerlendirme
    c. Kısa ve uzun vadeli hedefler belirleme
    d. Hedeflere ulaşma konusunda planlama yapma
    e. Hedeflere ulaşırken süreci gözden geçirme
    f. Başarıyı ve süreci değerlendirme
2. Bilgiyi anlama ve bilgiyi yapılandırma konusunda yeterli olma
    a. Disiplin içindeki bilgiyi fark etme
    b. Odaklı bir şekilde bilgiye (ders notu, kitaplar) başvurma
    c. Var olan bilgiyi değerlendirme
    d. Farklı kaynakları tutarlı bir şekilde bir araya getirme/sentezleme
    e. Kendini ifade etme

En yalın ifade ile öğrenmeyi öğrenen bireyler, kendi öğrenmelerini yönetebilen bireylerdir. Alanyazında da öğrenmeyi öğrenme becerisi; öz-yönetimin temel ilkelerinden biri olarak kabul edilmektedir (Salas, 2010). Öz-yönetim, bireyin öğrenmesinin kendi kontrolünde olması, diğer bir deyişle öğrenenin kendi öğrenmesini yönetebilmesidir.

Öz-yönetimli öğrenme ise en geniş anlamda, bireylerin öğrenme ihtiyaçlarını belirleme, öğrenme hedeflerini ortaya koyma, öğrenme konusunda ihtiyaç duydukları kaynakları tanımlama, uygun öğrenme stratejilerini seçme/kullanma ve öğrenme çıktılarını değerlendirme konusunda başkalarının yardımı ile ya da yardımı olmaksızın inisiyatif almalarını ifade etmektedir (Knowles, 1975). Öz-yönetimli öğrenme, öğrenme sorumluluğunun bir kaynaktan (öğretmen, vb.), bireyin kendisine doğru yön değiştirmesi sürecini içermektedir. Burada, öğrenenin öğrenme sürecindeki kontrolü ve etkin

katılımı büyük bir öneme sahiptir (Boyer ve Usinger, 2015; Grover, 2015). Bu kontrol sayesinde bireyler kendi öğrenmelerini düzenleyebilmekte ve hedeflerini belirleyebilmektedirler. Bireyler belirledikleri hedeflere ulaşabilmek için öğrenme sorumluluğunu üstlenerek bireysel tercihlerde bulunabilirler (Kaufman, 2003). Öz-yönetimli öğrenmede aynı zamanda, bireyin öğrenme konusundaki güdülenme düzeyi de büyük bir önem taşımaktadır (Song ve Hill, 2007).

Öz-yönetimli öğrenme, farklı amaçları içermektedir. Bu konudaki yazarların felsefi yönelimlerine göre, öz-yönetimli öğrenmenin üç temel amacından söz edilebilir. Öz-yönetimli öğrenmenin ilk amacı; öğrenenlerin öz-yönetimli öğrenme konusunda kendilerini geliştirmelerinin sağlanmasıdır. Bu amacı gerçekleştirebilmek için, bireylerin özelliklerine uygun olarak becerilerinin özelleştirilmesi gerekmektedir. İkinci amaç, öz-yönetimli öğrenme özelinde dönüştürücü öğrenmeyi (transformational learning) teşvik etmektir. Dönüştürücü öğrenme, Mezirow'un (1991) ortaya koyduğu, öğrenenlerin öğrenmede eleştirel yansıtmayı merkeze aldıkları bir süreçtir. Burada yer alan eleştirel yansıtma, bireyin ihtiyaçlarının, isteklerinin ve ilgilerinin tarihi, kültürel ve biyolojik nedenlerini anlama olarak ifade edilmektedir. Mezirow (1991) böyle bir öz-bilginin, öz-yönetimli öğrenmede bağımsızlığın sağlanabilmesi için bir ön koşul olduğunu belirtmektedir. Öz-yönetimli öğrenmenin üçüncü amacı, özgürlükçü öğrenmeyi ve sosyal etkinlikleri teşvik etmektir. Alanyazında Brookfield ve Collins gibi araştırmacılar, öz-yönetimli öğrenmenin, bireysel öğrenmeden çok sosyal ve politik etkinliklerinden söz etmekte ve öz-yönetimli öğrenmenin eleştirel ve politik analizleriyle ilgilenmektedirler (Merriam, Caffarella ve Baumgartner, 2007).

Öz-yönetimli öğrenme, bireylerin özgüvenlerinin, özerkliklerinin, güdülenmelerinin artması ve yaşam boyu öğrenme becerilerinin gelişmesini sağlamaktadır (O'Shea, 2003). Bunun yanı sıra, öz-yönetimli öğrenme öğrenenlerin etkin katılımcılar olmalarını sağlamakta ve derinlemesine öğrenen öğrenciler olmaları konusunda onları cesaretlendirmektedir (Spencer ve Jordan, 1999). Ancak bireylerin, öz-yönetimli öğrenenler olabilmeleri için sahip olmaları gereken bir takım yeterlikler bulunmaktadır. Bu yeterlikler Knowles (1977) tarafından şu şekilde açıklanmıştır:

- Öğrenenlerle yakın, saygılı ve öğrenmeyi kolaylaştırıcı bir ilişki içerisinde bulunma becerisi
- Fiziksel ve psikolojik olarak rahat, etkileşime açık, işbirliğine dayalı, açık ve karşılıklı güvene dayalı bir ortam oluşturma becerisi
- Bireylerin öğrenme konusundaki ihtiyaçlarını belirleme sorumluğunu almaları
- Bireylerin hedeflerini belirleme becerisi
- Bireylerin öğrenme etkinliklerini planlama, yürütme ve değerlendirme becerisi
- Öğrenenlere öz-yönetimli olabilmeleri konusunda yardım etme becerisi
- Küçük grup süreçlerini etkili bir şekilde kullanma becerisi
- Öğrenme süreçlerini ve çıktılarını değerlendirme becerisi (Knowles 1977; akt. Kasworm, 1983).

Knowles'un üzerinde durduğu bu yeterliklere sahip bireylerin öz-yönetimli öğrenen bireyler oldukları ifade edilebilir. Savin-Baden ve Major'a (2004) göre öz-yönetimli öğrenenler; açık hedefler koyan, bir plan doğrultusunda ilerleyen, görevlerini takip eden ve normal sınırlarını yüksek standartlara ulaşmak için zorlayan bağımsız, öz-güdülenmeye sahip bireylerdir. Ayrıca, öz-yönetimli öğrenenlerin, öğrenmeye açık, meraklı, düzenli, güdülenmiş ve hevesli bireyler olduğu; öğrenmeye değer verdikleri ve öz-kontrollü oldukları; bunun yanı sıra, belirsizlik ve değişiklik konusunda rahat oldukları ifade edilmektedir (Jennett, 1992).

Araştırmanın Amacı

Öz-yönetimli öğrenenler; kendilerine net hedefler koyan, planlı hareket eden, insiyatif alan, öğrenmeye açık, güdülenmiş, kendilerine güvenen ve öz-kontrollü bireylerdir. Bu özellikler bilginin

hızla değiştiği ve katlanarak arttığı günümüzde, çağa ayak uydurabilecek bireylerin sahip olmaları gereken nitelikler arasında yer almaktadır. Kendi öğrenmesini yönetebilen bireyler; bilgiye ulaşma yollarını kazanmış, üst düzey düşünebilen ve kendi öğrenmesini düzenleyebilen, kısaca öğrenmeyi öğrenmiş bireylerdir. Üniversite öğrencilerinin bu özelliklere sahip olmaları, onların üniversite yaşamından sonra da gerek kişisel gerek mesleki olarak kendini geliştirmelerini, öğrenme isteği taşımalarını, öğrenmeye açık olmalarını ve öğrenmelerini sürdürme eğiliminde olmalarını; kısaca yaşam boyu öğrenenler olmalarını sağlayacaktır. Bu bağlamda üniversite öğrencilerinin öz-yönetimli öğrenme becerilerinin belirlenmesi önemli görülmektedir. Bu becerilerin tespit edilebilmesi için ise geçerli ve güvenilir bir ölçme aracının kullanılması gerekmektedir. Bu kapsamda bu araştırmanın amacı, üniversite öğrencilerinin öz-yönetimli öğrenme becerilerinin belirlenmesini sağlayacak geçerli ve güvenilir bir ölçek geliştirmektir.

Alanyazında, öz-yönetimli öğrenme konusundaki ölçek geliştirme çalışmaları incelendiğinde yurt dışında geliştirilen çeşitli ölçekler bulunmakla birlikte; yurt içinde öz-yönetimli öğrenme becerilerini ölçmek amacıyla kullanılan ölçeklerin genellikle yurt dışından uyarlanan ölçekler olduğu görülmektedir (Demir ve Yurdagül, 2013; Kocaman, Dicle, Üstün ve Çimen, 2004; Salas, 2010; Sasa, 2011; Şahin, 2013; Şahin ve Erden, 2008). Bunun yanı sıra, yurt içinde geliştirilen az sayıda ölçeğin, öğrencilerin öz-yönetimli öğrenme becerilerine ilişkin belirli özellikleri ölçmek üzere geliştirilen alana özgü ölçekler olduğu ortaya konmuştur (Acar, 2014; Alkan ve Erdem, 2013; Aydede ve Kesercioğlu, 2009). Bu ölçeklerin, ilköğretim öğrencilerinin fen ve teknolojiye ilişkin öz-yönetimli öğrenme becerileri, öğretmen adaylarının öz-yönetimli öğrenme becerileri, hemşirelik bölümü öğrencilerinin öz-yönetimli öğrenme becerileri gibi öz-yönetimli öğrenmeye ilişkin daha özelleşmiş ölçekler olduğu, üniversite öğrencilerinin öz-yönetimli öğrenme becerilerini genel olarak ölçmeye yönelik ölçeklerin olmadığı görülmektedir. Uyarlanan ölçeklerden Guglielmino'nun (1977) ölçeğinin geçerliğine ilişkin sorunların ortaya konması (Field, 1989) ve Fisher, King ve Tague'nin (2001) ölçeğinin de hemşirelik ve eğitim fakültesi gibi alanlara uygun olarak uyarlanmış olması sebebiyle bu alanda yeni ve kapsamlı bir ölçek geliştirilmesi gerekli görülmüştür. Bu kapsamda araştırmada, tüm üniversite öğrencilerinin öz-yönetimli öğrenme becerilerini ortaya koymak amacıyla "Öz-yönetimli Öğrenme Becerileri Ölçeği"nin geliştirilmesi amaçlanmıştır.

YÖNTEM

Bu araştırma bir ölçek geliştirme çalışmasıdır. Araştırmada üniversite öğrencilerinin öz-yönetimli öğrenme becerilerinin tespit edilebilmesi amacıyla "Öz-Yönetimli Öğrenme Becerileri Ölçeği" geliştirilmiştir. Ölçek geliştirmeye ilişkin adımlar sırasıyla açıklanmıştır.

Çalışma Grupları

Araştırmada, açımlayıcı faktör analizi ve doğrulayıcı faktör analizi ayrı gruplar üzerinde yapılmıştır. Araştırmada devlet üniversitesi olarak gelişmiş ve araştırmacılar açısından kolay ulaşılabilir bir üniversite olduğu için Hacettepe Üniversitesi ve bu üniversite ile ortak fakülte ve bölümlere sahip olması nedeniyle özel üniversite olarak Başkent Üniversitesi seçilmiştir. Çalışma gruplarına her iki üniversitede ortak olan bölümlerde öğrenim gören öğrenciler dahil edilmiştir. Çalışma gruplarına ilişkin bilgiler aşağıda yer almaktadır:

Çalışma grubu 1

Araştırmanın çalışma grubunu Hacettepe ve Başkent Üniversitelerinde öğrenim gören öğrenciler oluşturmuştur. Öğrencilere ilişkin bilgiler Tablo 1'de sunulmuştur.

Tablo 1. Açımlayıcı Faktör Analizinin Uygulandığı Çalışma Grubu

Fakülte	Öğrenci Sayısı
Eğitim Fakültesi	173
Fen-Edebiyat Fakültesi	141
Güzel Sanatlar Fakültesi	40
Devlet Konservatuvarı	6
Hukuk Fakültesi	115
İktisadi ve İdari Bilimler Fakültesi	91
Mühendislik Fakültesi	90
Sağlık Bilimleri	110
Toplam	766

Ölçeğin deneme formu Tablo 1'de görüldüğü gibi 766 öğrenciye uygulanmıştır. Bu formlardan eksik ya da hatalı doldurulan Tablo 1 incelendiğinde, 13 form çıkarılmıştır. Sonuç olarak açımlayıcı faktör analizi 753 öğrencinin verisi ile gerçekleştirilmiştir.

Çalışma grubu 2

Doğrulayıcı faktör analizi açımlayıcı faktör analizinin yapıldığı gruptan farklı bir grup üzerinde uygulanmıştır. Ölçeğinin yapısının doğrulanması amacıyla, Ankara ilinde bulunan Hacettepe ve Başkent Üniversitelerinin Mühendislik, Tıp, Diş Hekimliği, Güzel Sanatlar, Hemşirelik, Sağlık Bilimleri, İktisadi ve İdari Bilimler, Eğitim ile Fen ve Edebiyat fakülteleri ile Devlet Konservatuvarının birinci ve dördüncü sınıflarında öğrenim görmekte olan 2600 üniversite öğrencinin verileri doğrulayıcı faktör analizine tabi tutulmuştur (Bilgisayar, Elektrik Elektronik, Endüstri Mühendislikleri, Tıp Fakültesi, Diş Hekimliği, Hemşirelik, Sosyal Hizmet, Beslenme ve Diyetetik, Fizyoterapi ve Rehabilitasyon, İşletme, İktisat, Psikoloji, Türk Dili ve Edebiyatı, Okul Öncesi, Sınıf, İlköğretim Matematik, İngilizce Öğretmenlikleri ve Psikolojik Danışma ve Rehberlik, İç Mimarlık ve Çevre Tasarımı, Grafik, Müzik/Sahne Sanatları). Çalışma grubuna ilişkin bilgiler Tablo 2'de yer almaktadır.

Tablo 2. Doğrulayıcı Faktör Analizinin Uygulandığı Çalışma Grubu

Konu Alanları	Bölümler	Öğrenci Sayısı
Fen Bilimleri	Bilgisayar Müh. Elektrik Elektronik Müh. Endüstri Müh. İlköğretim Matematik Öğrt.	444
Güzel Sanatlar	Grafik İç Mimarlık ve Çevre Tasarımı Müzik/Sahne Sanatları	142
Sağlık Bilimleri	Beslenme ve Diyetetik Diş Hekimliği Fizyoterapi ve Rehabilitasyon Hemşirelik Sosyal Hizmet Tıp	1105
Sosyal Bilimler	Türk Dili ve Edebiyatı Psikoloji Okul Öncesi Öğrt. Sınıf Öğrt. İngilizce Öğrt. Rehberlik ve Psikolojik Dan. İktisat, İşletme	909
Toplam		2600

Tablo 2'ye göre doğrulayıcı faktör analizinin uygulandığı çalışma grubunu, Bilgisayar Elektrik Elektronik, Endüstri Mühendislikleri ve İlköğretim Matematik Öğretmenliği bölümlerinde öğrenim gören olan 444; Grafik, İç Mimarlık ve Çevre Tasarımı ve Müzik/Sahne Sanatları bölümlerinde öğrenim gören 142; Beslenme ve Diyetetik, Diş Hekimliği, Fizyoterapi ve Rehabilitasyon, Hemşirelik, Sosyal Hizmet ve Tıp alanlarında öğrenim gören 1105; Türk Dili ve Edebiyatı, Psikoloji, Okul Öncesi, Sınıf, İngilizce Öğretmenlikleri, Rehberlik ve Psikolojik Danışma, İktisat ve İşletme bölümlerinde öğrenim görmekte olan 909 üniversite öğrencisi oluşturmuştur. Doğrulayıcı faktör analizi toplam 2600 öğrencinin verisi üzerinde uygulanmıştır.

İşlem

Ölçeğin geçerliğinin ortaya konması amacıyla, kapsam ve yapı geçerliği çalışmaları yapılmıştır. Ölçeğin kapsam geçerliği belirlenirken öncelikle ilgili alanyazını incelenmiş ve bu doğrultuda kavramsal çerçeve oluşturularak, ölçeğe ilişkin temel özellikler ortaya konmuştur. Bu özellikler belirlenirken bu alanda yurt içi ve yurt dışında yapılmış olan ölçek geliştirme çalışmaları incelenmiş ve ölçek maddelerini oluşturmak amacıyla bir grup üniversite öğrencisi ile görüşmeler yapılmıştır. Bunun sonucunda ölçeğe ilişkin temel ölçütler tespit edilmiştir. Bu ölçütler öğrenme ihtiyacını belirleme, öğrenme amacını belirleme, öğrenme sonuçlarını değerlendirme, farklı öğrenme stratejilerini kullanma, öğrenme sürecini izleme, planlama yapma, öğrenmeyi sevme, öğrenmekten vazgeçmeme/öğrenmede sebat, kendine güven/benlik algısı, öğrenme sorumluluğunu alma, girişkenlik ve bağımsızlık olarak belirlenmiştir. Ölçütlerin ortaya konmasının ardından, 72 maddelik madde havuzu oluşturulmuş ve yedi eğitim programları ve öğretim, üç ölçme ve değerlendirme bir Türk Dili uzmanına sunularak, uzman görüşleri alınmıştır. Alınan görüşler doğrultusunda ölçekte gerekli düzenlemeler yapılmış ve ölçeğin 53 maddelik deneme formu, 5'li likert tipine göre (her zaman, genellikle, bazen, nadiren, hiçbir zaman) puanlanmıştır.

Verilerin Toplanması

Araştırmanın verileri, Hacettepe ve Başkent Üniversitelerinin belirlenen bölümlerinin birinci ve dördüncü sınıflarında öğrenim gören öğrencilerinden, araştırmacı tarafından toplanmıştır. Uygulama öncesinde ilgili üniversitelerin rektörlüklerinden ve dekanlıklarından izin alınarak, uygulamalar, öğrencilerin gönüllü katılımları ile gerçekleştirilmiştir.

Verilerin Analizi

Toplanan verilerin faktör analizi için uygun olup olmadığının belirlenmesi amacıyla, verilere Kaiser-Meyer-Olkin (KMO) katsayısı ve Barlett Sphericity testi uygulanmıştır. Ardından, Varimaks rotasyonu ile Temel Bileşenler Analizi kullanılarak açımlayıcı faktör analizi yapılmıştır. Ölçeğin güvenirliğinin hesaplanması amacıyla ise Cronbach Alpha İç Tutarlılık katsayısı hesaplanmıştır. Açımlayıcı faktör analizi SPSS programı ile yapılmıştır.

Doğrulayıcı faktör analizi yapılmadan önce, 2600 öğrencinin verilerindeki uç değerlerin ortaya konması amacıyla Mahalonobis uzaklıkları hesaplanarak, uç değer veren veriler, çalışma grubundan çıkarılmıştır ve 2533 öğrencinin verisi analize dahil edilerek, ölçeğin yapısının doğrulanması amaçlanmıştır. Doğrulayıcı faktör analizi yapılırken, LISREL programından yararlanılmıştır.

BULGULAR

Açımlayıcı Faktör Analizi

Toplanan verilerin faktör analizi için uygun olup olmadığının belirlenmesi amacıyla, verilere Kaiser-Meyer-Olkin (KMO) katsayısı ve Barlett Sphericity testi uygulanmıştır. Analiz sonucunda KMO

değeri .95 olarak hesaplanmıştır. Bu değer örneklem büyüklüğünün faktör analizi için yeterli olduğu anlamına gelmektedir. Barlett Sphericity testinin sonucunda ise anlamlılık değeri ,000 bulunmuştur. Buna göre, verilerin faktör analizi için uygun olduğu görülmektedir. Tablo 3'de analiz sonuçları yer almaktadır.

Tablo 3. KMO ve Bartlett Testi Sonuçları

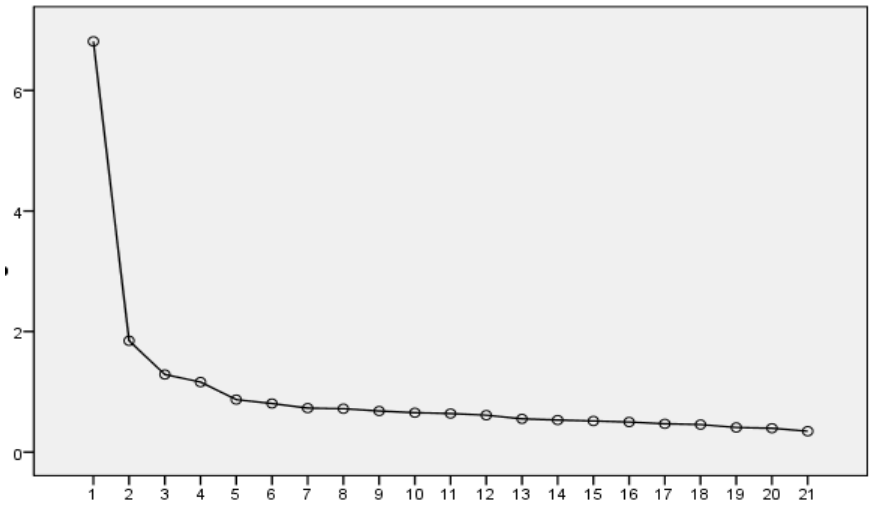
Kaiser-Meyer-Olkin Örneklem Yeterliği Ölçümü	,953	
Bartlett's Testi Sonuçları	Yaklaşık Kay-Kare	14795,919
	Serbestlik derecesi	1378
	Anlamlılık düzeyi	,000

Verilerin faktör analizi için uygun olmasının ortaya konması üzerine, verilere Varimaks rotasyonu ile Temel Bileşenler Analizi kullanılarak açımlayıcı faktör analizi yapılmıştır. Faktör analizi ile ölçeğin boyutlarının belirlenmesi amaçlanmıştır. Yapılan açımlayıcı faktör analizine göre, ölçeğin öz değeri 1'den büyük 10 faktörü olduğu görülmüştür. Bu durumda, birden fazla boyutta yük veren maddelerin toplam test korelasyonları, ortak varyansları ve ölçek için kritik olmaları göz önünde bulundurularak maddeler anlamlı bir yapı sergileyene kadar ölçekten sırayla çıkarılmışlardır. Bunun yanı sıra, faktör yükü değeri ,32'nin altında olan maddeler de ölçekten çıkarılmıştır. Tabachnick ve Fidell (2012), ölçekteki maddelerin faktör yükü değerlerinin en az ,32 olması gerektiğini ifade etmektedirler. Yapılan analizler sonucunda, özdeğeri 1'in üzerinde olan 4 faktör ortaya çıkmıştır. Bu faktörlerin ölçeğe ilişkin açıkladıkları toplam varyans %52,906 olarak belirlenmiştir. Faktörlerin yüklerine ilişkin bilgiler Tablo 4'de yer almaktadır.

Tablo 4. Ölçeğe İlişkin Açımlayıcı Faktör Analizi Sonuçları

Maddeler	Faktör 1	Faktör 2	Faktör 3	Faktör 4
m17	,758			
m30	,707			
m38	,693			
m15	,641			
m9	,637			
m10	,591			
m42	,555			
m46		,738		
m33		,708		
m48		,702		
m31		,651		
m45		,570		
m3			,691	
m5			,690	
m6			,681	
m13			,663	
m12			,473	
m26				,686
m36				,668
m19				,665
m25				,599

Ölçeğin boyutları alanyazını ile tutarlı olacak şekilde, güdülenme (7 madde), öz-kontrol (5 madde), öz-izleme (5 madde) ve özgüven (4 madde) olarak belirlenmiştir. Boyutların dağılımına ilişkin grafik ise Şekil 1'de yer almaktadır.



Şekil 1. "Öz-Yönetimli Öğrenme Becerileri" Ölçeğine İlişkin Özdeğer Grafiği

Faktörlere ilişkin özdeğerleri içeren grafik incelendiğinde, ilk faktörden sonra keskin bir düşüş olduğu göze çarpmaktadır. Bu durum ölçeğin genel bir faktöre sahip olabileceği anlamına gelmektedir. Devamındaki üç faktörden sonraki faktörlerin 1'in altında özdeğere sahip olduğu görülmektedir. Bu durumda, ölçeğin 4 boyutlu olduğu kabul edilmiştir.

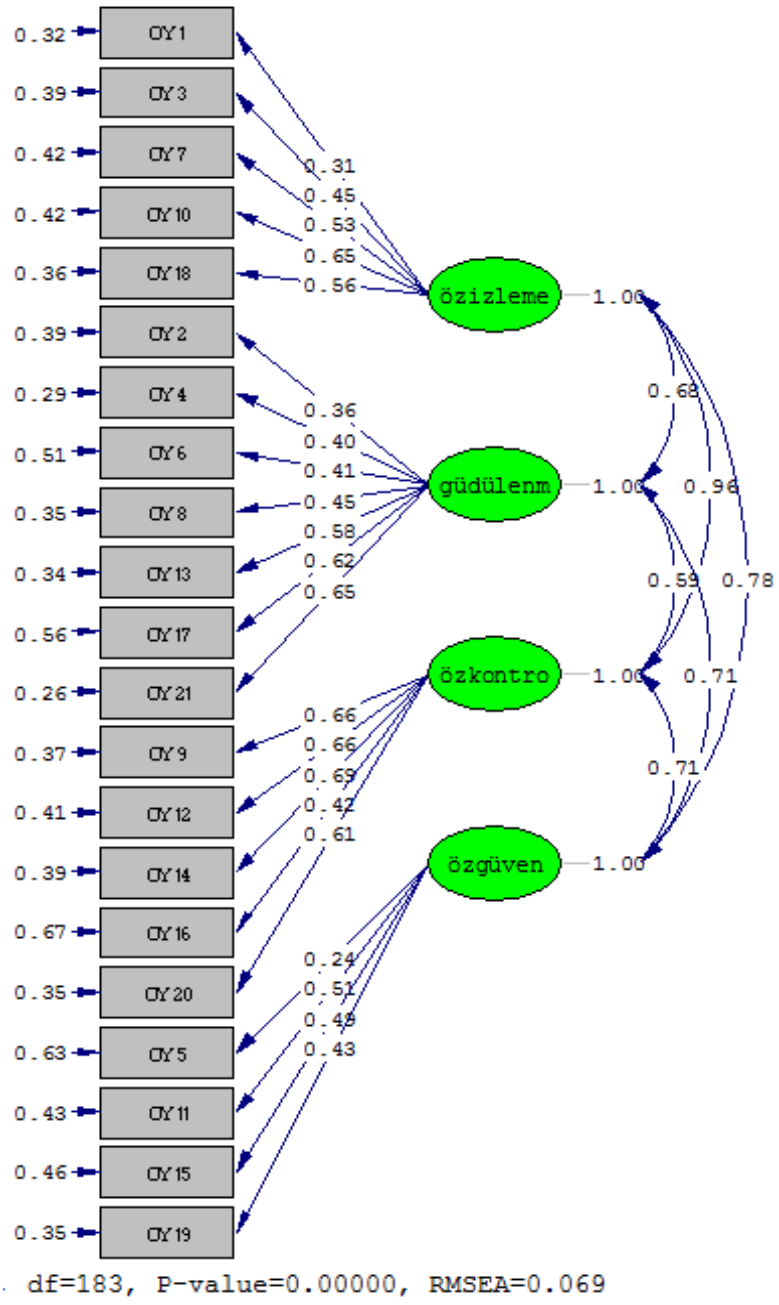
Doğrulayıcı Faktör Analizi

Ölçeğinin yapısının doğrulanması amacıyla, 2600 öğrencinin verileri doğrulayıcı faktör analizine tabi tutulmuştur. Doğrulayıcı faktör analizine ilişkin sonuçlar Tablo 5'te yer almaktadır.

Tablo 5. Doğrulayıcı Faktör Analizi Sonuçları

Boyut	Madde	Standartlaştırılmış değerler	t-değeri
Öz-İzleme	OY1	0,48	24,47
	OY3	0,59	30,68
	OY7	0,64	34,01
	OY10	0,71	39,10
	OY18	0,69	37,37
Güdülenme	OY2	0,49	24,60
	OY4	0,60	30,78
	OY6	0,50	24,74
	OY8	0,60	31,33
	OY13	0,70	37,92
	OY17	0,64	33,70
	OY21	0,78	44,11
Öz-Kontrol	OY9	0,73	40,95
	OY12	0,72	39,83
	OY14	0,74	41,60
	OY16	0,46	22,84
	OY20	0,72	40,05
Özgüven	OY5	0,29	12,84
	OY11	0,61	29,41
	OY15	0,58	27,72
	OY19	0,58	27,69

Tablo 5 incelendiğinde standartlaştırılmış yüklerin 0,29 ile 0,74 arasında değiştiği görülmektedir. Yine tabloya göre tüm t değerlerinin 1,96'nın üzerinde olduğu görülmektedir. Buna göre ölçekteki tüm maddelerin, ilgili olduğu boyutları anlamlı bir şekilde temsil ettiği söylenebilir (Şimşek, 2007). Maddelerin boyutlara göre dağılımına ilişkin veriler Şekil 2'de yer almaktadır.



Şekil 2. "Öz Yönetimli Öğrenme Becerileri" Ölçeğine İlişkin Doğrulayıcı Faktör Analizi Sonuçları

Şekil 2 incelendiğinde, RMSEA değerinin 0,069 olduğu görülmektedir. RMSEA değerinin 0,08'den küçük olması, "kabul edilebilir bir uyum"a işaret etmektedir (Browne ve Cudeck, 1992). Bunun dışında, diğer uyum indekslerine ilişkin bilgiler Tablo 6'da verilmiştir.

Tablo 6. Ölçeğe İlişkin Uyum İndeksleri

Uyum Ölçüleri	Ölçeğe İlişkin Uyum Değerleri
Kikare	2426,14
sd	183
Kikare/sd	13,26
GFI	0,92
AGFI	0,89
CFI	0,96
NFI	0,96
NNFI	0,96
SRMR	0,05
RMSEA	0,069

Tablo 6 incelendiğinde, kikare değerinin 2424,14 olduğu görülmektedir. Kikare değerinin serbestlik derecesine bölünmesi sonucunda elde edilen değer 13,26'dır. Bu değerin 5 ve altında çıkması iyi bir uyum olduğunun göstergesidir (Çokluk, Şekercioğlu ve Büyüköztürk, 2012). Değerin 5'ten yüksek çıkması istenen bir sonuç değildir. Ancak, kikare değeri örneklem büyüklüğüne çok duyarlı bir değerdir (Stapleton, 1997; Marsh, Herbert ve Balla, 1986). Bunu doğrulamak amacıyla, çalışma grubunun içerisinden random olarak seçilen 1000'er kişilik verilerle de doğrulayıcı faktör analizi yapıldığında, kikare değerinin düştüğü ve serbestlik derecesine bölündüğünde değerin 5'in altına düştüğü görülmektedir. Bu nedenle kikare değerinin serbestlik derecesine bölünmesi ile elde edilen değerin, beklenen değerlerden yüksek çıkmasının gerekçesi, çalışma grubunun büyüklüğü ile açıklanabilir. Bunun yanı sıra, doğrulayıcı faktör analizinde modele ilişkin değerlendirmenin tek bir değer üzerinden değil (özellikle kikare); birden çok uyum indeksi üzerinden yapılması gerektiği araştırmacılar tarafından da (Bentler ve Bonett, 1980; Jöreskog ve Sörbom, 1982; Stevens, 2002) önerilmektedir. Bu doğrultuda, ölçeğe ilişkin uyum indeksleri incelendiğinde; GFI değerinin 0,92 olduğu görülmektedir. Bu değerin 0,90-0,95 aralığında yer alması, kabul edilebilir bir uyum olduğu anlamına gelmektedir. AGFI değeri, 0,89 olarak hesaplanmıştır. Bu değerin 0,85 ile 0,90 arasında yer alması kabul edilebilir bir uyumunun olduğunun göstergesidir. Tabloda yer alan CFI değeri 0,96 bulunmuştur. CFI değerinin 0,95-0.97 aralığında olması kabul edilebilir uyum anlamına gelmektedir. NFI değeri 0,96'dır. Bu değerin 0,95 üzerinde olması iyi bir uyumun göstergesidir. NNFI değeri ise 0,96 olarak hesaplanmıştır. NNFI değerinin 0,95-0,97 arasında olması kabul edilebilir uyum olduğunu ifade etmektedir. Son olarak, SRMR değerinin 0,05 ve altında olması iyi bir uyumun göstergesi sayılmaktadır. Bu değerin 0,05 bulunması, ölçeğin iyi bir uyum gösterdiği anlamına gelmektedir (Schermelleh-Engel ve Moosbrugger, 2003). Sonuç olarak, ölçeğe ilişkin uyum indekslerinin iyi uyum gösteren ve kabul edilebilir değerler arasında bulunduğu ortaya konulmuştur. Buna göre, "Öz-Yönetimli Öğrenme Becerileri" ölçeğinin 21 maddeden oluşan 4 faktörlü yapısı, bir model olarak doğrulanmıştır.

Öz-Yönetimli Öğrenme Becerileri Ölçeğinin Güvenirliği

"Öz-yönetimli Öğrenme Becerileri Ölçeği"nin ölçeğin güvenirliğinin belirlenmesi amacıyla Cronbach Alpha iç tutarlık katsayısı kullanılmıştır. Ölçeğin madde-toplam korelasyonları Tablo 7'de verilmiştir.

Tablo 7. Madde Toplam Korelasyonları

	Madde Çıkarıldığında Ölçek Ortalaması	Madde Çıkarıldığında Ölçek Varyansı	Düzeltilmiş Madde-Toplam Korelasyonu	Madde Çıkarıldığında Cronbach Alpha Değeri
m3	80,52	95,231	,476	,891
m5	80,52	95,114	,527	,889
m6	80,46	94,975	,531	,889
m9	80,13	95,243	,542	,889
m10	80,22	96,018	,502	,890
m12	80,43	96,312	,507	,890
m13	80,65	94,500	,525	,889
m15	80,09	96,458	,467	,891
m17	80,36	94,266	,575	,888
m19	80,00	98,055	,360	,894
m25	80,15	95,875	,526	,890
m26	80,15	95,694	,496	,890
m30	80,26	94,938	,536	,889
m31	80,86	93,494	,553	,889
m33	80,79	92,979	,548	,889
m36	80,10	97,261	,461	,891
m38	80,22	94,554	,487	,891
m42	80,11	95,981	,464	,891
m45	80,37	96,089	,470	,891
m46	80,55	94,266	,566	,888
m48	80,79	93,771	,531	,889

Tablo 7 incelendiğinde, madde toplam korelasyonlarının ,360 ile ,566 arasında değiştiği görülmektedir. Her bir boyuta ilişkin güvenirlik katsayılarına ilişkin tablo ise aşağıda yer almaktadır.

Tablo 8. Maddelerin Güvenirliklerinin Boyutlara Göre Dağılımı

Maddeler	Faktör 1 Güdülenme	Faktör 2 Öz-Kontrol	Faktör 3 Öz-İzleme	Faktör 4 Özgüven	Toplam
m17	,487				
m30	,536				
m38	,575				
m15	,467				
m9	,542				
m10	,502				
m42	,464				
m46		,548			
m33		,531			
m48		,553			
m31		,470			
m45		,476			
m3			,527		
m5			,531		
m6			,525		
m13			,507		
m12			,496		
m26				,496	
m36				,461	
m19				,360	
m25				,526	
Cronbach Alpha	,826	,799	,768	,690	,895

Tablo 8 incelendiğinde, güdülenme boyutunun ,826; öz-kontrol boyutunun ,799; öz-izleme boyutunun ,768 ve özgüven boyutunun ,690 değerine sahip olduğu görülmektedir. Ölçeğin tamamına göre (21 madde) hesaplanan Cronbach Alpha iç tutarlık katsayısı ise ,895 olarak hesaplanmıştır. Bu sonuç, ölçeğin nihai formunun yüksek bir güvenirliğe sahip olduğu anlamına gelmektedir.

SONUÇLAR ve TARTIŞMA

Bu araştırmada, üniversite öğrencilerinin öz-yönetimli öğrenme becerilerinin belirlenmesine yönelik bir ölçek geliştirilmiştir. Ölçek geliştirilirken, geçerliğinin belirlenebilmesi amacıyla açımlayıcı faktör analizi yapılmıştır. Öncelikli olarak verilerin faktör analizi için uygun olup olmadığı Kaiser-Meyer-Olkin (KMO) katsayısı ve Barlett Sphericity testi ile belirlenmiştir. Verilerin faktör analizi için uygun olmasının ortaya konması üzerine, Varimaks rotasyonu ile Temel Bileşenler Analizi kullanılarak açımlayıcı faktör analizi yapılmıştır. Faktör analizi sonucunda, ölçeğin öz değeri 1'den büyük 10 faktörü olduğu belirlenmiştir.

Birden fazla boyutta yük veren maddeler toplam test korelasyonları, ortak varyansları, ölçek için kritik olmaları ve faktör yükü değeri ,32'nin altında olmalarına göre ölçekten çıkarılmışlardır. Bunun sonucunda, "Öz-Yönetimli Öğrenme Becerileri Ölçeği" 21 madde ve dört boyuttan oluşan son halini almıştır. Ölçeğin boyutları alanyazınla tutarlı olacak şekilde, güdülenme (Abd-El-Fattah, 2010; Alkan ve Erdem, 2013; Fisher, King ve Tague, 2001; Guglielmino, 1977; Stockdale ve Brockett, 2010; Teng, 2005), öz-izleme (Abd-El-Fattah, 2010; Alkan ve Erdem, 2013), öz-kontrol (Abd-El-Fattah, 2010; Alkan ve Erdem, 2013; Aydede ve Kesercioğlu, 2009; Fisher ve diğerleri, 2001; Guglielmino, 1977; Lee ve Kim, 2005; Lounsbury, Levy, Park, Gibson ve Smith, 2009; McCurdy, 1973; Teng, 2005; Williamson, 2007) ve özgüven (Alkan ve Erdem, 2013; Aydede ve Kesercioğlu, 2009; Oddi, 1984; Stockdale ve Brockett, 2010) olarak isimlendirilmiştir. Maddelerin boyutlara ilişkin dağılımları incelendiğinde güdülenme boyutunda yedi madde; öz-izleme boyutunda beş

madde; öz-kontrol boyutunda beş madde ve özgüven boyutunda ise dört maddenin yer aldığı görülmektedir. Bu faktörler ölçeğe ilişkin toplam varyans %52,906'sını açıklamaktadırlar. Ölçek beşli likert tipinde geliştirilmiştir ve ölçekten alınabilecek en düşük puan 21, en yüksek puan ise 105 olarak hesaplanmıştır.

Açımlayıcı faktör analizinin ardından ölçeğin yapısının doğrulanması amacıyla, 2600 üniversite öğrencisinden elde edilen veriler doğrulayıcı faktör analizi ile analiz edilmiş ve RMSEA değeri 0,069 bulunmuştur. Bu değer "kabul edilebilir bir uyum" olduğunun göstergesidir. Analiz sonucunda kikare değeri yüksek bulunmuştur. Bu durum çalışma grubunun büyüklüğü ile açıklanabilmektedir. Bunun doğrulanması amacıyla, çalışma grubunun içerisinden tesadüfi seçilen 1000'er kişinin verilerine de doğrulayıcı faktör analizi yapıldığında, kikare değerinin serbestlik derecesine bölündüğünde değerin 5'in altına düştüğü belirlenmiştir. Ölçeğin yapısının doğrulanması konusunda, diğer uyum indekslerinin de göz önünde bulundurulmasının önemli görülmesi nedeni ile birden çok uyum indeksi incelenmiştir. Uyum indekslerinin iyi uyum gösteren ve kabul edilebilir değerler arasında olması sebebiyle (GFI: ,92; AGFI: ,89; CFI: ,96; NFI: ,96; NNFI: ,96; SRMR: ,05) "Öz-Yönetimli Öğrenme Becerileri Ölçeği"nin 21 maddeden oluşan dört faktörlü yapısı, bir model olarak doğrulanmıştır. Ölçeğe ilişkin güvenirlik katsayısı ise ,895 olarak hesaplanmıştır. Bu değer, ölçeğin güvenirliğinin yüksek olduğunu göstermektedir.

Elde edilen bulgular ışığında, bu araştırmada geliştirilen "Öz-Yönetimli Öğrenme Becerileri Ölçeği"nin geçerli ve güvenilir bir ölçme aracı olduğu ve üniversite öğrencilerinin öz-yönetimli öğrenme becerilerinin belirlenmesinde kullanılabileceği düşünülmektedir. Ölçekten alınan yüksek puan öz-yönetimli öğrenme becerilerinin yüksek olduğuna, alınan düşük puan ise öz-yönetimli öğrenme becerilerinin düşük olduğuna işaret etmektedir. Bu ölçek aracılığıyla öz-yönetim becerisi düşük olan öğrenciler belirlenerek onların kendi kendilerine öğrenme becerilerini geliştirmeye yönelik uygulamalar yapılabilir.

KAYNAKÇA

Abd-El-Fattah, S. M. (2010). Garrison's model of self-directed learning: preliminary validation and relationship to academic achievement. *The Spanish Journal of Psychology, 13*(2), 586-596.

Acar, C. (2014). *Fen bilgisi öğretmen adaylarının kendi kendine öğrenme becerilerinin çeşitli değişkenler açısından incelenmesi* (Yüksek lisans tezi, Pamukkale Üniversitesi, Eğitim Bilimleri Enstitüsü, Denizli). Erişim adresi: http://tez2.yok.gov.tr

Alkan, F. ve Erden, E. (2013). Kendi kendine öğrenmenin laboratuvarda başarı, hazırbulunuşluk, laboratuvar becerileri tutumu ve endişeye etkisi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 44*, 15-26.

Aydede, M. N. ve Kesercioğlu, T. (2009). Fen ve teknoloji dersine yönelik kendi kendine öğrenme becerileri ölçeğinin geliştirilmesi. *Çukurova Üniversitesi Eğitim Fakültesi Dergisi, 3*(36), 53-61.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588-606.

Boyer, N. R., & Usinger, P. (2015). Tracking pathways to success: triangulating learning success factors. *International Journal of Self-Directed Learning, 12*(2), 22-48.

Browne, M.W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods Research. 21*(2), 230-258. doi:10.1177/0049124192021002005

Csapo, B. (2007). Research into learning to learn through the assessment of quality and organization of learning outcomes. *The Curriculum Journal, 18*(2), 195-210. doi:10.1080/09585170701446044

Çokluk, Ö., Şekercioğlu, G. ve Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik SPSS ve Lisrel uygulamaları* (3. bs). Ankara: Pegem Akademi.

Demir, Ö. ve Yurdagül, H. (2013). Self-directed learning with technology scale for young students: a validation study. *E-International Journal of Educational Research, 4*(3), 58-73.

Field, L. (1989). An investigation into the structure, validity, and reliability of Guglielmino's self-directed learning readiness scale. *Adult Education Querterly, 39*(3), 125-139. doi: 10.1177/0001848191041002004

Fisher, M., King, J., & Tague, G. (2001). Development of self-directed learning readiness scale for nursing education. *Nurse Education Today, 21*, 516-525. doi:10.1054/nedt.2001.0589

Grover, K. (2015). Online social networks and the self-directed learning experience during a health crisis. *International Journal of Self-Directed Learning, 12*(1), 1-15.

Guglielmino, L. M. (1977). *Development of the self-directed learning readiness scale* (Unpublished Doctoral dissertation). University of Georgia, Athens.

Jennett, P. A. (1992). Self-directed learning: a pragmative view. *The Journal of Contuinuing Education in the Health Professions*, *12*, 99-104. doi:10.1002/chp.4750120208

Jöreskog, K. G. & Sörbom, D. (1982). Recent developments in structural equation modeling. *Journal of Marketing Research, 19*(4), 404-416. doi: 10.2307/3151714

Kasworm, C. E. (1983). An examination of self-directed contract learning as an instructional strategy. *Innovative Higher Education*, *8*(1), 45-54. doi:10.1007/BF00889559

Kaufman, D. M (2003). Applying educational theory in practice. *British Medical Journal*, *326*, 213-216.

Knowles, M. S. (1975). *Self-directed learning: A guide for learners and teachers*. Cambridge: Englewood Cliffs.

Kocaman, G. Dicle, A. Üstün, B. ve Çimen, S. (2004*). Kendi kendine öğrenmeye hazıroluş ölçeği.* I. Aktif Eğitim Kurultayı, Dokuz Eylül Üniversitesi, İzmir.

Lee, C. H., & Kim, S. H. (2005). Development of the self-directed mathematics learning test based on Vygotsky. *Journal of Korea Society of Educational Studies in Mathematics: School Mathematics*, *7*(3), 253-268.

Lounsbury, J. W., Levy, J. J., Park, S.H., Gibson, L., W. & Smith, R. (2009). An investigation of the construct validity of the personality trait of self-directed learning. *Learning and Individual Differences, 19*(4), 411–418. doi:10.1016/j.lindif.2009.03.001

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1986). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103*(3), 391-410.

McCurdy, D. W. (1973). An analysis of qualities of self-directedness as related to selected characteristics of I.S.C.S. students. *ERIC.* Erişim adresi: https://eric.ed.gov/

Merriam, S. B., Caffarella, R. S., & Baumgartner, L. M. (2007). *Learning in adulthood*. San Francisco: Jossey-Bass.

O'Shea, E. A. (2003). Self-directed learning in nurse education: a review of the literature. *Journal of Advanced Nursing, 43*(1), 42-70. doi:10.1046/j.1365-2648.2003.02673.x

Oddi, L. F. (1984). *Department of an instrument to measure self-directed continuing learning.* (Unpublished Doctoral dissertation). Northern Illinois University, Illinois.

Salas, G. (2010). *Öğretmen adaylarının kendi kendine öğrenmeye hazırbulunuşlukları (Anadolu Üniversitesi örneği).* (Yüksek lisans tezi, Anadolu Üniversitesi, Eğitim Bilimleri Enstitüsü, Eskişehir). Erişim adresi: http://tez2.yok.gov.tr

Sasa, A. F. (2011). *Karma öğrenme temelli özel öğretim yöntemleri dersinin fen ve teknoloji öğretmen adaylarının öz yönetimli öğrenmelerine etkisi ve çevrimiçi tartışmaların içerik analizi.* (Yüksek lisans tezi, Fırat Üniversitesi, Eğitim Bilimleri Enstitüsü, Elazığ). Erişim adresi: http://tez2.yok.gov.tr

Savin-Baden, M., & Major, C. H. (2004). *Foundations of problem based learning.* Cornwall: MPG Books Ltd.

Schermelleh-Engel, K., & Moosbrugger, H. (2003). Structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*(2), 23-74.

Song, L., & Hill, J. R. (2007). A conceptual model for understanding self-directed learning in online environments. *Journal of Interactive Online Learning, 6*(1), 27-42.

Spencer, J. A., & Jordan, R. K. (1999). Learner centered approaches in medical education. *British Medical Journal*, *318*(7193), 1280-1283.

Stapleton, C. D. (1997). *Basic concepts and procedures of confirmatory factor analysis.* [Çevrimiçi: http://files.eric.ed.gov/fulltext/ED407416.pdf], Erişim tarihi: 23.06.2015.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences*. New Jersey: Lawrance Erlbaum Associates, Inc.

Stockdale, S. L., & Brockett, R. G. (2004). Development of the PROSDLS: A measure of self- direction in learning based on the personal responsibility orientation model. *Adult Education Quarterly. 20*(10), 1-20

Şahin, M. S. (2013). The adaptation of self-directed mathematics learning attitude scale into Turkish. *Eğitim ve Bilim, 38*(169), 209-223.

Şahin, E. ve Erden, M. (2008). Özyönetimli öğrenmeye hazırbulunuşluk ölçeği'nin (ÖYÖHÖ) geçerlik ve güvenirlik çalışması. *E-journal of New World Sciences Academy, 4*(3), 695-706.

Şimşek, Ö. F. (2007). *Yapısal eşitlik modellemesine giriş.* Ankara: Ekinoks.

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics*. Boston, MA: Allyn & Bacon/Pearson Education.

Teng, K. H. (2005). *Perceptions of Taiwanese students to english learning as functions of self-efficacy, motivation, learning activities and self-directed* learning (Unpublished Doctoral dissertation). University of Idaho, Idaho.

Wingate, U. (2007). A framework for transition: supporting 'learning to learn' in higher education. *Higher Education Quarterly, 61*(3), 391-405.

Williamson, S. N. (2007). Development of a self-rating scale of self-directed learning. *Nurse Researcher, 14*(2), 66-83.

EXTENDED ABSTRACT

Introduction

The ways of obtaining and using information have substantially changed as it has become accessible from the multiple sources. This, in turn, has challenged the belief that information is unchangeable, as well as the belief that authorities have access to absolute and correct information. Therefore, the importance attached to memorization faded as conceiving the nature of knowledge and learning how to reach information, in other words learning how to learn, gained ground. Learning to learn is considered basis of principles for self-directed learning.

Self-directed learning in its largest sense refers to individuals taking initiative to identify their own learning needs, determine their learning goals, define the sources they need in order to learn, choose/use appropriate learning strategies and evaluate learning outcomes with or without help from an outsider (Knowles, 1975). In self-directed learning, the responsibility to learn shifts from an external source (teacher, etc.) to the individual. The control and active involvement of the learner in the learning process is crucial in this process (Boyer and Usinger, 2015; Grover, 2015).

The aim of this study was to develop a psychometric scale that measures self-directed learning skills. In accordance with this, a validity and reliability study was conducted to develop the scale.

Method

Content and construct validity was tested to ensure the validity of the scale. For content validity, the literature was first reviewed and a conceptual framework was established to identify the basic characteristics of the scale. Besides these characteristics, similar scale development studies conducted in and outside Turkey were examined, and interviews were held with a group of university students to decide on the items of the scale. After revealing the criteria and dimensions, a 72-item pool was formed, and views were taken from 7 Curriculum and Instruction specialists, 3 Measurement and Evaluation specialists and one Turkish Language specialist. The scale was revised in line with expert views, and a 53 item version was implemented on 753 students attending Hacettepe and Başkent Universities. In order to decide whether the data were suitable for factor analysis, the data were subjected to the Kaiser-Meyer-Olkin (KMO) coefficient and Barlett's Sphericity Test.

In order to assess the construct validity of the scale, a confirmatory factor analysis was performed with the 2600 students who studied in Hacettepe and Başkent. Prior to the confirmatory factor analysis, the Mahalonobis distances were calculated to reveal the extreme values and remove them from the study group. As a result, confirmatory factor analysis was run with 2533 students.

The Cronbach Alpha internal consistency coefficient was used to explore the reliability of the 'Self-Directed Learning Skills Scale'.

Results and Discussion

After verifying fitness for factor analysis, the Varimax Rotation and Basic Components Analysis were used for the purposes of exploratory factor analysis. Following the factor analysis, the scale

was decided to have 10 factors with an eigenvalue over 1. As items that had a load in more than one dimension were critical to total test correlation and common variance, these were removed from the scale one by one until a meaningful structure was achieved. In addition, items with a factor loading value below .32 were also removed from the scale. The analyses resulted in four factors with an eigenvalue above 1. This means that the scale may have a general factor. The factors following the next three factors all have an eigenvalue below 1. These factors were ignored and the scale was accepted to have 4 dimensions. The total scale variance explained by these factors was 52,906%. Factors were named according to the literaure; motivation, self-control, self-monitoring and self-confidence.

Confirmatory factor analysis was run on data from 2533 students and RMSEA value was calculated of 0,069. An RMSEA value below 0,08 showed 'acceptable fit'. Chi square value was found 2424,14. The value obtained by dividing the chi square value by the degree of freedom was 13,26. This value being five or below shows good fit (Çokluk, Şekercioğlu and Büyüköztürk, 2012). A value over five is not desired. However, chi square value is highly sensitive to sample size (Stapleton, 1997; Marsh, Balla and McDonald, 1986). In order to confirm this, a confirmatory factor analysis was also run on data from 1,000-individual groups selected randomly from among the study sample, as a result of which the chi square value dropped and its division into the degree of freedom yielded in a value below five. Therefore, the high sum resulting from the division of the chi square value into the degree of freedom may be explained with the size of the study group. In addition, researchers also recommend that model evaluation in confirmatory factor analysis should be done not over a single value (particularly chi square), but over multiple fit indices (Bentler and Bonett, 1980; Jöreskog and Sörbom, 1982; Stevens, 2002). The fit indices of the scale were shown to range between good and acceptable values (Schermelleh-Engel and Moosbrugger, 2003). According to this, the 21-item and four-factor structure of the 'Self-Directed Learning Skills Scale' was confirmed as a model.

The internal consistency of the items for the different sub-dimensions of the scale were calculated: The motivation subdimension was .826, in self-control .799, in self-monitoring .768, and in self-confidence .690. The Cronbach Alpha internal consistency coefficient for the entire scale (21 items) was .895, suggesting that the items in the final version of the scale had high internal consistency. Exploratory factor analysis was implemented and it was found the scale consisted 21 item and 4 factors (motivation, self-control, self-monitoring and self-confidence).

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(2);101-115

Sınav Stresi Ölçeğinin Türkçeye Uyarlanması ve Ölçme Değişmezliğinin İncelenmesi

Adaptation of the Examination Stress Scale into Turkish and Examination of Measurement Invariance*

Büşra KARADUMAN** Sevilay KİLMEN***

Öz

Bu araştırmada Sung ve Chao (2015) tarafından geliştirilmiş olan Sınav Stresi Ölçeğinin (SSÖ) Türkçeye uyarlaması yapılmış, ilgili ölçekten elde edilen ölçümlerin geçerlik ve güvenirliğine ilişkin kanıtlar elde edilmeye çalışılmış ve ölçme değişmezliği test edilmiştir. Araştırmanın çalışma grubunu 2016-2017 eğitim-öğretim döneminde Balıkesir ilindeki liselerde öğrenim gören 1617 lise öğrencisi oluşturmaktadır. Araştırmanın ilk aşamasında yapılan doğrulayıcı faktör analizi sonucunda ölçeğin üç faktörlü yapısının doğrulandığı sonucuna ulaşılmıştır. Araştırmanın diğer aşamasında ise ölçeğin cinsiyet, okul türü ve sınıf düzeyi alt gruplarında ölçme değişmezliği incelenmiştir. Ölçeğin faktör yapısına ilişkin tanımlanan ölçme modelinin cinsiyet, okul türü ve sınıf düzeyi alt gruplarında ölçme değişmezliğinin sağlandığına ilişkin kanıtlar elde edilmiştir. Bu durum Sınav Stresi Ölçeğinin (SSÖ) cinsiyet, okul türü ve sınıf düzeyi alt gruplarında geçerli ve güvenilir ölçümler sağladığına işaret etmektedir.

Anahtar Kelimeler: Çoklu grup doğrulayıcı faktör analizi, ölçme değişmezliği, sınav stresi ölçeği

Abstract

In this study, Examination Stress Scale developed by Sung and Chao (2015) was adapted into Turkish, the proofs for validity and reliability were collected for the measures obtained from this scale, and measurement invariance was tested. The study group consists of 1617 students in the 9th, 10th, 11th and 12th grades of science high school, anatolian high school, technical high school and social sciences high school in Balıkesir in 2016-2017 education period. According to the confirmatory factor analysis conducted in the first phase of this research, it was found that the scale had three-factor structure. In the second phase of the study, the measurement invariance of the examination stress model was examined across the gender, school type and grade. Findings show that the Examination Stress Scale provides valid and reliable measures at different levels in sub-groups of gender, school type, and grade.

Keywords: Multi-group confirmatory factor analysis, measurement invariance, examination stress scale

GİRİŞ

Eğitim ve öğretim faaliyetleri içerisinde ölçme ve değerlendirme alanına hizmet eden sınavlar bireyin eğitim yaşantılarının her basamağında bulunmaktadır. Sözü edilen bu değerlendirilme durumunun bireyler üzerinde strese yol açtığı çeşitli araştırmalarla ortaya konmuştur (Arslan, 2016). Özellikle sınavlara gelecekteki kariyerlerini belirlemek için giren lise öğrencileri arasında sınav

**Bu çalışma, ilk yazarın, ikinci yazar danışmanlığında 2017 yılında tamamladığı "Sınav Stresi Ölçeğinin Uyarlanması ve Ölçme Değişmezliğinin İncelenmesi" isimli yüksek lisans tezinden üretilmiştir.*

***Doktora öğrencisi, Gazi Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara-Türkiye, e-posta: busra_karaduman@yahoo.com ORCID ID: orcid.org/0000-0001-7565-1025*
****Doç. Dr., Abant İzzet Baysal Üniversitesi, Eğitim Fakültesi, Bolu-Türkiye, e-posta:, sevilaykilmen@ibu.edu.tr, ORCID ID: orcid.org/0000-0002-5432-7338*

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, Cilt 9, Sayı 2, Yaz 2018, 101-115.
Journal of Measurement and Evaluation in Education and Psychology, Vol. 9, Issue 2, Summer 2018, 101-115.
DOI: 10.21031/epod.330426
Geliş Tarihi: 23.07.2017
Kabul Tarihi: 21.02.2018

stresi önemli bir konudur (Sung ve Chao, 2015). Başarı odaklı toplumlarda bireylerin sosyal ortamlarda sıklıkla değerlendirilmeye maruz kalmaları, stresin ortaya çıkmasında temel faktör olarak rol oynayabilmektedir (Arslan, 2016). Toplumun öğrenciler üzerindeki başarı odaklı zorlayıcı tutumu öğrenciler üzerinde stres ve kaygıyla sonuçlanmakta, bu durum kişinin duygusal ve akademik durumunu etkilemektedir. Bireyin yaşadığı sınav stresinin bireyin sağlığı üzerindeki olumsuz etkileri çeşitli araştırmalarda (Costarelli ve Patsai, 2012; Steptoe, Wardle, Pollard, Canaan ve Davies, 1996) rapor edilmektedir.

Kavramsal olarak dünyada sınavlara yönelik yapılan çalışmalar incelendiğinde sınav stresinden çok sınav kaygısı üzerine odaklanıldığı göze çarpmaktadır. Oysa ki, stres, fiziksel ve duygusal yüklenmeler sonucu meydana gelen zorlanmalardır ve meydana gelen gerilimler ile birlikte stres bazen istenmedik tepkileri meydana getirir. Bu tepkilerden sadece bir tanesi kaygıdır (Altuntaş, 2003; Sarason ve Sarason, 1990). En tipik kaygı tepkileri terleme, korku, endişe gibi uyarılmalardır. Birçok araştırmacı tarafından kaygı, stres tepkisinin bir boyutu olarak tanımlanmıştır (Sung ve Chao, 2015).

Türkiye'de sınavlara yönelik çalışmalar incelendiğinde sınav kaygısına odaklanıldığı, sınav stresine yer verilmediği görülmektedir. Örneğin, Pazarlı (2009) araştırmasında, öğrencilerin öğrenme stilleri ile sınav kaygıları arasındaki ilişkiyi araştırmıştır. Çapulcuoğlu ve Gündüz (2012) ise araştırmalarında, sınav kaygısı, öğrenci tükenmişliği, akademik yetkinlik ve anne-baba tutumları değişkenleri arasındaki ilişkiyi incelemiştir. Bazı araştırmalarda da çeşitli değişkenlerin sınav kaygısı üzerindeki etkileri araştırılmıştır. Örneğin, Ulusoy, Yavuz, Esen, Umut ve Karatepe (2016) araştırmalarında, bilişsel müdahalelerin davranışçı müdahalelerden bağımsız olarak sınav kaygısı üzerindeki etkililiğini araştırmışlardır. İlgili alan yazın genel olarak incelendiğinde, Türkiye'de sınav kaygısına ilişkin hem tarama hem de deneme türü araştırmaların yer aldığı ancak sınav stresi kavramına hiç değinilmediği göze çarpmaktadır. Bu araştırmada alan yazındaki bu boşluğa odaklanılmış, Sung ve Chao (2015) tarafından geliştirilen SSÖ'nün Türkçeye uyarlaması yapılmış, ardından farklı gruplarda ölçme değişmezliği incelenmiştir.

Bu araştırmanın yapılması en az iki nedenden dolayı önemlidir. Birincisi, Türkiye'de sınav stresine ilişkin araştırmalara rastlanmamış olmasıdır. Bu durumun olası nedenlerinden biri sınav stresini ölçmeye yönelik Türkiye'de geliştirilen veya Türkçeye uyarlaması yapılan herhangi bir ölçeğin olmayışı olabilir. Sınav stresi ölçeğinin geliştirilmesi veya başka bir kültürde geliştirilen sınav stresi ölçeğinin Türk kültürüne uyarlanması Türkiye'de sınav stresine ilişkin araştırmalara katkı sağlayabilir. Nitekim öğrencilerin sınav stres düzeyinin ölçülmesi için bir araç geliştirilmesi, sınav stresi ile ilgili konuları araştırmaya yönelik ilk adımdır (Sung ve Chao, 2015). İkincisi ise uyarlanan SSÖ'ye dayalı olarak yapılacak grup karşılaştırmaları için ön koşul olan ölçme değişmezliğine kanıt aranmasıdır. Araştırmacılar psikolojik ölçümlerin geçerlik ve güvenirliklerine dair kanıtlar elde ettiklerinde yapmış oldukları karşılaştırmalara güven duymaktadırlar. Oysa geçerlik ve güvenirlik ölçme aracına dayanmaktan ziyade araçtan elde edilecek ölçümlere dayanmaktadır. Bu sebepten geçerlik ve güvenirlik düzeyi ile ilgili bilgiler edinmek için hesaplanan test ve madde istatistikleri sadece grupta yer alan bireylerin özelliklerini yansıtmaktadır (Crocker ve Algina, 1986). Buradan farklı gruplarda yapılan ölçümlere ait geçerlik ve güvenirlik kanıtlarının farklılaşabileceği anlaşılmaktadır. Elde edilen ölçümlerin psikometrik nitelikleri, bireylerin farklı özellikler taşımasından kaynaklanabilir ya da ölçme aracı bu farklılaşmaların sebebi olabilir. Bu sınırlılık nedeniyle, gözlenen değişkenlerin hangi durumlarda gruplar arasında geçerli ve güvenilir olduğu yönünde bir sorun gündeme gelmiş ve alan yazında ölçme değişmezliği çalışmaları yapılmaya başlanmıştır (Vandenberg ve Lance, 2000). Bu araştırmada SSÖ için farklı gruplarda ölçme değişmezliği çalışması yapılarak ölçme aracının gruplar arası karşılaştırmalar için uygun olup olmadığı test edilmiştir. Aşağıda konuyla ilgili kavramsal çerçeveye değinilmiştir.

Stres

Stres, bedenin kendi üstündeki baskıya gösterdiği "genel uyum sendromu" adı verilen tepkidir (Selye, 1986). Selye (1973)'ye göre stres üç aşamada gerçekleşmektedir. İlk aşama organizmanın tehlikeyi hissettiği andaki durumunu anlatan alarm aşamasıdır. İkinci aşama, tehlike organizma için

devam ederken, organizmanın fiziksel olarak tepkide bulunduğu direnç aşamasıdır. Tehlikenin geçmediği durumlarda organizmanın tükendiği aşama ise tükenme aşamasıdır.

Organizmanın bedensel ve ruhsal olarak algıladığı tehditler sonucunda stres meydana gelir. Bu durum organizmada bedensel ve psikolojik olarak değişikliklere sebep olmaktadır. Stres belirtileri kendi içinde dörde ayrılmaktadır:

Duygusal (psikolojik) belirtiler

Organizma stresin neden olduğu etkilere karşı bazı duygusal tepkiler verir. Rowshan (2000) duygusal stres belirtilerine; aşırı ağlama, depresyon, duygu değişikliği, hastalıklı gibi hissetme, kabus görme, kızgınlık, sinirsel gülmeler ve üzüntü örneklerini vermiştir.

Fiziksel belirtiler

Rowshan (2000) bir bireyde stresin başlayıp başlamadığının anlaşılması için fiziksel belirtilerin gözlenmesi gerektiğini ileri sürmektedir. Fiziksel stres belirtileri için solunum hızlanması, kalp atışı ve kan basıncının yükselmesi, kasların gerilmesi, göz bebeklerinin büyümesi, baş ağrısı, mide bulantısı, göğüste ağrı, yorgunluk, uykusuzluk, zararlı alışkanlıklara yönelme, iştah bozukluğu gibi belirtiler örnek olarak verilebilir (Baltaş ve Baltaş, 1999; Hançerlioğlu, 1988; Köknel, 1998).

Zihinsel belirtiler

Bireyin verdiği duygusal tepkilerin yanı sıra, birey ciddi stres nedenleriyle karşılaştığında bilişsel bozukluklar yaşayabilir (Atkinson, Atkinson, Smith ve Bem, 1999). Zihinsel stres belirtileri için iş kalitesindeki düşüş, hatalarda artış örnek olarak verilebilir (Braham, 1998).

Sosyal belirtiler

Stres bireyin çevresindeki insanlarla ilişkilerini etkileyebilir ve bu durumda stresin sosyal belirtileri ortaya çıkar. Sosyal stres belirtileri için topluma uyum sağlayamamak, ben merkezli olmak, yalnızlık hissi, insanlara karşı hoşgörülü olamamak, insanlarla kurulan iletişimde meydana gelen bozukluklar (Rowshan, 2000), işe devamsızlık gibi (Sabuncuoğlu ve Tüz, 1995) örnekler verilebilir.

Sınav stresi, sınavlarla ilgili olarak yaşanan stresi tanımlama çabasındadır. Bu nedenle özellikleri sınavlara özgü olarak tanımlanmıştır. Aşağıda sınav stresi ile ilgili genel bilgiler verilmektedir.

Sınav Stresi

Günümüzde yaşamın her döneminde, bireyler problemli olaylarla veya durumlarla karşılaşmakta, sık sık stres kavramı dile getirilmekte ve bu kavram pek çok alanda kullanılmaktadır (Eryılmaz, 2009). Kullanılan alanlardan biri de eğitimdir. Eğitim bireyin doğduğu andan başlayarak yaşam boyu devam eden bir süreçtir. Genel olarak eğitimde stresin öğrenciler üzerinde çabuk sinirlenme, fazla uyuma, tahammül edememe, yalnızlık, okul başarısında düşme, arkadaşları ile iletişimde meydana gelen kopukluklar, öfke patlamaları gibi sonuçları olabilmektedir (Motavallı, 1997). Sınav stresi eğitim öğretim ortamlarında yaşanan stresin özel bir şeklidir. Sung ve Chao (2015), sınav stresini "fizyolojik kaygı tepkileri", "bilişsel ve davranışsal tepkiler" ve "algılanan sosyal beklenti ve sosyal kıyas" olmak üzere üç boyutta ele almaktadırlar.

Fizyolojik kaygı tepkileri boyutu fiziksel hastalık veya rahatsızlık, uyku bozuklukları ve duygusal problemler olarak tanımlanmıştır. Bilişsel ve davranışsal tepkiler boyutu sınav stresiyle tetiklenen düşünceler ve davranışlar (örneğin; sınav puanlarına ilişkin endişeler ve kendilerini yoğun bir şekilde çalışmaya zorlamak) olarak ele alınmıştır. Algılanan sosyal beklenti ve sosyal kıyas boyutu birey

tarafından algılanan sosyal beklentiler ve kişilerarası karşılaştırmayı ele almaktadır (Sung ve Chao, 2015).

Ölçme Değişmezliği

Klasik test kuramında (KTK) hesaplanan test ve madde istatistikleri uygulandığı gruptan etkilenmektedir. Araştırmacılar, aynı ölçme aracını farklı gruplar üzerinde uygulayabilirler ve cinsiyet, okul türü gibi demografik özelliklerin etkilerinin ortadan kaldırılmadığı durumlarda elde edilen sonuçlar yorumlanırken hatalar yapılabilir. Bu durum KTK'nın bir sınırlılığıdır. Bu sınırlılık günümüzde de önemli bir konu olan ölçme değişmezliğini gündeme getirmiştir (Crocker ve Algina, 1986). Ölçme değişmezliği, bir ölçeğe ait özelliklerin farklı gruplarda değişmez olup olmadığı ile ilgilenir.

Karşılaştırmalarda gizil/örtük değişkenler olarak adlandırılan bilişsel yetenekler, kişilik gibi özellikler incelenmektedir (Somer, Korkmaz, Dural ve Can, 2009). Bireysel farklılıklardan hareketle, ölçme sonuçlarında gözlemlenen farklılığın sadece bireyin özelliğinden kaynaklandığını söylemek doğru değildir çünkü ölçümlerdeki farklılıklar ölçme aracının kendisinden de kaynaklanıyor olabilir (Cheung ve Rensvold, 2002).

Bu araştırmada ölçme değişmezliği, literatürde yaygın olarak kullanılan yapısal değişmezlik, metrik değişmezlik, ölçek değişmezliği ve katı değişmezlik olmak üzere dört ayrı hipotezin test edilmesi ile yürütülmüştür. Bahsedilen dört aşama (Milfont ve Fischer, 2010) aşağıda açıklanmaktadır:

Yapısal değişmezlik (configural invariance)

Ölçme değişmezliği çalışmalarının en temel düzeyidir ve daha sınırlayıcı modellerin karşılaştırılması için test edilmesi gerekmektedir (Wu, Li ve Zumbo, 2007). Ölçme değişmezliği hipotezler geliştirilerek aşamalı olarak test edilmektedir (Steenkamp ve Baumgartner, 1998). Test edilen ilk hipotez, psikolojik ölçme aracının faktör yapısının gruplar arasında değişmez olduğudur. Yapısal değişmezliğe ilişkin kanıt elde edilirse "gruplar arasında kavramsal yapı aynıdır ve maddeler gruplar arasında aynı psikolojik yapıyı ölçüyor" denilebilir. Bu durum psikolojik ölçümler için yapı geçerliğine ilişkin kanıtlar sağlamaktadır (Vandenberg ve Lance, 2000). Eğer yapısal değişmezliğe ilişkin kanıt sağlanamazsa ölçülen yapılar gruplar arasında farklılaştığı için grup farklılıkları testleri anlamlı olmayacaktır. Yapısal değişmezliğe ilişkin kanıtlar sağlandıktan sonra ölçme model parametrelerinin değişmezliğinin yorumlanabilmesi için bir sonraki aşama olan metrik değişmezliğe ilişkin hipotez test edilir (Vandenberg ve Lance, 2000).

Metrik değişmezlik (metric invariance)

Metrik değişmezlik farklı grupların maddeleri aynı şekilde yanıtlayıp yanıtlamadıklarını test eder. Diğer bir deyişle, belli ölçek maddeleri ve bu maddelerin temelini oluşturan yapılar arasındaki ilişkinin gücünün aynı olup olmadığını sınar. Metrik değişmezlik faktör yüklerinin (λ) gruplar arasında aynı olacak şekilde sınırlandırılmasıyla test edilir (Milfont ve Fischer, 2010). Bollen (1989) faktör yüklerinin gözlenen değişkenleri gizil/örtük değişkenlere bağladığını ve bu nedenle gizil/örtük değişkendeki en küçük değişimin, gözlenen değişkeni etkilediğini belirtmiştir. Eğer metrik değişmezliğe ilişkin kanıt elde edilemezse, maddelerin tüm gruplar için anlamlarının aynı olmadığına yani maddelerin bir ya da birden fazla gruba karşı yanlı davrandığına ilişkin yorum yapılabilir.

Ölçek değişmezliği (scalar invariace)

Ölçek değişmezliği gizil/örtük ortalamaların karşılaştırılmasını gerektirir. Ölçek değişmezliğinin sağlanması gözlenen puanların ve gizil/örtük puanların ilişkili olduğunu göstermektedir. Diğer bir deyişle, gizil/örtük yapıda aynı puanı alan bireylerin –hangi gruba mensup olup olmadığına

bakılmaksızın- gözlenen değişkenler için de aynı puanı alacağı anlamına gelmektedir. Ölçek değişmezliği sabitin gruplar arasında aynı olacak şekilde sınırlandırılmasıyla test edilir (Milfont ve Fischer, 2010).

Katı değişmezlik (strict invariance)

Hata varyansları, doğrudan gözlemlenemeyen gizil/örtük değişkenlerin altında yatan değişkenliğin açıklanamayan kısmıdır ve gözlenen değişkenler arasındaki korelasyon büyüklüğünü etkilemektedirler. Katı değişmezlik aşamasında gruplar arasında ölçeğe ait hata varyanslarının değişip değişmediği test edilir. Burada önceki değişmezlik aşamalarındaki faktör yapısı, faktör yükleri ve madde sabitlerine ek olarak, hata varyanslarının da gruplar arası aynı olduğu sınırlaması getirilmektedir (Hirschfeld ve von Brachel, 2014; Milfont ve Fischer, 2010).

Araştırmanın Amacı

Bu araştırmanın amacı Sung ve Chao (2015) tarafından geliştirilen SSÖ'nün Türkçeye uyarlamasını gerçekleştirdikten sonra bu ölçme aracının faktör yapısı için oluşturulan ölçme modelinin veri ile uyum düzeyini inceleyip, aracın farklı gruplarda (cinsiyet, okul türü ve sınıf düzeyi) aynı yapıyı ölçüp ölçmediğini diğer bir deyişle gruplar arası ölçme değişmezliğini belirlemektir. Bu genel amaca ulaşmak için dört araştırma sorusu oluşturulmuştur:

1. SSÖ'nün üç faktörlü yapısı doğrulanmakta mıdır?
2. SSÖ'nün cinsiyet açısından ölçme değişmezliği sağlanmakta mıdır?
3. SSÖ'nün okul türü açısından ölçme değişmezliği sağlanmakta mıdır?
4. SSÖ'nün sınıf düzeyi açısından ölçme değişmezliği sağlanmakta mıdır?

YÖNTEM

Çalışma Grubu

Araştırmanın çalışma grubunu 2016-2017 eğitim-öğretim yılının güz döneminde Balıkesir ilindeki fen lisesi, anadolu lisesi, mesleki ve teknik anadolu lisesi ve sosyal bilimler liselerinde 9., 10., 11. ve 12. sınıfa devam eden 1617 öğrenci oluşturmuştur. Tablo 1'de öğrencilerin cinsiyet, sınıf düzeyi ve okul türlerine göre dağılımları verilmiştir.

Tablo 1. Çalışma Grubundaki Öğrencilerin Cinsiyet, Okul Türü ve Sınıf Düzeylerine göre Dağılımı

		9. sınıf		10. sınıf		11. sınıf		12. sınıf		Toplam
		E	K	E	K	E	K	E	K	
Okul türü	Fen lisesi	45	61	38	57	36	63	48	33	381
	Anadolu lisesi	58	52	53	81	80	87	37	46	494
	Mesleki ve teknik anadolu lisesi	96	40	61	73	48	80	36	52	486
	Sosyal bilimler lisesi	35	74	27	64	-	-	25	31	256
	Genel toplam	461		454		394		308		1617

E: Erkek K: Kız

Çalışma grubuna 894 (%55.3) kız öğrenci ve 723 (%44.7) erkek öğrenci dahil edilmiştir. Tablo 1'de görüldüğü gibi, öğrencilerin 381'i (%23.6) fen lisesinde, 494'ü (%30.5) Anadolu lisesinde, 486'sı (%30) mesleki ve teknik anadolu lisesinde ve 256'sı (%15.8) da sosyal bilimler lisesinde öğrenim görmektedir. Öğrencilerin sınıf düzeylerine ilişkin dağılımları incelendiğinde 9. sınıflardan 461

(%28.5), 10. sınıflardan 454 (%28.1), 11. sınıflardan 394 (%24.4) ve 12. sınıflardan 308 (%19) öğrencinin çalışmaya dahil edildiği görülmektedir.

Veri Toplama Aracı

Araştırmada Sung ve Chao (2015) tarafından geliştirilmiş olan SSÖ veri toplama aracı olarak kullanılmıştır. SSÖ beşli likert tipinde (hiç katılmıyorum:1, tamamen katılıyorum:5), 27 maddeden oluşmaktadır. Ölçekte sınav stresi; fizyolojik kaygı tepkileri, bilişsel ve davranışsal tepkiler ve algılanan sosyal beklenti ve sosyal kıyas şeklinde toplam üç boyutta ele alınmaktadır.

Fizyolojik kaygı tepkileri: Ölçekte fizyolojik kaygı tepkileri boyutu fiziksel hastalık veya rahatsızlık, uyku bozuklukları ve duygusal sıkıntı gibi tepkiler olarak tanımlanmıştır (Örnek: Sınavlara hazırlanırken, kendimi çoğunlukla fiziksel olarak rahatsız hissederim). Bahsedilen boyutta 10 madde yer almaktadır. Bu boyuta ait Cronbach alfa değeri 0.89 olarak bulunmuştur.

Bilişsel ve davranışsal tepkiler: SSÖ'de bilişsel ve davranışsal tepkiler boyutu sınav stresiyle tetiklenen düşünceler ve davranışlar olarak tanımlanmıştır (Örnek: Tüm sınavlardan aldığım puanları önemserim). Bahsedilen boyutta 8 madde yer almaktadır. Bu boyuta ait Cronbach alfa değeri 0.85 olarak bulunmuştur.

Algılanan sosyal beklenti ve sosyal kıyas: Ölçekte algılanan sosyal beklenti ve sosyal kıyas boyutu öğrenciler tarafından algılanan beklentiler ve başkalarıyla karşılaştırılmaya ilişkin olarak tanımlanmıştır (Örnek: Sınav puanlarım ile ilgili olarak ailemin beklentileri beni rahatsız eder). Bahsedilen boyutta ise 9 madde yer almaktadır. Bu boyuta ait Cronbach alfa değeri 0.88 olarak bulunmuştur (Sung ve Chao, 2015).

Verilerin Analizi

Ölçeğin dilsel eşdeğerliğini incelenmek amacıyla ölçeğin orijinal formu ile Türkçe formundan elde edilen puanlar arasındaki ilişki hesaplanmıştır. Toplam 50 kişiden elde edilen verilerle ölçeğin toplam puan ve alt boyutları için Pearson Momentler Çarpım Korelasyonu hesaplanmış ve bağımlı örneklemler *t*-testi yapılmıştır. Ölçümlerin güvenirliğine ilişkin kanıtlar elde etmek için Cronbach alfa değeri hesaplanmıştır. Ölçekteki maddelerin faktör yükleri ve hata varyanslarının değiştiği göz önüne alınarak çok boyutlu ölçekler için hesaplanan ve Cronbach alfa katsayısına göre daha güçlü bir güvenirlik değeri olarak belirtildiği için bileşik güvenirlik katsayısı (Raykov, 1997) da ek olarak hesaplanmıştır. Geçerliğe ilişkin kanıtlar elde etmek için ise doğrulayıcı faktör analizi (DFA) kullanılmıştır. Bu analiz için LISREL 8.7 paket programından yararlanılmıştır. Model veri uyumunu kestirebilmek için Ki kare, Ki kare/sd, RMSEA, CFI, NFI, IFI ve NNFI model uyum indeksleri kullanılmıştır (Hu ve Bentler, 1999). Ölçme değişmezliği ise çoklu grup doğrulayıcı faktör analizi (ÇGDFA) ile incelenmiştir (Jöreskog ve Sörbom, 1993). Ölçme değişmezliği ile ilgili analiz sonuçlarının doğru bir şekilde yorumlanabilmesi için verilerle analize geçilmeden önce varsayımların karşılanıp karşılanmadığına yönelik bazı ön analizlerin yapılması gerekmektedir. Bu nedenle araştırmada kayıp değerler, uç değerler ve normalliğe bakılmış ayrıca çoklu bağlantı sorunu da incelenmiştir. Veriler varsayımları karşıladığı için tahmin yöntemi olarak en çok olabilirlik yöntemi kullanılmıştır. ΔCFI ve ΔRMSEA (Chen, 2007; Cheung ve Rensvold, 2002) ise farklı gruplarda ölçme değişmezliğine ilişkin aşamalı testlerin analizinde karar ölçütü olarak kullanılmıştır. Chen (2007) tarafından yapılan simülasyon araştırması sonucunda 300'den büyük örneklemler söz konusu olduğunda $-0.010 \leq \Delta CFI$ ve $\Delta RMSEA \leq 0.015$ değerleri değişmezlik kararı için kesim noktası olarak önerilmektedir. Bu nedenle bu araştırmada bu değerler ölçme değişmezliğinin sağlanıp sağlanmadığına ilişkin değerlendirmelerde kesim noktası olarak kullanılmıştır.

BULGULAR

SSÖ'nün Dilsel Eşdeğerliğine İlişkin Bulgular

Uyarlama çalışması kapsamında 27 maddelik SSÖ İngilizce ile Türkçe dillerine hakim beş uzman tarafından, birbirinden bağımsız olacak şekilde Türkçeye çevrilmiştir. Çevirisi yapılan formlar birbiri ile karşılaştırılmış, benzer çeviriye sahip olduğu düşünülen maddeler ise görüş birliği ile belirlenmiştir. İngilizceden Türkçeye çevrilen form dört uzman tarafından tekrardan geri çeviri yöntemi kullanılarak İngilizceye çevrilmiştir. Düzenlenen ölçek ön uygulama için 50 kişilik bir öğrenci grubuna uygulanmış, ön uygulamadan alınan geri bildirimler sonucu düzeltmeler yapılarak ölçeğin asıl formu oluşturulmuştur. SSÖ'nün orijinal formu (İngilizce) ile Türkçe formundan elde edilen puanlar arasındaki ilişki Pearson Momentler Çarpım Korelasyonu ile bu puanlar arasında fark olup olmadığı ise bağımlı örneklemler *t* testi ile incelenmiştir.

Tablo 2. SSÖ'nün Dilsel Eşdeğerliğine İlişkin Bulgular

	Fizyolojik kaygı tepkileri	Bilişsel ve davranışsal tepkiler	Algılanan sosyal beklenti ve sosyal kıyas	Toplam
Fizyolojik kaygı tepkileri	0.93	-	-	-
Bilişsel ve davranışsal tepkiler	-	0.87	-	-
Algılanan sosyal beklenti ve sosyal kıyas	-	-	0.96	-
Toplam	-	-	-	0.97

Tablo 2 incelendiğinde ölçeğin orijinal formu ile çeviri formunun fizyolojik kaygı tepkileri, bilişsel ve davranışsal tepkiler ve algılanan sosyal beklenti ve sosyal kıyas boyutlarının korelasyon katsayıları sırasıyla 0.93, 0.87 ve 0.96 olarak bulunmuştur. Tüm ölçekten elde edilen korelasyon katsayısı ise 0.97 olarak bulunmuştur. Korelasyon katsayılarına bakıldığında katsayıların pozitif yönde yüksek olduğu görülmektedir. Ölçeğin orijinal formu ile Türkçe formunun uygulamaları arasında ilişkinin yüksek olması, ölçeğin dilsel eşdeğerliğinin sağlandığının göstergesidir. Bu bulguya ek olarak, orijinal ve alt ölçek formlarından elde edilen uygulamaların ortalamaları arasındaki farkın anlamlılığını test etmek için bağımlı örneklemler *t* testi yapılmıştır. Tablo 3'de elde edilen bulgulara yer verilmiştir.

Tablo 3. SSÖ'nün Orijinal ve Türkçe Formundan Elde Edilen t Testi Sonuçları

Faktör	Formlar	n	sd	$\bar{X}$	*t*
1.Boyut	Orijinal	50	49	31.50	1.06
	Türkçe	50	49	32.00	
2.Boyut	Orijinal	50	49	16.58	-1.65
	Türkçe	50	49	16.08	
3.Boyut	Orijinal	50	49	26.92	-1.80
	Türkçe	50	49	26.34	
Toplam	Orijinal	50	49	75.20	-1.33
	Türkçe	50	49	74.42	

p >0.05

Tablo 3'te *t* değerleri incelendiğinde SSÖ'nün orijinal formu ile çeviri formundan elde edilen puanların farklarının manidar olmadığı görülmektedir. Bu durum SSÖ'nün orijinal dilinden Türkçeye uygun şekilde çevrildiğinin ve dilsel eşdeğerliğinin sağlandığının bir göstergesidir.

Betimsel İstatistiklere ve Ayırt Edicilik Değerlerine İlişkin Bulgular

SSÖ'nün dilsel eşdeğerliği sağlandıktan sonra ölçeğin geçerliğini ve güvenirliğini sınamak için ölçeğin asıl uygulaması gerçekleştirilmiştir. Ölçek, 1617 kişiden oluşan çalışma grubuna uygulanmıştır. Tablo 4'te ölçek uygulamasından elde edilen betimsel istatistiklere yer verilmiştir. Tablo 4'e göre, SSÖ'nün her bir maddesi için hesaplanan ortalama puan 2.01 ile 3.62 arasında değişmektedir. SSÖ'den elde edilen veriler üzerinde her bir maddeye ve alt boyutlara ilişkin test istatistikleri hesaplandıktan sonra, ölçekteki her bir maddenin ayırt ediciliğini belirlemek için madde toplam korelasyonları hesaplanmıştır. Maddelerin ayırt edicilik değerlerinin 0.28 ile 0.68 arasında değiştiği görülmektedir. Maddelerin genelde yüksek ayırt edicilik düzeylerine sahip olduğu gözlenmiştir.

Tablo 4. SSÖ'nün Madde ve Alt Ölçeklerine İlişkin Hesaplanan Betimsel İstatistikler

	$\bar{X}$	Mod	Medyan	SS	Min	Max	Madde toplam korelasyonu
M1	3.26	4.00	3.00	1.26	1	5	0.55
M2	2.07	2.00	2.00	1.08	1	5	0.55
M3	2.92	3.00	3.00	1.21	1	5	0.59
M4	3.38	4.00	4.00	1.26	1	5	0.61
M5	3.62	4.00	4.00	1.18	1	5	0.65
M6	3.77	4.00	4.00	1.16	1	5	0.68
M7	3.33	4.00	4.00	1.27	1	5	0.63
M8	3.36	4.00	4.00	1.33	1	5	0.52
M9	3.10	4.00	3.00	1.32	1	5	0.64
M10	2.32	2.00	2.00	1.25	1	5	0.45
M11	2.56	1.00	1.00	0.82	1	5	0.43
M12	2.54	2.00	2.00	1.14	1	5	0.46
M13	2.01	1.00	2.00	1.09	1	5	0.45
M14	2.64	1.00	1.00	0.93	1	5	0.45
M15	2.72	2.00	3.00	1.25	1	5	0.48
M16	2.25	1.00	2.00	1.22	1	5	0.58
M17	2.04	1.00	2.00	1.17	1	5	0.49
M18	2.24	1.00	2.00	1.21	1	5	0.28
M19	2.77	2.00	3.00	1.29	1	5	0.51
M20	2.96	1.00	3.00	1.54	1	5	0.38
M21	3.09	4.00	3.00	1.27	1	5	0.48
M22	2.61	2.00	2.00	1.32	1	5	0.55
M23	2.59	2.00	2.00	1.32	1	5	0.37
M24	2.97	4.00	3.00	1.38	1	5	0.55
M25	2.68	2.00	3.00	1.33	1	5	0.56
M26	3.39	5.00	4.00	1.35	1	5	0.36
M27	3.26	5.00	3.00	1.42	1	5	0.49

Ölçümlerin Güvenirliğine İlişkin Bulgular

27 maddelik SSÖ'den elde edilen ölçümlerin güvenirliğine ilişkin kanıtlar sağlamak amacıyla Cronbach alfa katsayısı ile bileşik güvenirlik katsayısı (Raykov, 1997) hesaplanmıştır. Ölçeğin alt boyutlarından elde edilen Cronbach alfa değerlerinin 0.75 ile 0.87 arasında değiştiği gözlenmiştir. Ölçümlerin güvenirliğine ilişkin bir diğer güvenirlik ölçüsü olarak bileşik güvenirlik katsayısı hesaplanmıştır. Bu değerler Tablo 5'te gösterilmektedir.

Tablo 5. Ölçümlerin Farklı Gruplardan Elde Edilen Bileşik Güvenirlik Katsayıları

Gruplar	Fizyolojik kaygı tepkileri	Bilişsel ve davranışsal tepkiler	Algılanan sosyal beklenti ve sosyal kıyas
Kız	0.86	0.76	0.80
Erkek	0.87	0.75	0.77
Fen lisesi	0.91	0.76	0.80
Anadolu lisesi	0.88	0.78	0.80
Mesleki ve teknik anadolu lisesi	0.83	0.75	0.72
Sosyal bilimler lisesi	0.89	0.76	0.84
9. sınıf	0.88	0.78	0.78
10.sınıf	0.86	0.76	0.79
11.sınıf	0.88	0.77	0.79
12.sınıf	0.86	0.75	0.80
Tüm grup	0.88	0.77	0.79

Tablo 5 incelendiğinde hesaplanan bileşik güvenirlik katsayılarının tamamının güvenirlik düzeyinin alt sınırı olan 0.70'in üzerinde değerler aldığı görülmektedir. Bu durum ölçümlerin güvenirliğinin bir kanıtı olarak gösterilebilir.

Yapı Geçerliğine İlişkin Bulgular

Yapı geçerliğine ilişkin kanıtlar sağlamak için öncelikle ölçeğin faktör yapısına ilişkin tanımlanan ölçme modelinin veri ile uyumu DFA kullanılarak test edilmiştir. Analiz sonuçlarını doğru yorumlamak ve anlamlı çıkarımlar yapabilmek için ilk olarak veri içerisinde, kayıp verilerin olup olmadığı incelenmiştir. Araştırmada, başlangıçta lise öğrencilerinden 1640 adet veri toplanmıştır. Ancak veri incelendiğinde 23 öğrencinin, birden fazla maddeyi yanıtsız bıraktığı görülmüştür. Bu durumun analiz sonuçlarını etkileyeceği göz önünde tutulmuş, az sayıda katılımcının kayıp değere sahip olduğu düşünülerek araştırmada kayıp veri yöntemlerinden liste bazında silme yöntemi kullanılarak araştırma 1617 öğrenci üzerinden yürütülmüştür.

Kayıp veriler analiz dışında bırakıldıktan sonra araştırmada tek değişkenli uç değerler belirlenmiştir. N>100 gibi geniş örneklemlerde normal dağılım için z puan aralığının ±4 arasında olması istenmektedir (Çokluk, Şekercioğlu ve Büyüköztürk, 2010). Araştırmada kullanılan veri sayısı 1617 olduğu için örneklem büyüklüğü dikkate alınarak hesaplanan z puanlarının ±4 puan aralığında olup olmadığı kontrol edilmiştir. Elde edilen z puanlarının -2.28 ile 3.34 arasında değiştiği belirlenmiş ve ±4 sınırları dışında olan herhangi bir uç değere rastlanmamıştır. Çok değişkenli normallik varsayımı için her gözlenen değişkende tek değişkenli normallik varsayımının gözlenmesi gerekmektedir (Çokluk, Şekercioğlu ve Büyüköztürk, 2010). Bu doğrultuda araştırmada yer alan her bir bağımsız değişkendeki gözlenen değişken için basıklık, çarpıklık değerleri dikkate alınmıştır. Çarpıklık ve basıklık değerlerinin $p<0.05$ için ±1.96'dan büyük; $p<0.01$ için ±2.58'den büyük değerler alması, verideki manidar basıklık ve çarpıklığa işaret etmektedir (Harrington, 2009). Araştırmada yer alan bağımsız değişkenlerdeki gözlenen değişkenlere ilişkin basıklık ve çarpıklık değerleri hesaplanmış ve ±1.96 değerini aşan herhangi bir değer bulunmadığı için normallik varsayımı karşılanmıştır.

Çoklu bağlantı problemi değişkenler arası korelasyonların yüksek olması durumunda ortaya çıkmaktadır. Bu nedenle değişkenlere ilişkin VIF (Variance Inflation Factor; varyans şişkinlik faktörü) değeri incelenmiş ve 10'dan oldukça küçük olduğu görülmüştür. VIF değerinin 10'a eşit olması ya da daha büyük olması durumunda çoklu bağlantı probleminden bahsedilebilir. Elde edilen VIF sonuçlarına göre araştırmada çoklu bağlantı sorununun veriler için söz konusu olmadığı görülmektedir.

Ölçeğin 27 maddelik Türkçe formunun üç faktörlü yapısına ilişkin tanımlanan ölçme modelinin veriye uyumunu incelemek için yapılan DFA sonucunda elde edilen *t* değerleri, hata varyansları ve faktör yüklerine ait değerler Tablo 6'da belirtilmiştir. Tablo 6 incelendiğinde, maddelere ait faktör yüklerinin 0.29 ile 0.75 arasında değiştiği görülmektedir. Maddelere ait t değerleri 2.56'yı aştığından

27 madde için elde edilen tüm t değerleri 0.01 düzeyinde manidardır. Buradan hareketle yapılacak analizde, ölçekte bulunan herhangi bir maddenin çıkarılmasına gerek yoktur. Ancak kesin karara varmadan önce hata varyansları da incelenmelidir. Gözlenen değişkenlere ilişkin hata varyansları incelendiğinde 0.90 üzerinde sadece M18'in (0.92) yer aldığı görülmektedir. Hata varyansı yüksek olan bir madde düşük faktör yükü vermektedir. M18'in faktör yükünün 0.29 olduğu görülmektedir. Diğer maddeler ile kıyaslandığında M18 en düşük faktör yüküne sahip maddedir. M18'in hata varyansı yüksek olmasına rağmen söz konusu maddeye ilişkin manidar t değerleri elde edildiği için ve faktör yük değerinin 0.25 üzerinde olmasından dolayı maddenin model içinde yer almasına ilişkin karar verilebilmektedir (Çokluk, Şekercioğlu ve Büyüköztürk, 2010; Önen, 2009). Bu nedenle M18 ölçekten çıkarılmamıştır. Model uyum indeksleri incelendiğinde elde edilen uyum iyiliği indekslerinin kabul edilebilir aralıklarda yer aldığı söylenebilir (RMSEA <0.08, CFI>0.90, NFI >0.90, IFI >0.90 ve NNFI >0.90). Modeldeki gözlenen değişkenlerin (maddelerin) ilgili yapının iyi birer temsilcisi olduğu ve veriye uyum sergilemesi, ölçekten elde edilen ölçümlerin yapı geçerliğine ilişkin kanıtlar sağlamaktadır.

Tablo 6. Sınav Stresi Ölçme Modeline İlişkin Tüm Veriden Elde Edilen *t* Değerleri, Hata Varyansları ve Faktör Yükleri

Maddeler	*t*	Hata varyansları	Standartlaştırılmış beta katsayıları
M1		0.63	0.61
M2	22.87	0.63	0.61
M3	22.90	0.58	0.65
M4	24.33	0.54	0.68
M5	25.30	0.48	0.72
M6	26.49	0.43	0.75
M7	24.15	0.52	0.69
M8	19.38	0.68	0.56
M9	25.03	0.52	0.69
M10	17.45	0.75	0.50
M11		0.72	0.53
M12	14.96	0.71	0.54
M13	14.06	0.74	0.51
M14	15.44	0.73	0.52
M15	14.63	0.69	0.56
M16	16.98	0.51	0.70
M17	17.07	0.60	0.64
M18	8.89	0.92	0.29
M19		0.70	0.55
M20	14.49	0.83	0.41
M21	16.02	0.71	0.54
M22	17.73	0.62	0.61
M23	13.73	0.81	0.44
M24	20.75	0.62	0.62
M25	18.86	0.54	0.68
M26	13.69	0.79	0.46
M27	17.38	0.64	0.60

χ^2=2808.25, sd=321, χ^2/sd=8.74, RMSEA=0.068 (0.066-0.071), CFI=0.94, NFI=0.93, NNFI=0.93, IFI=0.94

Modelin tüm grup ile veri uyumu incelendikten sonra, ayrıca modelin her bir grupta ayrı ayrı veri ile uyumunun incelenmesi gerekmektedir. Araştırmanın amacı doğrultusunda cinsiyete, okul türüne ve sınıf düzeyine göre model veri uyumuna ilişkin değerler Tablo 7'de verilmiştir.

Tablo 7. Sınav Stresi Ölçme Modelinin Cinsiyet, Okul Türü ve Sınıf Düzeyine İlişkin Veri Uyum Değerleri

	χ^2	s.d.	χ^2/s.d.	RMSEA (CI)	CFI	NFI	IFI	NNFI
Kız	1682.78	321	5.24	0.068 (0.065-0.071)	0.93	0.92	0.93	0.93
Erkek	1471.56	321	4.58	0.069 (0.065-0.072)	0.93	0.91	0.93	0.92
Fen	956.70	321	2.98	0.068(0.063-0.074)	0.95	0.93	0.95	0.95
Anadolu	113.60	321	3.53	0.069 (0.065-0.074)	0.94	0.92	0.94	0.93
Meslek	1044.24	321	3.25	0.064 (0.059-0.068)	0.92	0.88	0.92	0.91
Sosyal	772.09	321	2.40	0.068 (0.061-0.075)	0.95	0.92	0.95	0.95
9.sınıf	1157.15	321	3.60	0.072 (0.067-0.077)	0.94	0.91	0.94	0.93
10.sınıf	1041.93	321	3.24	0.064 (0.059-0.069)	0.94	0.91	0.94	0.93
11.sınıf	973.37	321	3.03	0.070 (0.065-0.075)	0.94	0.91	0.94	0.93
12.sınıf	789.82	321	2.46	0.065 (0.058-0.071)	0.94	0.90	0.94	0.94

RMSEA güven aralığı parantez içerisinde verilmiştir.

Tablo 7 incelendiğinde, sınav stresi ölçme modelinin cinsiyet, okul türü ve sınıf düzeyi değişkenlerine ilişkin uyum indekslerinin, model veri uyumunu değerlendirmede kullanılan ölçütlerin, mesleki ve teknik anadolu lisesindeki NFI değeri hariç, kabul edilebilir aralıklarda olduğu görülmektedir (RMSEA<0.08, CFI>0.90, NFI>0.90, IFI >0.90, NNFI >0.90). Modelde kızlar ve erkekler; fen lisesi, anadolu lisesi, mesleki ve teknik anadolu lisesi ve sosyal bilimler lisesi; 9., 10., 11. ve 12. sınıflar için hesaplanan uyum indekslerinde ilgili gruplar birbiri ile karşılaştırıldığında model uyum indekslerinin benzer olduğu görülmektedir. Tüm gruplardan ayrı ayrı elde edilen model veri uyumu indeksleri sonuçlarına bakılarak cinsiyet, okul türü ve sınıf düzeyi gruplarına göre üç boyutlu bir yapı olarak sınav stresi ölçme modeli, 27 madde ve üç boyuttan oluşan bir model olarak doğrulanmıştır. Bu bulgular kızlar ve erkekler; fen lisesi, anadolu lisesi, mesleki ve teknik anadolu lisesi ve sosyal bilimler lisesi; 9., 10., 11. ve 12. sınıf öğrenci grupları için yapı geçerliğine ilişkin kanıtlar sağlamaktadır.

Ölçme Değişmezliğine İlişkin Bulgular

Modelin tüm alt gruplarda veri ile iyi uyum sergilemesinin ardından farklı gruplarda ölçme değişmezliği çalışmalarına geçilmiştir. Alt gruplara göre ölçme değişmezliği bulguları Tablo 8'de gösterilmektedir.

Tablo 8. Alt Gruplardan Elde Edilen Ölçme Değişmezliği Bulguları

	χ^2	s.d.	RMSEA (CI)	ΔRMSEA	CFI	ΔCFI
Cinsiyet						
Yapısal	3073.5	642	0.075 (0.073-0.078)	-	0.93	-
Metrik	3126.4	666	0.074 (0.072-0.077)	-0.001	0.93	0
Ölçek	3144.1	672	0.074 (0.072-0.077)	0	0.93	0
Katı	3401.4	699	0.077 (0.074-0.079)	0.003	0.92	-0.01
Okul türü						
Yapısal	3852.3	1284	0.076 (0.073-0.078)	-	0.93	-
Metrik	3959.9	1356	0.074 (0.072-0.077)	-0.002	0.93	0
Ölçek	4012.5	1374	0.074 (0.072-0.077)	0	0.93	0
Katı	4319.8	1455	0.075 (0.073-0.078)	0.001	0.93	0
Sınıf düzeyi						
Yapısal	3856.5	1284	0.076 (0.074-0.079)	-	0.93	-
Metrik	4168.2	1356	0.072 (0.070-0.074)	-0.004	0.93	0
Ölçek	3998.0	1374	0.075 (0.072-0.077)	0.003	0.93	0
Katı	4319.8	1455	0.074 (0.072-0.077)	-0.001	0.93	0

RMSEA güven aralığı parantez içerisinde verilmiştir.

Tablo 8 incelendiğinde yapısal değişmezlik aşaması için, model uyumunun değerlendirilmesinde kullanılan uyum ölçütlerinin tüm gruplar için kabul edilebilir sınırlar içerisinde yer aldığı söylenebilir (RMSEA<0.08, CFI>0.90, NFI>0.90, NNFI>0.90, IFI>0.90). Yapısal değişmezlikte modele ilişkin faktör yükleri, faktörler arası korelasyon ve hata varyansları parametreleri alt gruplarda serbest bırakıldığı için, sınav stresi ölçme modelinin yapısının cinsiyet, okul türü ve sınıf düzeyi alt gruplarında aynı olduğu söylenebilir ve söz konusu alt gruplarda yer alan öğrencilerin, ölçek maddelerine cevap verirken kullanmış oldukları kavramsal bakış açısı aynıdır. Yapısal değişmezlik aşaması sağlandığı için bir sonraki metrik değişmezlik aşamasına geçilmiştir.

Metrik değişmezlik aşamasında, alt gruplarda faktör yüklerinin aynı olma sınırlaması getirilmiştir. Elde edilen uyum ölçütleri incelenmiş ve modelin veri ile iyi uyum sergilediği sonucuna ulaşılmıştır. Metrik değişmezliği test etmek için yapısal değişmezlik ve metrik değişmezlik aşamalarında elde edilen CFI ve RMSEA değerleri arasındaki fark incelenmiş ve metrik değişmezlik için ΔCFI ve ΔRMSEA'nın kabul edilebilir sınırlar içerisinde yer aldığı görülmüştür (ΔCFI ≤0.01; ΔRMSEA ≤ 0.015). Bu bulgu modele alınan değişkenlerin faktör yüklerinin cinsiyet, okul türü ve sınıf düzeyi alt gruplarında değişmediğine işaret etmektedir. Metrik değişmezlik aşaması sağlandığı için bir sonraki ölçek değişmezliği aşamasına geçilmiştir.

Tablo 8'e göre, ölçek değişmezliği aşamasında, uyum indekslerinin kabul edilebilir sınırlar içerisinde yer aldığı söylenebilir. Ölçek değişmezliğini test etmek için yapı değişmezliğinden elde edilen CFI ve RMSEA değerleri ile ölçek değişmezliğinden elde edilen CFI ve RMSEA değerleri arasındaki farklar incelenmiştir. Bulgular incelendiğinde sınav stresi ölçme modelinin ölçek değişmezliğini sağladığı görülmüştür (ΔCFI ≤ 0.01; ΔRMSEA ≤ 0.015). Maddeler için oluşturulan regresyon denklemlerindeki sabitlerin cinsiyet, okul türü ve sınıf düzeyi alt gruplarında değişmez olduğu doğrulanmıştır. Bu bulgudan hareketle maddeler bazında herhangi bir yanlılığın bulunmadığı ve gözlenen değişkenlerdeki ortalama farklılıkların gizil/örtük yapılardaki farklılıktan kaynaklandığı sonucuna ulaşılabilir. Ölçek değişmezliği aşaması sağlandıktan sonra bir sonraki katı değişmezlik aşamasına geçilmiştir.

Tablo 8'e göre, katı değişmezlik için uyum indekslerinin kabul edilebilir sınırlar içerisinde yer aldığı söylenebilir. Yapı değişmezliği ve katı değişmezlik aşamalarından elde edilen CFI ve RMSEA değeri arasındaki fark incelendiğinde cinsiyet, okul türü ve sınıf düzeyi alt gruplarında sınav stresi ölçme modelinin katı değişmezlik aşamasını sağladığı görülmüştür (ΔCFI ≤ 0.01; ΔRMSEA ≤ 0.015).

Sınav stresine ilişkin oluşturulan ölçme modeli, tüm değişmezlik aşamalarını sağlamıştır. Bu durum ölçme modelinin cinsiyet, okul türü ve sınıf düzeyi alt gruplarında ölçme değişmezliğinin sağlandığına işaret etmektedir.

SONUÇ

Bu araştırmanın amacı doğrultusunda, ilk olarak ölçeğin İngilizceden Türkçeye çevirisinin uygun olup olmadığını belirlemek amacıyla dilsel eşdeğerlik çalışması yapılmıştır. Ölçeğin İngilizce formundan elde edilen puanlar ile Türkçe formundan elde edilen puanlar karşılaştırılmıştır. Elde edilen bulgular çerçevesinde SSÖ'nün İngilizce formunun Türkçeye uygun bir biçimde çevrildiği sonucuna ulaşılmıştır.

27 maddelik SSÖ'den ölçümlerin güvenirliğine ilişkin kanıtlar sağlamak amacıyla Cronbach alfa katsayısı ile bileşik güvenirlik katsayısı hesaplanmıştır. Ölçeğin alt boyutlarından elde edilen ve tüm ölçekten elde edilen Cronbach alpha ve birleşik güvenirlik katsayıları incelendiğinde katsayı değerlerinin yüksek olduğu görülmüştür. Elde edilen güvenirlik katsayıları, ölçeğin orijinal formundan elde edilen katsayılar ile yakın değerlere sahiptir. Elde edilen katsayıların 0.70'den düşük değerler almadığı görülmektedir. Psikolojik bir test için hesaplanan güvenirlik katsayısının 0.70 ve üstü olması durumunda elde edilen güvenirlik düzeyi yeterli görülmektedir (Büyüköztürk, 2003). Bu bulgulardan hareketle ölçeğin güvenilir bir araç olduğu söylenebilir.

27 maddelik SSÖ'nün geçerliğine ilişkin kanıtları sağlamak için tüm gruplardan elde edilen veri üzerinde DFA yapılmıştır. Oluşturulan sınav stresi ölçme modelinin tüm grup ile veri uyumu incelendikten sonra ayrıca her bir grupta ayrı ayrı veri ile uyumu incelenmiştir. Tüm gruplardan ayrı ayrı model veri uyumu indeksleri sonuçlarına bakılarak cinsiyet, okul türü ve sınıf düzeyi gruplarına göre ölçeğin sınav stresini üç boyutlu bir yapı olarak ölçebildiğine işaret eden yapı eşitliği sağlanmıştır ve sınav stresi ölçme modeli, 27 madde ve üç boyuttan oluşan bir ölçme modeli olarak doğrulanmıştır. Aynı zamanda kızlar ve erkekler; fen lisesi, anadolu lisesi, mesleki teknik ve anadolu lisesi ve sosyal bilimler lisesi; 9., 10., 11. ve 12. sınıf öğrenci grupları için yapı geçerliğine ilişkin ayrı ayrı kanıtlar sağlanmıştır.

Model veri uyumu sağlandıktan sonra cinsiyet, okul türü ve sınıf düzeyine dayalı olarak yapılan ölçme değişmezliği sonuçları ΔCFI ve ΔRMSEA değerleri incelenerek belirlenmiştir (ΔCFI <0.01, ΔRMSEA≤ 0.015). Sınav stresi ölçme modelinin cinsiyet, okul türü ve sınıf düzeyine göre alt gruplarda yapısal, metrik, ölçek ve katı değişmezlik aşamalarının tümünü sağlayarak tam değişmezlik koşulunu yerine getirdiği sonucuna ulaşılmıştır.

Genel olarak bulgular değerlendirildiğinde bu araştırmadan çıkarılacak sonuçlar şöyle sıralanabilir: SSÖ Türkçe formu kullanılmaya uygun geçerli ve güvenilir bir ölçme aracıdır. Ölçeğin orijinal ve Türkçe formundan elde edilen sonuçlar benzerdir. Ayrıca yapılan ölçme değişmezliği analizleri sonucunda ölçeğin cinsiyet, okul türü ve sınıf düzeyi değişkenlerine dayalı olarak karşılaştırılabilir olduğu ifade edilebilir.

Her araştırmada olduğu gibi bu araştırmanın da bazı sınırlıkları bulunmaktadır. Bu sınırlıklardan yola çıkılarak gelecekte yapılacak olan araştırmalara bazı önerilerde bulunulabilir. Bu çalışma Balıkesir ilinde yer alan ortaöğretim kurumlarının oluşturduğu bir çalışma grubu üzerinden yürütülmüştür. Sonuçların genellenebilirliğini arttırmak amacıyla araştırmalar, farklı okul türleri üzerinde yürütülebilir. Araştırma, çalışma grubu üzerinden gerçekleştirilmiştir. Farklı örneklem büyüklüklerinde, örneklem büyüklüğünün ölçme modeline etkisini test etmek amacıyla araştırmalar yürütülebilir. Araştırmada SSÖ kullanılarak cinsiyet, okul türü ve sınıf düzeyi değişkenleri ile farklı gruplarda ölçme değişmezliği test edilmiştir. Cinsiyet, okul türü ve sınıf düzeyi değişkenleri dışında farklı değişkenler kullanılarak ölçme değişmezliğine ilişkin araştırmalar yapılabilir.

KAYNAKÇA

Altuntaş, E. (2003). *Stres yönetimi*. İstanbul: Alfa Yayınları.

Arslan, S. (2016). *Üniversitelere hazırlanan öğrencilerde stres düzeylerinin duyguları yönetme becerisine etkisi.* Yayımlanmamış Yüksek Lisans Tezi, Nişantaşı Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul.

Atkinson, R., Atkinson, R., Smith, E., & Bem, D. (1999) *Psikolojiye giriş*. (Çev. Y. Alogan). Ankara: Arkadaş Yayınevi.

Baltaş, A., & Baltaş, Z. (1999). *Stres ve başa çıkma yolları*. İstanbul: Remzi Kitabevi.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley Interscience.

Braham, B. (1998). *Stres yönetimi* (Çev. V. Diker). İstanbul: Hayat Yayıncılık.

Büyüköztürk, Ş. (2003). *Sosyal bilimler için veri analizi el kitabı*. Ankara: Pegem.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464–504.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9* (2), 233-255.

Costarelli, V., & Patsai, A. (2012). Academic examination stress increases disordered eating symptomatology in female university students. *Eat Weight Disord, 17* (3), 164-169.

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

Çapulcuoğlu, U., & Gündüz, B. (2012). Öğrenci tükenmişliğini yordamada stresle başa çıkma, sınav kaygısı, akademik yetkinlik ve anne-baba tutumları. *Eğitim Bilimleri Araştırmaları Dergisi Uluslararası E-Dergi, 3* (1), 201–218.

Çokluk, Ö., Şekercioğlu, G. & Büyüköztürk, Ş. (2010). *Sosyal bilimler için çok değişkenli istatistik (SPSS ve LISREL uygulamaları)*. Ankara: Pegem Akademi.

Eryılmaz, A. (2009). Ergenlik döneminde stres ve başa çıkma. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi, 6* (2), 20-37.

Hançerlioğlu, O. (1988). *Ruhbilim sözlüğü.* İstanbul: Remzi Kitabevi.

Harrington, D. (2009). *Confirmatory factor analysis*. New York: Oxford University Press, Inc.

Hirschfeld, G. & von Brachel, R. (2014). Multiple-Group confirmatory factor analysis in R –A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment, Research & Evaluation, 19*(7), 1-11.

Hu, L., & Bentler, P. (1999). Cut off criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Jöreskog, K. G. & Sörbom, D. (1993). *Lisrel 8: Structural equation modeling with the simplis command language.* Lincolnwood: Scientific Software International, Inc.

Köknel, Ö. (1998). *Zorlanan insan (Kaygı çağında stres)*. İstanbul: Altın.

Milfont, T. L. & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3*(1), 111-121.

Motavallı, N. (1997). Çocukluk çağında görülen "Travma sonrası stres bozukluğunun" klinik özellikleri ve seyri. *Yeni Sempozyum, 35*, 92-95.

Önen, E. (2009). *Ölçme değişmezliğinin yapısal eşitlik modellemesi teknikleri ile incelenmesi.* Yayımlanmamış Doktora Tezi, Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Pazarlı, S. (2009). *Öğrenme stilleri ile sınav kaygısı arasındaki ilişki (İstanbul ili örneği)*. Yayımlanmamış Yüksek Lisans Tezi, Yeditepe Üniversitesi, Eğitim Bilimleri Enstitüsü, İstanbul.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*(2), 173-184.

Rowshan, A. (2000). *Stres yönetimi*. (Çev. Ş. Cüceloğlu). İstanbul: Sistem Yayıncılık.

Sabuncuoğlu, Z., & Tüz, M. (1998). *Örgütsel psikoloji*. Bursa: Alfa.

Sarason I. G., & Sarason B. R. (1990). Test anxiety. In Leitenberg H. (Ed.), *Handbook of social and evaluation anxiety* (pp. 475–496). New York, NY: Plenum Press.

Selye, H. (1973). The evolution of the stress concept. *American Scientist, 61*(6), 692-699.

Selye, H. (1986). History and present status of the stress concept. Goldberger, L. S. & Breznitsz, S. (Ed.), *Handbook of stress*. New York: The Free Press.

Somer, O., Korkmaz, M., Dural, S., & Can, S. (2009). Ölçme eşdeğerliğinin yapısal eşitlik modellemesi ve madde cevap kuramı kapsamında incelenmesi. *Türk Psikoloji Dergisi, 24*(64), 61-75.

Steenkamp, E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*(1), 78-90.

Steptoe, A., Wardle, J., Pollard, T. M., Canaan, L., & Davies, G. J. (1996). Stress, social support and health-related behavior: a study of smoking, alcohol consumption and physical exercise. *Journal of Psychosomatic Research, 41*(2), 171–180.

Sung, T. S., & Chao, T. Y. (2015). Construction of the examination stress scale for adolescent students. *Measurement and Evaluation in Counseling and Development, 48*(1), 44-58.

Ulusoy, S., Yavuz, K. F., Esen, F. B., Umut, G. & Karatepe, H. T. (2016). Sınav kaygısına yönelik bilişsel grup terapisi. *Journal of Cognitive-Behavioral Psychotherapy and Research, 5(1)*, 28-37.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-70.

Wu, D. A., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariances and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment Research and Evaluation, 12*(3), 1-26.

EXTENDED ABSTRACT

Introduction

Within the scope of this research, validity and reliability studies are carried out by adapting the Examination Stress Scale developed by Sung and Chao (2015) into Turkish and a model for examination stress is tested. In the second phase of the study, the measurement invariance was examined across gender, school type and grade.

Method

Data collection tool, the Examination Stress Scale, is a five-point likert scale consisting of 27 items. Firstly, the original form of the scale was translated into Turkish by five experts. The English and Turkish form of the scale were applied to a group knowing both languages well. Pearson Moment Product Correlation Coefficient was estimated to examine the linguistic equivalence between the scores obtained from the original form and Turkish form.

The study group consists of 1617 students in the 9th, 10th, 11th and 12th grades of science high school, Anatolian high school, technical high school and social sciences high school in Balıkesir in 2016-2017 education period. Test and item analyses were performed on the data. Item –total correlations were calculated for item discrimination and reliability was calculated by the Cronbach alpha internal consistency coefficient for all subscales of the scale separately. In addition, considering the differentiation of factor loadings and error variances, the composite reliability coefficient is calculated because it is considered that the composite reliability coefficient calculated for multidimensional scales has an appropriate reliability value than the Cronbach alpha coefficient. Confirmatory factor analysis, for the validity of the scale, was conducted to test the three-factor structure of the scale. Measurement invariance of the examination stress measurement model was examined according to the gender, school type and grade. Multi-group confirmatory factor analysis technique was used to examine the measurement invariance of the examination stress model across the gender, school type and gender. The measurement invariance tests were interpreted based on ΔCFI and ΔRMSEA values.

Results

Analyses that were conducted to examine the linguistic equivalence between the scores obtained from the original form of the scale and the scores obtained from the Turkish form showed that correlation coefficients were quite high. Correlation coefficients obtained from subscales were found as .93, .87, .96, respectively. Difference between the scores obtained from Turkish and English form was examined via the paired sample t-test. According to the result of the analysis, significant difference was not found. After the linguistic equivalency was achieved, the examination stress measurement model was tested by confirmatory factor analysis. Findings revealed that the Examination Stress Scale was composed of 27 items and three dimensions confirmed in Turkish sample. After it was determined that all the data fit well with the model, the model data fit was tested separately for each group in the sub-groups. It has been seen that modeling in each of sub-groups is well adapted. These findings were taken as evidence of validity of the scale. After this phase, multi-group confirmatory factor analysis technique was performed to examine the measurement invariance of the examination stress model across gender, school type and grade. Measurement invariance results showed that acceptable measurement invariance (configural, metric, scalar and strict invariance) exists across gender, school type and grade. The results were evaluated generally, it has been suggested that the Examination Stress Scale is a valid and reliable measurement tool and it can be used to measure examination stress at different levels in sub-groups of gender, school type, and grade.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(2);116-135

Sosyal Medya Kullanım Bozukluğu Ölçeği'nin Türk Kültürüne Uyarlanması: Geçerlik ve Güvenirlik Çalışması

The Adaptation of the Social Media Disorder Scale to Turkish Culture: Validity and Reliability Study

Hakan SARIÇAM* Fatıma Firdevs ADAM KARDUZ**

Öz

Bu çalışmada Sosyal Medya Kullanım Bozukluğu Ölçeği'nin (SMKBÖ-9) Türk ergenlerde geçerlik ve güvenirliğini test etmek ve psikometrik özelliklerini incelemek amaçlanmıştır. Araştırma kapsamına yaşları 13 ile 18 arasında değişen üç farklı çalışma grubunda yer alan toplam 586 (202+204+180) ergen alınmıştır. Açımlayıcı faktör analizi sonuçlarına göre SMKBÖ'nün Türkçe formunun orijinalindeki gibi tek boyuta sahip olduğu ve bu tek boyutlu yapının ölçtüğü özellikle ilgili toplam varyansın %48.11'ini açıkladığı görülmüştür. Doğrulayıcı faktör analizi sonucu ölçeğin uyum iyiliği değerleri χ^2/sd=1.87, RMSEA=.066, CFI=.98, GFI=.98, IFI=.98, NFI=.96, RFI=.97 ve SRMR=.039 olarak hesaplanmıştır. Ayrıca madde faktör yükleri de .35 ile .76 arasında sıralanmaktadır. Eşdeğer ölçek (ölçüt) geçerliği çalışmasında Sosyal Medya Kullanım Bozukluğu Ölçeği ile Young İnternet Bağımlılığı Testi-Kısa Formu ve Ergenler için Akıllı Telefon Bağımlılığı-Kısa Formu arasında sırasıyla *r*= .64, .66 ilişkiler tespit edilmiştir. Cronbach alfa iç tutarlık katsayısı .75, Guttman iki yarı test güvenirlik katsayısı .64 olarak bulunmuştur. Düzeltilmiş madde toplam korelasyon katsayıları .29 ile .73 arasında değişmektedir. Tüm bu sonuçlar, Sosyal Medya Kullanım Bozukluğu Ölçeği'nin Türk ergenlerde kullanılabilecek geçerli ve güvenilir bir ölçme aracı olduğunu göstermiştir.

Anahtar Kelimeler: Sosyal medya, bağımlılık, ölçek, geçerlik, güvenirlik.

Abstract

The present study aimed to test the reliability and validity of Social Media Disorder (SMD-9) Scale in Turkish adolescents and to examine its psychometric properties. Three study groups were conducted among a total of 586 (202+204+180) Turkish adolescents aged from 13 to 18. According to explanatory factor analysis results, it was found that the Turkish form of the SMDS had one-dimension just like the original version and the unidimensional scale explained 48.11% of variance related to the attribute it measured. In confirmatory factor analysis, fit index values were found as χ^2/df=1.87, RMSEA=.066, CFI=.98, GFI=.98, IFI=.98, NFI=.96, RFI=.97, and SRMR=.039. Factor loadings ranged from .35 to .76. In the concurrent validity, the Social Media Disorder Scale had significant relationships with Young's Internet Addiction Test-Short Form and Smart Phone Addiction Scale-Short Form for Adolescent (*r*= .64, .66, respectively). Cronbach alpha internal consistency coefficient was found as .75. Guttman split-half reliability coefficient was found as .64. Corrected item-total correlations ranged from .29 to .73. Overall results demonstrated that Social Media Disorder Scale can be used for Turkish adolescents as a valid and reliable instrument.

Keywords: Social media, addiction, scale, validity, reliability.

GİRİŞ

Sosyal medya ağları ya da siteleri internet teknolojisinin günlük yaşamımıza kattığı ve her geçen gün kullanıcı sayısının arttığı popüler kültürün öğelerinden biridir. Gittikçe kullanımı artan sosyal medya, karşılıklı konuşmaları geliştirerek insanların iletişim biçimini değiştiren web tabanlı bir

*Dr. Öğr. Üyesi, Dumlupınar Üniversitesi, Eğitim Fakültesi, Kütahya, Türkiye, e-posta: hakansaricam@gmail.com, ORCID ID: orcid.org/0000-0002-8723-1199
**Doktora öğrencisi, Sakarya Üniversitesi, Eğitim Bilimleri Enstitüsü, Sakarya, Türkiye, e-posta: karduzfirdevs@gmail.com, ORCID ID: orcid.org/0000-0003-1765-6287

DOI: 10.21031/epod.335607
Geliş Tarihi: 22.08.2017
Kabul Tarihi: 06.03.2018

teknolojidir ve temelde kişinin şahsi kullanımlarını, performansını ve üretkenliğini kişiselleştirebilecekleri bir alandır (Cabral, 2011). Teknoloji çağının en temel göstergelerinden biri olan sosyal paylaşım ağları her yaştan her kesimin ilgisini çekmekte; sunulan uygulamalar aracılığıyla sanal dünya reel hayatın önüne geçmektedir (Kırık, Arslan, Çetinkaya, & Gül 2015). Bunun nedenlerinden biri sosyal ağ sitelerinin (Social Network Sites-SNS), web üzerinden ücretsiz ve kolayca erişilebilir olması ve bunun da kullanıcıları cezbetmesidir. Ayrıca bu sanal ortam, yer-mekân sınırlaması olmadığı için insanlara birbirleriyle etkileşimde bulunma fırsatları oluşturmaktadır (Mahmood & Farooq, 2014). Buna ek olarak son yıllarda mobil cihaz teknolojisinin hızla büyümesi ve ilerlemesiyle birlikte, geniş kapsamlı mobil veri servislerine erişim kolaylığı, düzenli profil güncellemeleri, yorumlara ve mesajlara gerçek zamanlı olarak cevap vermeyi kolaylaştırarak günlük sosyal etkileşimin hızla gelişmesine katkı sağlamaktadır (Otu, 2015). Öte yandan günlük hayatımızda bu kadar hızlı bir şekilde kendine yer bulan sosyal medyanın eğlence, oyun, sosyal etkileşim, bilgiye kolay ulaşma gibi birçok fırsat sunmasının yanı sıra tıpkı internet, akıllı telefon, bilgisayar gibi araçlara bağımlılık oluşturulabildiği ve buna bağlı olarak Karpal Tünel Sendromu (KTS) (Collier, 2009), Göz Yorgunluğu Sendromları (Kozeis, 2009; Rosenfield, 2016) gibi sağlık problemleri oluşturma riski taşıdığı söylenebilir.

İlgili literatür incelendiğinde sosyal medyaya ve internete yönelik bağımlılık konusunun sıklıkla ele alındığı görülmektedir (Aboujaoude, 2010; Byun et al., 2009; Douglas et al., 2008; Griffiths, 2013; Kempa, 2015; Kırık ve diğerleri, 2015; Li, Zhang, Li, Zhou, Zhao, & Wang, 2016; Milani, Osualdella, & Di Blasio, 2009; Sioni, Burleson, & Bekerian, 2017; Wang & Qi, 2017; Yuan et al., 2011). Patolojik internet kullanımı ya da internet kullanım bozukluğu akıllı telefonların akıl almaz yükselişi (Erfanmanesh & Hosseini, 2015), genetik olarak bağımlılığa yatkınlık (Kormas, Critselis, Janikian, Kafetzis, & Tsitsika, 2011), mutsuz olunan gerçek hayattan kaçma arzusu (Ang, Chan, & Lee, 2017), zihinsel sağlık sorunları (Islam & Hossin, 2014; Robbins, Gillan, Smith, de Wit, & Ersche, 2012), utangaç olma (Scealy, Phillips, & Stevenson, 2002), ebeveyn sosyoekonomik statüsü (Hur, 2006), kişilerarası problemler, sosyal beceri eksikliği (Torrente, Piqueras, Orgilés, & Espada, 2014), sosyal olma arzusu (Esen & Siyez, 2011), stresli durumlardan kaçma (Shaw & Black, 2008) gibi nedenlerden kaynaklanabilmektedir. İnternet bağımlılığı kavramı ile daha çok çevrimiçi veya çevrimdışı oyun, kumar veya cinsel aktiviteler kastedilir. Kavramın öncüsü Young (1998a) ise internet bağımlılığını bir teknoloji bağımlılığı olarak adlandırmış ve internette geçirilen süreyi ayırıcı tanı olarak ele almıştır. Bununla birlikte internet bağımlılığı kavramsal olarak çok heterojen olduğu için fonksiyonel olmayan ya da bozukluk olarak adlandırılan birçok farklı bileşene sahiptir (Morahan-Martin & Schumacher, 2000; Starcevic & Aboujaoude, 2017). Bunlardan bazıları arasında video oyunları, çevrimiçi kumar oyunları ve sosyal medya kullanım süresi (Erfanmanesh & Hosseini, 2015; Griffiths, 2010; Van Rooij, Schoenmakers, Van den Eijnden, & Van de Mheen, 2010) gösterilebilir.

İnternet teknolojisinin bir yansıması olarak bilinen sosyal medya sitelerine özellikle ergenler ve gençler yoğun bir şekilde talep gösterirken; gerek Türkiye'de gerekse de dünyada sosyal medya bağımlıların (Aftab, Çelik, & Sarıçam, 2015; Andreassen, 2015; Andreassen, Pallesen, & Griffiths, 2017; Caumont, 2014; Chaffey, 2017; Shensa et al., 2017) sayısı hızla artmaktadır. Sosyal medya bağımlılığı, internet bağımlılığının evrilmiş bir başka tipi olmasının yanı sıra madde bağımlılığında gözüken benzer belirtilere sahiptir (Griffiths, 2013). Kuss ve Griffiths (2011) bu ortak belirtileri ruh halinde değişimler, duygu durumu çatışmaları, davranışsal, bilişsel ve duygusal olarak zihnin sürekli internet ile meşgul olması, sosyal medya kullanımı kısıtlandığında ya da durdurulduğunda olumsuz fiziksel ve duygusal belirtilerin görülmesi, kişisel ve kişiler arası çatışma yaşanması halinde kullanımın artması şeklinde aktarmıştır. Ryan ve diğerlerinin (2014) yaptığı araştırmada Facebook kullanımının en önemli nedenlerinin ilişki kurma, zaman geçirme, eğlence ve arkadaşlık olduğu; ayrıca bazı bağımlıların negatif ruh hallerinden kaçmak için Facebook kullandıkları bulunmuştur. Amerikan Psikiyatri Birliği (APA, 2013) DSM-V Tanı Ölçütleri Başvuru El Kitabında patolojik internet/kumar bağımlılığını bir dürtü kontrol bozukluğu ve geçici bağımlılık türü olarak belirtmiştir. Bununla birlikte sosyal medya bağımlılığını kapsam dışında bırakması sosyal medya bağımlılığının meşru bir zihinsel bozukluk olmadığı fikrini oluştursa da (Starcevic & Aboujaoude, 2017) bunun

aksini iddia eden birtakım çalışmalar da bulunmaktadır (Pantic, 2014; Ryan, Chester, Reece, & Xenos, 2014; Van den Eijnden, Lemmens & Valkenburg, 2016).

Sosyal medya bağımlılığının ele alış biçimi incelendiğinde günümüzde sosyal medya bağımlılığının sıradan bir problem olmaktan çıktığı, küresel salgın bir hastalık haline geldiği söylenebilir. Dünyanın dört bir yanındaki insanlar sosyal medyaya fazlasıyla ilgili olabilmekte ve sosyal medyayı kullanırken çok fazla zaman harcayabilmektedirler. Bundan dolayı sosyal medya dünyadaki milyonlarca insanın hayatını olumsuz etkilemektedir (Andreassen, 2015; Khan et al., 2017). İnsanlar artık her yerde, her zaman hatta hareket halindeyken bile mobil cihazlardan sosyal medyaya giriş yapmaktadır. Küresel sosyal medya araştırma raporunda Ocak 2017 itibariyle dünya genelinde 2.789 milyar aktif sosyal medya kullanıcısının olduğu, 2.549 milyar kişinin ise akıllı telefonlarından aktif olarak sosyal medya kullandığı aktarılmaktadır. Yine aynı araştırma bulgularına göre ülkemizde 48 milyon aktif sosyal medya kullanıcısı olduğu, 42 milyon kişinin aktif olarak akıllı telefonlardan sosyal medya kullandıkları tespit edilmiştir. Raporda aktarılan daha da çarpıcı bir bulgu: 1.871 milyon aktif kullanıcısı ile sosyal medya sitelerinin ilk sırasında yer alan Facebook kullanıcılarının %63'ünden fazlası Facebook sitesine günde en az sekiz defa girmektedir (Chaffey, 2017). Vishwanath'ın (2015) belirttiğine göre Facebook'ta bazı insanlar günde 8 saatini harcamaktadır. Yukarıdaki ifadelerden de anlaşılacağı üzere birçok insan için günlük sosyal medya kullanım süresinin, temel fizyolojik gereksinimleri için ayırdığı süreden daha fazla olduğu söylenebilir.

Araştırmanın Önemi ve Amacı

Sosyal medya kullanımının, günümüzdeki çocukların ve ergenlerin en yaygın faaliyetleri arasında olması (O'Keeffe, 2011), beraberinde problemli internet kullanımı sorununu ortaya çıkarmaktadır (Günlü & Ceyhan, 2017). Sosyal medya kullanım süresi arttıkça genç yetişkinlerde depresyon (Lin et al., 2016; Moreno et al., 2011; Sarıçam, Tarhan, & Soyuçok, 2015), anksiyete (Seabrook, Kern, & Rickard, 2016; Vannucci, Flannery, & Ohannessian, 2017), stres (Nabi, Prestin, & So, 2013; Rus & Tiemensma, 2017), ruminasyon (Davila et al., 2012; Shaw, Timpano, Tran, & Joormann, 2015), uyku bozukluğu (Levenson et al., 2016; Tavernier & Willoughby, 2014), sosyal kaygı (Chiou, Lee, & Liao, 2015; Dobrean & Păsărelu, 2016; Sarıçam et al., 2015), narsizim (Andreassen, Pallesen, & Griffiths, 2017), kıskançlık (Appel, Crusius, & Gerlach, 2015; Tandoc, Ferrucci, & Duffy, 2015), akademik başarısızlık (Owusu-Acheaw & Larson, 2015) gibi duygu, davranış ve kişilik bozukluklarının görülme olasılığı artmaktadır. Bununla birlikte yine sosyal medya kullanımının artması ergenlerde benlik saygılarının düşmesine (Hawi & Samaha, 2016), olumsuz benlik imajı oluşmasına (Dumitrache, Mitrofan & Petrov, 2012), ruhsal sağlık problemlerine (Best, Manktelow, & Taylor, 2014; Frost & Rickwood, 2017; Saraji & Fini, 2017; Sarıçam, Yaman, & Çelik, 2016, Wegmann, Stodt, & Brand, 2015), akademik başarısızlığa (Ahn, 2011; Owusu-Acheaw, & Laron, 2015), kortizol eksikliğine (Morin-Major et al., 2016) sebep olabilmektedir. Dolayısıyla sosyal medyanın yaygınlaşması göz önüne alındığında, sosyal medyanın davranış bozuklukları, kişilik problemleri, psikolojik sağlık ve akademik performans ile ilişkisinin nedenlerini belirlemek ve açıklamak, sosyal medya kullanımını ve bu problemleri ele alan müdahale yöntemlerini belirlemede kritik önem taşımaktadır.

Her yaş grubundaki bireylerin sosyal medyada fazla zaman geçirmeleri ve buna bağlı olarak kişisel, sosyal, eğitsel ve mesleki sorumluluklarını yerine getirmede problem yaşamaları kaçınılmaz olmaktadır (Aftab, Çelik, & Sarıçam, 2015; Şahin & Yağcı, 2017). Benzer durum tüm teknoloji kullanım orijinli bağımlılıklar için söylenebilir. Kırık ve diğerlerinin (2015) yaptığı çalışmada, 14 yaş grubundaki bireylerin sosyal medya bağımlılığı puanları düşük, fakat 17 yaşına kadar yaş arttıkça bağımlılığın arttığı, 18 yaşında tekrar düşüşe geçtiği bulunmuştur. Ekşi ve Çiftçi'ye (2017) göre ergenlerde problemli internet kullanımının en büyük nedenleri arasında ahlaki olgunluğun düşük olması gelmektedir. Sosyal medyayı kullanan ergenlerde yetişkinlerin fark ettiğinden de fazla risk faktörleri bulunmaktadır. Özellikle bu risk faktörlerine akranlar arası uygunsuz içerik, çevrimiçi gizlilik konularında bilgi eksikliği ve reklam gruplarının dış etkileri alanlarında rastlanılmaktadır (O'Keeffe, 2011). Bu bilgilerden yola çıkarak çocuk ve gençlerin fiziksel, psikolojik ve sosyal sağlıkları açısından uygun ortamlar olmaması nedeniyle anne-babaların sosyal medya sitelerinin

içinde barındırdığı risklerin farkında olmaları gerekmektedir. Ergenlerin sosyal medyaya ne ölçüde bağımlı olduklarını tespit edip, doğru tanımlayabilmek için geçerlik ve güvenirlik çalışmaları yapılmış ölçeklere ihtiyaç duyulmaktadır.

Türkiye'de konu ile ilgili yapılmış benzer ölçek uyarlama ve geliştirme çalışmaları bulunmaktadır (Otrar & Argın, 2015; Şahin & Yağcı, 2017; Şişman Eren, 2014; Türkyılmaz, 2015), fakat sosyal medya bağlamındaki bu ölçme araçları incelendiğinde bir kısmı tutuma yönelik (Otrar & Argın, 2015), bir kısmı yetişkinlerdeki bağımlılığı değerlendirmeye yönelik (Şahin & Yağcı, 2017), bir kısmı kullanım amacına yönelik (Şişman Eren, 2014), bir kısmı da spesifik sosyal medya sitelerine (ör: Facebook bağımlılığı) yöneliktir (Türkyılmaz, 2015).

Bu çalışmanın amacı, ergenlerin sosyal medya bağımlılıklarının tespit edilmesini sağlayacak Sosyal Medya Kullanım Bozukluğu Ölçeği'ni (SMKBÖ-9) Türkçeye uyarlamaktır. Ölçeğin yukarıdaki ölçme araçlarından farkları şu şekilde sıralanabilir: 1-Ergenlere yönelik olması, 2-Tutumları değerlendirmiyor olması, 3-Maddelerin DSM-IV ve DSM-V tanı kriterleri ve önceki bağımlılık ölçek çalışmalarına (Griffiths, 2005; Griffiths et al., 2016; Griffiths, Kuss, & Demetrovics, 2014; Lemmens, Valkenburg, & Gentile, 2015; Lemmens, Valkenburg, & Peter, 2009; Van Rooij et al., 2010; Yen et al., 2007; Young, 1998a) dayandırılarak hazırlanması, 4-Muhtemelen en önemli farkı, madde puanlamasının (derecelendirmesinin) kesin net aralıklar içermesi (Ör: "7"= Günde 40 kereden fazla), bundan dolayı teşhis ve tanılama yapabilmesi, 5-Madde sayısının az olması, 6-Basit ve kolay değerlendirilebilir olması. Dolayısıyla SMKBÖ-9'un Türkiye'deki alanyazına önemli katkılar sağlayacağı düşünülmektedir.

YÖNTEM

Çalışma Grubu

Ölçeğin psikometrik özelliklerinin tespiti için bir tanesi büyükşehir, iki tanesi orta ölçekli olmak üzere üç farklı ilden (202+204+180) yaşları 13 ile 18 arasında değişen kolay ulaşılabilirlik örnekleme ile seçilmiş 586 ortaokul ve lise öğrencisinden elde edilen verilerden yararlanılmıştır. Katılımcıların 91'i (%15.53) 13 yaşında, 80'i (%13.65) 14 yaşında, 118'i (%20.14) 15 yaşında, 104'ü (%17.75) 16 yaşında, 105'i (%17.92) 17 yaşında ve 88'i (%15.02) 18 yaşında olup yaş ortalaması 15.54'tür (SD=1.65). İlk grupta 113 (%19.28) kız ve 89 (%15.19) erkek, ikinci grupta 108 (%18.43) kız ve 96 (%16.38) erkek, üçüncü grupta 66 (%11.26) kız ve 114 (%19.45) erkek öğrenci bulunmaktadır. Ergenlerin 549'unun Facebook, WhatsApp, Instagram, Twitter vb sosyal medya sitelerinde en az bir tane hesabı bulunmaktadır. Katılımcıların 37 tanesi herhangi bir sosyal medya hesabının olup olmadığı ile ilgili dönüt vermemiştir. Çalışma grubundan 485 katılımcı sosyal medya sitelerine akıllı telefonlardan, 64 katılımcı ise bilgisayar, tablet, internet kafe gibi araçlar ile ulaşmaktadır.

Veri Toplama Araçları

Sosyal Medya Kullanım Bozukluğu Ölçeği-SMKBÖ (The Social Media Disorder Scale)

Ergenlerin sosyal medyaya bağımlılık düzeylerini ölçmek için Van den Eijnden, Lemmens ve Valkenburg (2016) tarafından geliştirilen SMKBÖ, dokuz maddelik kısa ve 27 maddelik uzun olmak üzere iki ayrı formdan oluşmaktadır. Maddeler hazırlanırken genellikle DSM-IV'teki Patolojik Kumar Oynama Bağımlılığı başlığındaki ve DSM-V'teki Internet Kumar Bozuklukları başlığındaki ölçütler temel alınmış ve toplamda dokuz ölçüte (Meşguliyet, dayanma, yoksunluk, ısrar, kaçış, problemler, aldatma, yer değiştirme, çatışma) göre madde havuzu oluşturulmuştur. Dokuz maddelik kısa formda her ölçüt için birer madde; 27 maddelik uzun formda her ölçüt için üçer madde bulunmaktadır. Ölçek "0=Hiçbir zaman"- ile "7= Günde 40 kereden fazla" arasında 8'li derecelendirmeye sahip olup; ölçekten 0 ile 63 arasında puanlar alınmaktadır. Yazarlar ölçeğin geçerlik ve güvenirlik çalışmaları için 3 farklı gruptan yaşları 10-17 arasında değişen toplam 2198 (724+873+601) Hollandalı ergenden çevrimiçi olarak elde ettikleri verilerden yararlanmışlardır.

Katılımcılara öncelikle Facebook, WhatsApp, Instagram, YouTube, Twitter vb sosyal medya sitelerinde hesapları olup olmadıkları ile bu hesapları ne kadar sıklıkla ve nasıl kullandıkları sorulmuştur. Yakınsak ve ölçüt geçerliği için Kompulsif Internet Ölçeği, Sosyal Medya Bağımlılığı Ölçeği, Benlik Saygısı Ölçeği, Depresyon Ölçeği, Dikkat Eksikliği Ölçeği, Dürtüsellik Ölçeği ve Yalnızlık Ölçeği kullanılmıştır. Ölçeğin 27 maddelik 9 alt boyutlu uzun formu için ikinci düzey faktör analizi sonucu uyum indeksi değerleri birinci grupta χ^2(288, n=724)=672.424, p< 0.001, CFI=0.963, RMSEA= 0.043 (90% CI:0.039=0.047), üçüncü grupta χ^2(288, n=601)=570.681, p<0.001, CFI=0.973, RMSEA=0.040 (90% CI: 0.036=0.045) olarak hesaplanmıştır. Cronbach alfa iç tutarlık güvenirlik katsayıları ilk grup için .90, üçüncü grup için .92 olarak bulunmuştur. Ölçeğin 9 maddelik kısa formu için birinci düzey faktör analizi sonucu uyum indeksi değerleri birinci grupta χ^2(27, n= 873)=62.852, p=0.001, CFI= 0.997, RMSEA= 0.041 (90% CI:0.028=0.055), üçüncü grupta χ^2(27, n=601)=54.129, p=0.002, CFI= 0.989, RMSEA= 0.041 (90% CI:0.025=0.057).olarak hesaplanmıştır. Ayrıca Cronbach alfa iç tutarlık güvenirlik katsayıları ilk grup için .76, üçüncü grup için .82 olarak bulunmuştur. Ölçeğin kısa ve uzun formu arasındaki korelasyon katsayısı *r*= .94'tür. Kısa formun 238 ergenden 2 ay arayla hesaplanan test tekrar test güvenirlik katsayısı *r*=.50 olarak hesaplanmıştır. Dokuz maddelik ölçeğin kısa formunun duyarlık ve ayırt edicilik analizlerinde 3 gruptan sosyal medya bağımlısı olan 217 (53, 101, 63) ergen seçilmiş ve bu ergenlerden elde edilen verilerden yararlanılmıştır. Duyarlılık katsayıları ilk grup için .59 ile .87, ikinci grup için .50 ile .79, üçüncü grup için .47 ile .81 arasında değişmektedir. Ayırt edicilik katsayıları ise ilk grup için .82 ile .97, ikinci grup için .83 ile .93, üçüncü grup için .82 ile .96 arasında sırlanmaktadır (Van den Eijnden, Lemmens, & Valkenburg, 2016). Tüm bu bulgulara dayanarak ölçeğin sosyal medya kullanım bozukluğunu ölçmede geçerli, güvenilir, hassas ve ayırt ediciliği yüksek bir ölçme aracı olduğu söylenebilir.

Young İnternet Bağımlılığı Testi Kısa Formu (YİBT-KF)

Ergenlerin internet bağımlılığı düzeylerini ölçmek amacıyla kullanılan *YİBT-KF* Young (1998a; 1998b) tarafından geliştirilmiş, Pawlikowski, Altstötter-Gleich ve Brand (2013) tarafından kısa forma dönüştürülmüştür. YİBT-KF, 12 maddeden oluşmakta olup beşli Likert (1=Hiçbir zaman, 5=Çok sık) tipi bir ölçektir. Ölçme aracının Türk kültürüne uyarlaması Kutlu, Savcı, Demir ve Aysan (2016) tarafından yapılmıştır. Açımlayıcı faktör analizi (AFA) sonucunda, üniversite öğrencilerinde özdeğeri 4.7 olan ve toplam varyansın %39.52'sini, ergenlerde ise özdeğeri 5.7 olan ve toplam varyansın %48.9'unu açıklayan bir yapı elde edilmiştir. Üniversite öğrencilerinde YİBT-KF'nin faktör yükleri .46 ile .72, ergenlerde ise .56 ile .82 arasında değişmektedir. Doğrulayıcı faktör analizi (DFA) sonucu uyum indeksi değerleri χ^2=144.93, sd=52, RMSEA=.072, RMR=.70, GFI=.93, AGFI=.90, CFI=.95 ve IFI=.91 olarak bulunmuştur. Madde faktör yükleri üniversite öğrencilerin de .33 ile .67 arasında sıralanmaktadır. Ergenlerde yapılan DFA sonucunda tek boyutlu YİBT-KF modelinin uyum indeksi değerleri χ^2=141.93, sd=51, RMSEA=.080, GFI=.90, CFI=.90 ve IFI=.90 olarak saptanmıştır. YİBT-KF'nin ergenlerde faktör yükleri .49 ile .71 arasında değişmektedir. Cronbach alfa iç tutarlık katsayısı üniversite öğrencilerinde .91, ergenlerde .86 olarak bulunmuştur. YİBTKF'nin test-tekrar test güvenilirlik (korelasyon) katsayısı üniversite öğrencilerinde .93, ergenlerde ise .86 olarak bulunmuştur. Ölçme aracının düzeltilmiş madde toplam korelasyon katsayılarının üniversite öğrencilerinde .36 ile .62 arasında, ergenlerde ise .47 ile .65 sıralandığı tespit edilmiştir (Kutlu ve diğerleri, 2016). Bu çalışmada ölçeğin Cronbach alfa iç tutarlık güvenirlik katsayısı .81 olarak hesaplanmıştır.

Ergenler için Akıllı Telefon Bağımlılık Ölçeği - Kısa Formu

Ölçeğin orjinali ergenlerin akıllı telefon bağımlılık düzeylerini ölçmek amacıyla Kwon, Kim, Cho ve Yang (2013) tarafından geliştirilmiş olup; toplamda 10 madde ve tek boyuttan oluşan beşli Likert tipi (1- Kesinlikle Katılmıyorum, 2- Kısmen Katılmıyorum, 3-Kararsızım, 4-Kısmen Katılıyorum, 5-Kesinlikle Katılıyorum) bir değerlendirme aracıdır. Ölçeğin Türk kültürüne uyarlaması Akın, Altundağ, Turan ve Akın (2014) tarafından gerçekleştirilmiştir. DFA sonucu ölçeğin uyum indeksi

değerleri χ^2= 56.92, sd= 31, RMSEA=.052, NFI=.96, CFI=.98, IFI=.98, RFI=.94, GFI=.96 ve SRMR=.040 olarak bulunmuştur. Ölçeğin Cronbach alfa iç tutarlık güvenirlik katsayısı .88 ve düzeltilmiş madde toplam korelasyonlarının .43 ile .76. arasında sıralandığı görülmüştür (Akın ve diğerleri, 2014). Bu çalışmada ölçeğin Cronbach alfa iç tutarlık güvenirlik katsayısı .79 olarak hesaplanmıştır.

İşlem

Sosyal Medya Kullanım Bozukluğu Ölçeğinin uyarlama çalışması için 28 Şubat 2017 tarihinde ölçeği geliştiren Regina Van den EIJNDEN ile e-mail yoluyla iletişim kurulmuş ve ölçeğin Türkçeye uyarlanabileceğine ilişkin gerekli izin alınmıştır. Ölçeğin Türkçeye çevrilme süreci belli aşamalardan oluşmaktadır. Öncelikle ölçek maddeleri araştırmacılar tarafından Türkçeye çevrilmiş ve ilgili alandan yurtdışı doktoralı dokuz uzmana gönderilerek görüş istenmiş; dilsel kapsam eşdeğerliğine bakılmıştır. Aynı uzmanlardan elde edilen uzman görüş formları üzerinde beyin fırtınası yöntemi ile anlam ve gramer açısından gerekli düzeltmeler yapılmış ve deneme Türkçe formu elde edilmiştir. Hazırlanan deneme formu, Kişisel Bilgi Formu ve diğer ölçme araçları bir araya getirilerek uygulama formu oluşturulmuş; bu form çoğaltılarak ortaokul ve lise öğrencilerine okul rehber öğretmenleri tarafından dağıtılmıştır. Uygulama sırasında gerekli açıklamalar yapılmış, kendilerine uygun bir şekilde yanıtlamaları için yönerge okunarak açıklanmış, isim yazmalarının zorunlu olmadığı, isteyenlere araştırma sonucu hakkında bilgi verilebileceği açıklanarak, cevaplarda içten olunmasının en doğru sonuca ulaşmayı sağlayacağı ifade edilmiştir. Uygulamalar yaklaşık olarak 15 dakika sürmüştür. Dağıtılan bu formlar toplanarak, verilerin bilgisayar ortamına aktarılması sağlanmıştır.

Verilerin Analizi

Veri setinin normal dağılım sergileyip sergilemediğini test etmek ve uç değerleri tespit etmek için Kolmogorov-Smirnov, Cook's ve Leverage değerlerine bakılmıştır. Veri setinin normal dağılım sergilemediği ve uç değer olmadığı görülmüştür. SMKBÖ'nün psikometrik özelliklerini tespit etmek için kapsam dil geçerliği, yapı geçerliği, ölçüt bağıntılı geçerlik, ayırt edicilik, iç tutarlık güvenirliği, iki yarı test güvenirliği (eşdeğer yarılar) ve madde analizi yöntemleri kullanılmıştır. SMKBÖ'nün yapı geçerliği için elde edilen veriler üzerinde AFA ve DFA yapılmıştır. Bir veri grubuna faktör analizi uygulanabilmesi için verinin faktör analizine uygunluğu ve örneklemin yeterli olması gerekmektedir (Özdamar, 2013). Bundan dolayı öncelikle Bartlet Küresellik Testi ve Kaiser- Meyer Olkin (KMO) Testi sonuçlarına bakılmıştır. AFA yapılmasının bir diğer nedeni sürecin doğası hakkında kuramı test etmek ve gözlenen değişkenleri kullanarak, sürecin temeli için bir işevuruk tanım yapmaktır (Tabachnick & Fidell, 2014). DFA kullanılmasının nedeni ise kuramsal olarak ortaya konan faktörleri belirlemede rol oynayan değişkenler ile AFA ile belirlenen faktörleri oluşturan orijinal değişkenler arasında uyumluluk bulunup bulunmadığını DFA ile test etmektir (Özdamar, 2013). DFA uyum indeksleri değerlendirilirken CFI (Comparative Fit Index-Karşılaştırmalı Uyum İndeksi), GFI (Goodness of Fit Index-Düzeltilmiş İyilik Uyum İndeksi), IFI (Incremental Fit Index-Artan Uyum İndeksi), NFI (Normed Fit Index -Normlaştırılmış Uyum İndeksi), RFI (Relative Fit Indix-Göreceli Uyum İndeks), RMSEA (Root Mean Square Error of Approximation-Yaklaşık Hataların Ortalama Karekökü) ve SRMR (Standardized Root Mean Square Residual-Standartize Edilmiş Hataların Ortalama Karelerinin Karekökü) değerlerinden faydalanılmıştır. Modelin iyi uyum kriterleri olarak $\chi^2/sd \leq 3$, CFI, GFI, IFI, NFI, RFI $\geq .90$, RMSEA $\leq .80$ değerleri kabul edilmiştir (Çokluk, Şekercioğlu ve Büyüköztürk, 2010; Schermelleh-Engel, Moosbrugger, & Müller, 2003). Thompson'a (2004) göre uyum iyiliği değerleri $\chi^2/sd \leq 2$, CFI, GFI, IFI, NFI, RFI $\geq .95$, RMSEA $\leq .50$ olursa çok iyi uyumu göstermektedir. Ölçüt bağıntılı geçerlik için YİBT-KF ve EATBÖ-KF kullanılmıştır. İç tutarlık güvenirlik testi için Cronbach alfa değeri ve iki yarı test güvenirliği için Spearman-Brown katsayıları (Guttman split-half katsayısı) hesaplanmıştır. Bland ve Altman (1997) Cronbach alfa değerinin karşılaştırma gruplarında .70 ile .80 arasında, klinik uygulamalarda .90 ve üzeri bir değer alması gerektiğini vurgulamıştır. Bununla

birlikte Spearman-Brown iki yarı test güvenirlik katsayısının .70'ten büyük olması gerektiği aktarılmıştır (Walker, 2006). Madde analizlerinde düzeltilmiş madde toplam korelasyon katsayıları ve alt-üst %27 değerleri incelenmiştir. Analizlerde güven aralığı olarak %95 ($p<.05$) ölçüt alınmış; verilerin çözümlenmesinde SPSS 17 ve Lisrel 9.1 istatistik paket programları kullanılmıştır.

BULGULAR

Geçerlik çalışmaları kapsamında dil kapsam ve yapı geçerliği sınanmıştır. Güvenirlik çalışmaları kapsamında iç tutarlık güvenirliği ve yarı değer eşler güvenirliği hesaplanmıştır. Bununla birlikte madde analizleri de gözlemlenmiştir.

Dil Kapsam Geçerliği

Ölçeğin dil kapsam geçerliği çalışmasında Lawshe (1975) ve Davis (1992) teknikleri kullanılmıştır. Lawshe (1975) tekniğinde her bir madde için uzmanlar maddeleri üçlü derecelendirmeyle puanlamaktadır. Kapsam geçerlik oranlar (KGO) değeri, herhangi bir maddeye ilişkin "uygun" görüşünü belirten uzman sayılarının, maddeye ilişkin görüş belirten toplam uzman sayısına oranının 1 eksiği ile elde edilir. Bu bağlamda şimdiki çalışmada dörtlü derecelendirme kullanılmış olsa da "uygundur" diyen uzman sayısı önemli olduğundan Lawshe tekniğini kullanmada bir sakınca görülmemiştir. Uzmanlardan gelen dönütlerde maddelere ait dil kapsam geçerlik oran değerleri 0.78 ile +1.00 arasında sıralanmaktadır. Veneziano ve Hooper'a (1997) göre 9 uzman için KGO değerlerinin $p<.05$ önem düzeyinde minimum 0.75 olması gerekmektedir. Bu bağlamda değerlendirildiğinde SMKBÖ'nün dil kapsam geçerliği için kabul edilebilir ölçütlere sahip olduğu söylenebilir. Davis (1992) tekniğinde ise dörtlü derecelendirme olarak 1)"maddenin uyarlaması uygun", 2)"maddenin uyarlaması uygun fakat biraz düzeltme gerekli", 2)"maddenin uyarlaması uygun fakat ciddi düzeltme gerekli" 4)"maddenin uyarlaması uygun değil" derecelendirme ifadeleri kullanılmakta, kapsam geçerlik indeksi (a) ve (b) seçeneğini işaretleyen uzmanların sayısı toplam uzman sayısına bölünerek hesaplanmaktadır. Maddelere ait dil kapsam geçerlik indeks değerleri 0.89 ile 1.00 arasında değişmektedir. Davis'e (1992) göre kapsam geçerlik indeks değerlerinin .80'den büyük olması gerekmektedir. Dolayısıyla SMKBÖ'nün dil kapsam geçerlik indekslerinin yeterli olduğu söylenebilir.

Yapı geçerliği

Açımlayıcı faktör analizi (AFA)

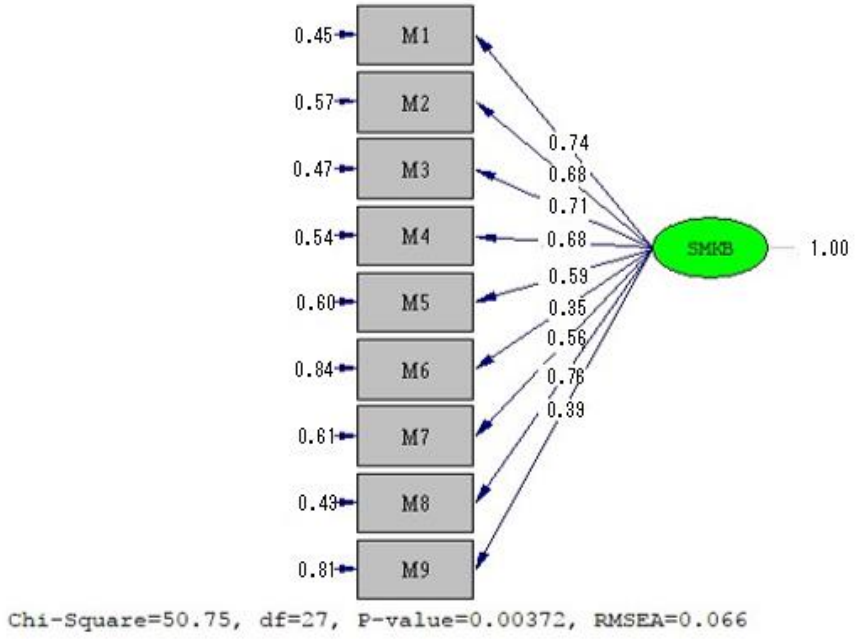
İlk gruptan elde edilen verilere uygulanan AFA sonucu Bartlet Küresellik Testi değerinin χ^2=450.74 sd=36 ($p=.000$) ve KMO örneklem uygunluk katsayısının .84 olduğu tespit edilmiştir. Mulaik'e (2010) göre bu sayının .60'tan büyük olması gerekmektedir. Hutcheson ve Sofroniou'ya (1999) göre ise KMO değerlerinin 0.80 ile 0.90 arasında olması ve Bartlet Küresellik Testi değerinin anlamlı olması örneklemin uygulama için elverişli olduğunu göstermektedir. Ayrıca ölçeğin orijinal yapısıyla uygun olarak dokuz maddenin tek faktörlü yapısı, ölçeğin toplam varyansının %48.11'ini açıklamaktadır. AFA sonucu madde faktör yük değerleri .28 ile .81 arasında sıralanmaktadır (Bkz Tablo 1)

Tablo 1. AFA Sonucu Madde Faktör Yükleri

Madde No	Madde faktör yükleri	**Madde No**	Madde faktör yükleri
m1	.77	**m6**	.28
m2	.60	**m7**	.58
m3	.77	**m8**	.75
m4	.81	**m9**	.33
m5	.61		
Toplam varyans oranı=48.11%			

Doğrulayıcı faktör analizi (DFA)

AFA sonucu oluşan yapının doğrulanması için ikinci gruptan elde edilen verilere uygulanan DFA sonucu uyum indeksi değerleri χ^2=50.65, sd=27 (χ^2/sd=1.87), RMSEA=.066, CFI=.98, GFI=.98, IFI=.98, NFI=.96, RFI=.97 ve SRMR=.039 olarak bulunmuştur (Bkz. Şekil 1).



Şekil 1. DFA Madde Faktör Yük Değerleri

Şekil 1'de görüldüğü üzere farklı örneklem gruplarında ölçeğin yapısının test edildiği, tek boyuttan ve dokuz maddeden oluşan yapısının DFA sonucu madde faktör yük değerleri (λ) .35 ile .76 arasındadır.

Ölçüt bağıntılı geçerlik

Ölçeğin ölçüt bağıntılı geçerlik çalışmasında YİBT-KF ve EATBÖ-KF ile ilişki düzeyleri Spearman-Brown korelasyon testiyle incelenmiştir. SMKBÖ-9 ile YİBT-KF ve EATBÖ-KF arasında p<.01 önem düzeyinde sırasıyla r_s=.64, .66 pozitif ilişkiler bulunmuştur (Bkz. Tablo 2).

Tablo 2. Ölçme Araçlarının Ortalama, Standart Sapma, Cronbach Alfa ve İlgileşim Değerleri

Ölçme araçları	1. SMKBÖ	2. YİBT-KF	3. EATBÖ-KF
1. SMKBÖ-9	-	.64**	.66**
2. YİBT-KF		-	.65**
3. EATBÖ-KF			-
Ortalama	28.29	25.02	20.32
Standart sapma	19.29	10.60	9.93
Cronbach alfa	.76	.81	.79

**p<.01 SMKBÖ-9=Sosyal Medya Kullanım Bozukluğu Ölçeği, YİBT-KF=Young İnternet Bağımlılık Testi-Kısa Formu, EATBÖ-KF= Ergenler için Akıllı Telefon Bağımlılık Ölçeği-Kısa Formu

Ayırt edici geçerlik

Ölçeğin geçerliğini sınamak için yapılan bir diğer çalışmada ölçeğin ayırt ediciliği incelenmiştir. Ayırt edicilik analizlerinden yararlanılırken önceki analizde sosyal medya kullanım bozukluğunun, internet bağımlılığı ve akıllı telefon bağımlılığı ile yüksek düzeyde ilişkili olmasından dolayı SMKBÖ'nün, internet bağımlılığı ve akıllı telefon bağımlılığı bağlamında ayırt ediciliğinin incelenmesine karar verilmiştir. Bunun için internet bağımlısı olan ve olmayan; akıllı telefon bağımlısı olan ve olmayan iki ayrı veri seti oluşturulmuştur. Veri setleri oluşturulurken ölçeklerden elde edilebilecek en yüksek ve en düşük puanlardan %33'lük dilimler ölçüt alınmıştır. İnternet bağımlılığı için 28 puan ve aşağısına "bağımlı değil", 44 puan ve yukarısına "bağımlı" denilmiştir. Akıllı telefon bağımlığı için 23 puan ve altı "bağımlı değil", 37 ve yukarısı "bağımlı" olarak gruplar oluşturulmuştur. Diskriminant analizi için belirlenen bu grupların SMKBÖ' den aldıkları puanları açısından Wilks' Lambda istatistiğiyle Kanonik (Cannonical) Diskriminant Fonksiyonları değerlendirilmiştir. Elde edilen bulgular Tablo 3 ve 4'te gösterilmiştir.

Tablo 3. Young İnternet Bağımlılık Testi-Kısa Formunun Gruplarına Göre SMKBÖ Wilks' Lambda İstatistiği

Fonksiyon	Wilks' Lambda	F	Ki-kare	sd	p	Kanonik korelasyon
1	.60	66.69**	102.301	2	.000	.63

**p<.01

Tablo 4. Ergenler İçin Akıllı Telefon Bağımlılık Ölçeğini Gruplarına Göre SMKBÖ Wilks' Lambda İstatistiği

Fonksiyon	Wilks' Lambda	F	Ki-kare	sd	p	Kanonik korelasyon
1	.50	98.45**	137.265	2	.000	.70

**p<.01

Tablo 3'te verilen YİBT-KF fonksiyonun Wilks'Lambda istatistiğine ilişkin ki-kare değeri [$\chi^2_{(2)}$ = 102.301; p<.01] ve Wilks' Lambda değeri (λ) .60 istatistiksel olarak anlamlı bulunmuştur. Ayrıca kanonik küme korelasyon katsayıları incelendiğinde (.63), internet bağımlılığı sosyal medya kullanım bozukluğu için paylaşılan varyansın %40'ını açıklamaktadır. Tablo 4'de verilen EATBÖ-KF fonksiyonun Wilks' Lambda istatistiğine ilişkin ki-kare değeri [$\chi^2_{(2)}$ = 137.265; p<.01] ve Wilks' Lambda değeri (λ) .50 istatistiksel olarak anlamlı bulunmuştur. Bununla birlikte kanonik küme korelasyon katsayıları incelendiğinde (.70), akıllı telefon bağımlılığı sosyal medya kullanım bozukluğu için paylaşılan varyansın %49'unu açıklamaktadır.

Güvenirlik

Ölçeğin güvenirlik çalışmaları üçüncü gruptan elde edilen verilerle yapılmıştır. Cronbach alfa iç tutarlık güvenirlik katsayısı α=.76 olarak hesaplanmıştır. Yarıya bölme analiz bulgularında ise Spearman-Brown iki yarı arasındaki korelasyon katsayısı *r*=.47, Guttman split-half katsayısı .64, her iki yarı için Cronbach α değerleri, ilk yarı için (5 soru) α= .69 , ikinci yarı için (5 soru) α= .51 olarak bulunmuştur. Alanyazında Cronbach α değerinin minimum olarak .70 olması gerektiği (Cortina,1993; Cronbach, 1951), fakat ölçek çok spesifik veya sınırlı bir özelliği ölçüyorsa .70'ten düşük bir değer alabileceği belirtilmiştir (Kline,1979, s., 292).

Madde analizleri

SMKBÖ'nin madde analiz sonuçlarına göre düzeltilmiş madde toplam korelasyon değerleri .29 ile .73 arasında sıralanmaktadır. Alanyazında düzeltilmiş madde toplam korelasyon değerleri ile ilgili

ölçüt ilgili değerlerin .30'dan büyük olmasıdır (Kline, 2000:523). Altıncı madde hariç diğer tüm bu maddeler için ölçütün sağlandığı görülmektedir. Ayrıca her bir maddenin alt-üst %27 madde ayırt edicilik indeks değerleri incelenmiş ve t değerlerinin p<01 önem düzeyinde 3.92 ile 19.51 arasında değiştiği gözlenmiştir (Bkz Tablo 5).

Tablo 5. Madde Toplam İlgileşim ve Madde Ayırt Edicilik İndeks Değerleri

Madde No	R_{jx}	Eğer madde silinirse Cronbach alfa değerleri	Alt-üst %27 madde ayırt edicilik indeksi (t)
m1-meşguliyet	.670	.537	13.70**
m2-dayanma	.649	.564	13.97**
m3-yoksunluk	.663	.546	14.96**
m4-ısrar	.561	.542	13.17**
m5-kaçış	.476	.661	11.66**
m6-problem	.288	.766	3.92**
m7-aldatma	.474	.636	7.07**
m8-yer değiştirme	.726	.534	19.51**
m9-çatışma	.321	.762	4.39**

**p<.01

TARTIŞMA ve SONUÇLAR

Sosyal Medya Kullanım Bozukluğu Ölçeği'nin ergenler üzerinde psikometrik özelliklerini incelemenin amaçlandığı bu çalışmada ölçeğin geçerlik, güvenirlik değerleri ile madde analizleri sınanmıştır. İlk olarak dil kapsam geçerliği iki ayrı teknikle değerlendirilmiştir. Her bir maddenin Türkçe ve İngilizceleri için 9 uzmandan alınan değerlerle hesaplanan dil kapsam geçerlik indeks değerleri alanda ölçüt kabul edilen .75 (Veneziano & Hooper, 1997) ve .80'den (David, 1992) büyüktür. Dolayısıyla ölçeğin dil kapsam geçerliğinin sağlandığı söylenebilir. Yapı geçerliği için iki farklı veri setine AFA ve DFA yapılmıştır. AFA sonucunda KMO değeri 0.80-0.90 arası olduğu için bu veri grubu için örneklem sayısının faktör analizine uygun olduğu söylenebilir (Kaiser, 1974). Ayrıca Bartlett küresellik testinin manidarlık değeri .05 alfa değerinden küçük olduğu için yokluk hipotezi ret edilir. Bu durumda korelasyon matrisi birim matris olmayıp faktör analizi için uygundur (Dziuban & Shirkey, 1974). Bununla birlikte ölçeğin AFA sonucunda maddelerin faktör yük değerleri (6. madde hariç) alanyazında ölçüt kabul edilen .30 (Stevens, 2002) ya da .32'den (Tabacnick & Fidell, 2014) büyüktür. Böyle bir durumda madde 6'nın atılması gereklidir, fakat bu madde bağımlılık kriterlerinden birine ait olduğundan ve ölçek bütünlüğü bozulacağından uzman görüşlerine dayanarak maddenin atılmaması kararlaştırılmıştır (Tekeş & Hasta, 2015). Madde 6'nın faktör yükünün düşük çıkması örneklem büyüklüğünün yetersiz olmasından kaynaklanabilir (MacCallum, Widaman, Zhang & Hong, 1998; Velicer & Fava, 1998). Ölçeğin orijinalindeki yapının Türk ergenlerde doğrulanıp doğrulanmadığını belirlemek için DFA sonucu uyum indeksi değerlerinden χ2/sd, CFI, GFI, IFI, NFI, RFI, SRMR değerleri alanyazında belirtilen ölçütlere (χ^2/sd<2, CFI> .95, GFI>.95, IFI>.95, NFI>.95, RFI>.95, SRMR<.05) göre iyi özellik gösterirken; RMSEA değeri (RMSEA<.08) kabul edilebilir düzeydedir (Barrett, 2007; Erkorkmaz vd., 2013; Harrington, 2009; Kline, 2011; Schermelleh-Engel, Moosbrugger, & Müller, 2003; Vieira, 2011). Ölçüt bağıntılı geçerlik çalışmasında SMKBÖ'nin YİBT-KF ve EATBÖ-KF ile p<.01 önem düzeyinde .60 ile .70 arasında ilişki katsayılarına sahip olması alanyazında yüksek düzeyde ilişki (Alper, 2006) olarak adlandırıldığından eşdeğer ölçek geçerliliğinin sağlandığı söylenebilir.

Ölçeğin ayırt edici geçerlik çalışmasında Wilks' Lambda değeri p<.01 önem düzeyinde manidar olduğundan (Ceyhan, Ceyhan, & Gürcan, 2007; Garson, 2012) SMKBÖ'nün internet bağımlılığı ve akıllı telefon bağımlılığına göre oluşturulmuş grupları ayırt edicilik etkisinin yüksek olduğu

söylenebilir. Bir başka ifade ile ölçek sosyal medya kullanım bozukluğu olanlar ile olmayanları, internet ve akıllı telefon bağımlılığı olanlar ile olmayanları ayırt edebilmektedir. Çalışmanın bu bulgusu "bir madde, cisim, nesne ya da şeye bağımlılık beraberinde birçok bağımlılık türünü ortaya çıkarabileceği" gerçeğini teknolojik yakınsama bağlamında gözler önüne sermiştir. Bir başka ifadeyle sosyal medya kullanım bozukluğu beraberinde internet bağımlılığı ve akıllı telefon bağımlılığını da tetiklemektedir (Kırcaburun, 2016; Saied, Elsabagh & El-Afandy, 2016; Sali, 2013; Şimşek & Balaban Sali, 2014; Veronica & Samuel, 2015).

Ölçeğin güvenirlik analizlerinde Cronbach alfa iç tutarlık güvenirlik katsayısının .70'ten büyük (Nunnally, 1974; Özdamar, 2002; Şencan, 2005; Tavşancıl, 2014; Traub, 1994) olması iç tutarlık güvenirliğinin psikolojik testlerde kabul edilebilir olduğunu göstermektedir. İki yarı test bulgularında Spearman-Brown, Guttman split-half ve Cronbach α katsayılarının alanyazında ölçüt kabul edilen .30 ile .70 arasında olmasının (Soğuksu ve Alıcı, 2016) kabul edilebilir fakat çok düşük değerler olduğu söylenebilir. Alanyazında kişilik ve kişilik bozukluğu gibi çok özel yapıları değerlendiren ölçeklerde Cronbach α katsayısının .43 ile .70 arasında değiştiği çalışmalar da bulunmaktadır (Cattell, 2007; Cattell & Schuerger, 2003; Cattell, Eber, & Tatsuoka, 1970; Grygier & Grygier, 1976; Soyer, Rovenpor, Kopelman, Mullins, & Watson, 2001). Patolojik internet kullanımının kişilik bozukluklarında önemli bir gösterge olduğunu aktaran çalışmalar (Muller, Beutel, Egloff, & Wolfling, 2014; Muller, Koch, Dickenhorst, Beutel, Duven, E., & Wolfling 2013) ışığında sosyal medya kullanım bozukluğunun spesifik bir kavram olduğu söylenebilir.

Madde analizlerinde düzeltilmiş madde toplam korelasyonlarının (madde 6 hariç) alanyazında ölçüt olarak bilinen .30'dan yüksek (Kline, 2000) olması her bir maddenin ölçeğin bütünü ile yeteri şekilde tutarlı olduğunu göstermektedir. Madde 6 için bu değerin .29 olması bir sınırlılık arz etse de ölçütün .30'dan küçük tutulduğu çalışmalar da mevcuttur (Acat, Tüken, & Karadağ, 2010; Tuğut & Gölbaşı, 2010). Ayrıca her bir maddenin alt-üst %27 madde ayırt edicilik indeks değerlerinin (t) $p<.01$ önem düzeyinde manidar olması (Brennan, 1972) maddelerin ayırt edicilik gücünün yüksek olduğunu göstermektedir. Diğer taraftan Van den Eijnden, Lemmens ve Valkenburg (2016) tarafından geliştirilen ölçeğin orijinal formu ile Türkçe sürümünün psikometrik özelliklerinde farklılaşmalar vardır; orijinal formun geçerlik ve güvenirlik değerlerinin daha yüksek olduğu söylenebilir. Bunun sebebinin ise kültürel farklılıklar (Korkmaz, 2007), kullanılan psikolojik terimlerin yanlış karşılığı (Erkuş, 2000; 2010), sosyal medyanın kullanım süresi ve amaçlarının farklılaşmasından kaynaklandığı söylenebilir. Bağımlılık terimi çok hassas bir psikolojik kavram olduğundan, bu kavramı içeren ifadelerin kullanımında da hassas olmak gerekir, çünkü psikolojik kavramlar kültüre duyarlıdır (Gözüm ve Aksayan, 2003; Hambleton, Merenda, & Spielberger, 2005). Dolayısıyla bir ölçeğin farklı kültürlerde geçerlik ve güvenirlik değerlerinin farklı olması olağan bir durumdur.

Sonuç olarak ölçeğin dil kapsam geçerliği, yapı geçerliği, ölçüt geçerliği, ayırt edici geçerlik ve güvenirlik değerlerine dayanarak Sosyal Medya Kullanım Bozukluğu Ölçeği'nin ergenlerin sosyal medya bağımlılığını ölçme ve değerlendirmede kullanmada geçerli ve güvenilir bir ölçme aracı olduğu söylenebilir. Bununla birlikte ölçeğin geçerlik ve güvenirlik değerlerine güç katmak amacıyla aşağıdaki öneriler yapılabilir:

- İki yarıdaki maddelerin homojen olarak aynı şeyi ölçmediği görülmektedir. Bu ifade olumsuzmuş gibi anlaşılsa da ölçeğin orijinal yapısı düşünüldüğünde mantıklı gelmektedir, çünkü ölçek maddeleri toplam bir puan verse de aslında her bir madde farklı ölçütleri değerlendirmek için hazırlanmıştır. Bundan dolayı eşdeğer yarılar güvenirliğinin Spearman-Brown formülüyle ile hesaplanması yerine Flanagan formülüyle hesaplanması iki yarı testlere ilişkin varyansları birbirine yaklaştıracak ve yarı testler arasındaki korelasyon katsayısını arttıracaktır.

- Ölçeğin test-tekrar-test güvenirlik analizleri, ilk uygulama ile ikinci uygulama arasına son sınavlar ve yaz tatili girdiğinden dolayı yapılamamıştır. Farklı çalışmalar ile bu çalışmanın test tekrar test güvenirlik katsayıları hesaplanabilir. Özellikle çalışmanın farklı illerde, farklı liselerde ergen grupları ile genişletilmesi ölçeğin genellebilirliği açısından önemli katkılar sağlayacaktır.

KAYNAKÇA

Aboujaoude, E. (2010). Problematic internet use: An overview. *World Psychiatry, 9*(2), 85-90.

Acat, M. B., Tüken, G. & Karadağ, E. (2010). Bilimsel Epistemolojik İnançlar Ölçeği: Türk kültürüne uyarlama, dil geçerliği ve faktör yapısının incelenmesi. *Türk Fen Eğitim Dergisi, 7*(4), 67-89.

Aftab, R., Çelik, İ., & Sarıçam, H. (2015). Facebook addiction and social emotional learning skills. *Ozean Journal of Social Science, 8*(2), 109-120. doi: 10.13140/RG.2.2.21774.66887

Ahn, J. (2011). The effect of social network sites on adolescents' social and academic development: Current theories and controversies. *Journal of the American Society for Information Science and Technology, 8*(62), 1435-1445. doi: 10.1002/asi.21540

Akın, A., Altundağ, Y., Turan, M. E., & Akın, U. (2014). The validity and reliability of the Turkish version of the Smart Phone Addiction Scale-Short Form for Adolescent. *Procedia-Social and Behavioral Sciences, 152,* 74-77. doi: 10.1016/j.sbspro.2014.09.157

Alper, R. (2006). *Spor bilimlerinde uygulamalı istatistik.* Ankara: Nobel Akademik.

Ang, C. S., Chan, N. N., & Lee, C. S. (2017). Shyness, loneliness avoidance, and internet addiction: What are the relationships? *The Journal of Psychology, 13,* 1-11. doi: 10.1080/00223980.2017.1399854

Andreassen, C. S. (2015). Online social network site addiction: A comprehensive review. *Current Addiction Reports, 2*(2), 175-184. doi: 10.1007/s40429-015-0056-9

Andreassen, C. S., Pallesen, S., & Griffiths, M. D. (2017). The relationship between addictive use of social media, narcissism, and self-esteem: Findings from a large national survey. *Addictive Behaviors, 64,* 287–293. doi: 10.1016/j.addbeh.2016.03.006

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: American Psychiatric Association.

Appel, H., Crusius, J., & Gerlach, A. L. (2015). Social comparison, envy, and depression on Facebook: A study looking at the effects of high comparison standards on depressed individuals. *Journal of Social and Clinical Psychology, 34*(4), 277-289. doi: 10.1521/jscp.2015.34.4.277.

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences, 42*(5), 815-24. doi:10.1016/j.paid.2006.09.018

Best, P., Manktelow, R., & Taylor, B. (2014). Online communication, social media and adolescent wellbeing: A systematic narrative review. *Children and Youth Services Review, 41,* 27-36. doi: 10.1016/j.childyouth.2014.03.001

Blackwell, D., Leaman, C., Tramposch, R., Osborne, C., & Liss, M. (2017). Extraversion, neuroticism, attachment style and fear of missing out as predictors of social media use and addiction. *Personality and Individual Differences, 116,* 69-72. doi: 10.1016/j.paid.2017.04.039

Bland, J., & Altman, D. (1997). Statistics notes: Cronbach's alpha. *BMJ, 314,* 572. doi: 10.1136/bmj.314.7080.572

Brennan, R. (1972). A generalized upper-lower item discrimination index. *Educational and Psychological Measurement, 32*(2), 289-303. doi: 10.1177/001316447203200206

Byun, S., Ruffini, C., Mills, J. E., Douglas, A. C., Niang, M., Stepchenkova., S. … Blanton, M. (2009). Internet addiction: Metasynthesis of 1996-2006 quantitative research. *CyberPsychology & Behavior, 12*(2), 203-207. doi: 10.1089/cpb.2008.0102.

Cabral, J. (2011). Is generation Y addicted to social media. *The Elon Journal of Undergraduate Research in Communications, 2*(1), 1-10.

Cattell, H. E. P. (2007). *Exploring your 16PF profile.* Oxford: Oxford Psychologist Press.

Cattell, H. E. P., & Schuerger, J. M. (2003). *Essentials of the 16PF assessment.* New York: John Wiley & Sons.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the Sixteen Personality Factor Questionnaire (16PF).* Institute for Personality and Ability Testing, Champaign, Ill.

Caumont, A. (2014). Americans increasingly view the internet, cellphones as essential. Retrieved from http://www.pewresearch.org/fact-tank/2014/02/27/americans-increasingly-view-the-internet-cellphones-as-essential/

Ceyhan, E., Ceyhan, A. A. & Gürcan, A. (2007). Problemli İnternet Kullanımı Ölçeği'nin geçerlik çalışmaları. *Kuram ve Uygulamada Eğitim Bilimleri, 7*(1), 387-416.

Chaffey, D. (2017). Global social media research summary 2017. [Online] Retrieved from http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/ [Accessed 1 May 2017].

Charter, R. A. (2001). It's time to bury the Spearman-Brown "prophecy" formula for some common applications. *Educational and Psychological Measurement, 61*(4), 690-696.

Chiou, W.-B., Lee, C.-C., & Liao, D.-C. (2015). Facebook effects on social distress: Priming with online social networking thoughts can alter the perceived distress due to social exclusion. *Computers in Human Behavior, 49,* 230-236. doi:10.1016/j.chb.2015.02.064

Collier, R. (2009). Internet addiction: New-age diagnosis or symptom of age-old problem? *CMAJ: Canadian Medical Association Journal, 181*(9), 575-576. doi: 10.1503/cmaj.109-3052

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *The Journal of Applied Psychology, 78*(1), 98-104. doi:10.1037/0021-9010.78.1.98

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. doi:10.1007/BF02310555

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2010). *Sosyal bilimler için çok değişkenli istatistik. SPSS ve Lisrel uygulamalı.* Ankara: Pegem Akademi.

Davila, J., Hershenberg, R., Feinstein, B. A., Gorman, K., Bhatia, V., & Starr, L. R. (2012). Frequency and quality of social networking among young adults: Associations with depressive symptoms, rumination, and corumination. *Psychology of Popular Media Culture, 1*(2), 72-86. doi: 10.1037/a0027512.

Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research, 5*(4), 194-197. doi: 10.1016/S0897-1897(05)80008-4.

Dobrean, A., & Păsărelu, C. R. (2016). Impact of social media on social anxiety: A systematic review. In Federico Durbano (Ed.), *New Developments in Anxiety Disorders* (pp.129-149). doi: 10.5772/65188. Retrieved from https://www.intechopen.com/books/new-developments-in-anxiety-disorders/impact-of-social-media-on-social-anxiety-a-systematic-review

Douglas, A. C., Mills, J. E., Niang, M., Stepchenkova, S., Byun, S., Ruffini, C. … Blanton, M. (2008). Internet addiction: Meta-synthesis of qualitative research for the decade 1996-2006. *Computers in Human Behavior, 24*(6), 3027-3044. doi:10.1016/j.chb.2008.05.009

Dumitrache, S. D., Mitrofan, L., & Petrov, Z. (2012). Self-image and depressive tendencies among adolescent Facebook users. *Revista de Psihologie, 58*(4), 285-295.

Dziuban, C. D., & Shirkey, E. C. (1974). When is a correletion matrix appropriate for factor analysis? Some decision rules. *Psychological Bulletin, 81*(6), 358-361.

Ekşi, H. & Çiftçi, M. (2017). Lise öğrencilerinin problemli internet kullanım durumlarının dinî inanç ve ahlaki olgunluk düzeylerine göre yordanması. *Addicta: The Turkish Journal on Addictions, 4,* 181–206. doi:10.15805/addicta.2017.4.2.0013

Erfanmanesh, M., & Hosseini, E. (2015). [Review of the book Internet and Social Media Addiction *by* Nakaya, Andrea C.]. *Webology*, 12(2). Retrieved from http://www.webology.org/2015/v12n2/bookreview25.pdf

Erkorkmaz, Ü., Etikan, İ., Demir, O., Özdamar, K. & Sanisoğlu, S. Y. (2013). Doğrulayıcı faktör analizi ve uyum indeksleri. *Turkiye Klinikleri Journal of Medical Sciences, 33*(1), 210-223. doi: 10.5336/medsci.2011-26747

Erkuş, A. (2000). Bazı psikometrik terimlerin Türkçe karşılıklarında yaşanan sorunlar. *Türk Psikoloji Yazıları, 3*(6), 31-40.

Erkuş A. (2010). Psikometrik terimlerin Türkçe karşılıklarının anlamları ile yapılan işlemlerin uyuşmazlığı. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*,*1*(2),72-77.

Esen, E., & Siyez, D. M. (2011). An investigation of psycho-social variables in predicting internet addiction among adolescents. *Turkish Psychological Counseling & Guidance Journal, 36,* 127-138.

Frost, R. L., & Rickwood, D. J. (2017). A systematic review of the mental health outcomes associated with Facebook use. *Computers in Human Behavior, 76,* 576-600. doi: 10.1016/j.chb.2017.08.001

Garson, G. D. (2012). *Discriminant function analysis.* Asheboro, NC: Statistical Associates Publishers.

Gözüm, S. & Aksayan, S. (2003). Kültürlerarası ölçek uyarlaması için rehber II: Psikometrik özellikler ve kültürlerarası karşılaştırma. *Hemşirelikte Araştırma Geliştirme Dergisi, 1,* 3-14.

Griffiths, M. (2005). A 'components' model of addiction within a biopsychosocial framework. *Journal of Substance Use, 10*(4), 191-197. doi: 10.1080/14659890500114359

Griffiths, M. D. (2010). The role of context in online gaming excess and addiction: Some case study evidence. *International Journal of Mental Health and Addiction, 8*(1), 119-125.

Griffiths, M. D. (2013). Social Networking Addiction: Emerging themes and issues. *Journal of Addiction Research & Therapy, 4*(5), e188. doi: 10.4172/2155-6105.1000e118

Griffiths, M., Van Rooij, A. J., Kardefelt-Winther, D., Starcevic, V., Kiraly, O., Pallesen, S. … Demetrovics, Z. (2016). Working towards an international consensus on criteria for assessing Internet Gaming Disorder: A critical commentary on Petry, et al (2014). *Addiction, 111*(1), 167-175. doi:10.1111/add.13057

Griffiths, M. D., Kuss, D. J., & Demetrovics, Z. (2014). Social networking addiction: An overview of preliminary findings. In *Behavioral addictions. Criteria, evidence, and treatment* (pp. 119-141). New York: Elsevier.

Grygier, J. G., & Grygier, P. (1976). *Dynamic Personality Inventory*. Montreal: Institute of Psychological Research.

Günlü, A. & Ceyhan, A. A. (2017). Ergenlerde internet ve problemli internet kullanım davranışının incelenmesi. *Addicta: The Turkish Journal on Addictions, 4*(1), 75–117. http://dx.doi.org/10.15805/addicta.2017.4.1.0016

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

Harrington, D. (2009). *Assessing confirmatory factor analysis model fit and model revision. Confirmatory factor analysis*. New York: Oxford University Press.

Hawi, N. S., & Samaha, M. (2016). The relations among social media addiction, self-esteem, and life satisfaction in university students. *Social Science Computer Review*, 35(5), 576-586 doi: 10.1177/0894439316660340

Hur, M. H. (2006). Demographic, habitual, and socioeconomic determinants of internet addiction disorder: An empirical study of Korean teenagers. *CyberPsychology & Behavior, 9*(5), 514–525. doi: 10.1089/cpb.2006.9.514

Hutcheson, G., & Sofroniou, N. (1999). *The multivariate social scientist: Introductory statistics using generalized linear models.* Thousand Oaks, CA: Sage Publications.

Islam, A., & Hossin, M. Z. (2014). Prevalence and risk factors of problematic internet use and the associated psychological distress among graduate students of Bangladesh. *Asian Journal of Gambling Issues and Public Health, 6*(1), 11. doi:10.1186/s40405-016-0020-1

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*, 31–36.

Kempa, E. P. (2015). *Social media addiction-The paradox of visibility & vulnerability.* (Master's thesis). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:hb:diva-1030

Khan, S. F., Ullah, F., Khan, M. K., Raza, A. J. S., & Shah, H. (2017). Effect of social media addiction on compliance in the patients of District Bannu, Khyber PakhtunKhwa. *International Journal of Basic Medical Sciences and Pharmacy (IJBMSP), 6*(2), 21-25.

Kırcaburun, K. (2016). Self-esteem, daily internet use and social media addiction as predictors of depression among Turkish adolescents. *Journal of Education and Practice, 7*(24), 64-72.

Kırık, A. M., Arslan, A., Çetinkaya, A., & Gül, M. (2015). A quantitative research on the level of social media addiction among young people in Turkey. *International Journal of Science Culture and Sport (IntJSCS), 3*(3), 108-122.

Kline, R. B. (2011). *Principles and practice of Structural Equation Modeling.* New York: The Guilford Press.

Kline, P. (2000). *Handbook of psychological testing* (2nd Ed.). London: Routledge.

Kline, P. (1979). *Psychometrics and psychology.* London, United Kingdom: Academic Press.

Koçer, M. (2012). Erciyes Üniversitesi öğrencilerinin internet ve sosyal medya kullanım alışkanlıkları. *Akdeniz İletişim, 18*, 70-85.

Korkmaz, M. (2007). Psikolojik ölçmenin yeni kuralları ve Türkiye'deki durumu. *Türk Psikoloji Bülteni, 13*(40), 8-14.

Kormas, G., Critselis, E., Janikian, M., Kafetzis, D., & Tsitsika, A. (2011). Risk factors and psychosocial characteristics of potential problematic and problematic internet use among adolescents: A crosssectional study. *BMC Public Health, 11*, 595. doi: 10.1186/1471-2458-11-595

Kozeis, N. (2009). Impact of computer use on children's vision. *Hippokratia, 13*(4), 230-231.

Kuss, D. J., & Griffiths, M. D. (2011). Online social networking and addiction-A review of the psychological literature. *International Journal of Environmental Research and Public Health, 8*(9), 3528-3552. doi: 10.3390/ijerph8093528

Kutlu, M., Savcı, M., Demir, Y. & Aysan, F. (2016). Young İnternet Bağımlılığı Testi Kısa Formunun Türkçe uyarlaması: Üniversite öğrencileri ve ergenlerde geçerlilik ve güvenilirlik çalışması. *Anatolian Journal of Psychiatry, 17*(Suppl. 1), 69-76.

Kwon, M., Kim, D. J., Cho, H., & Yang, S. (2013). The Smartphone Addiction: Development and validation of a short version for Adolescents (SAS-SV). *PLOS ONE, 8*(12), e83558. doi: 10.1371/journal.pone.0083558

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563-575. doi: 10.1111/j.1744-6570.1975.tb01393.x

Lemmens, J. S., Valkenburg, P. M., & Gentile, D. A. (2015). The internet gaming disorder scale. *Psychological Assessment, 72*, 567-582. doi:10.1037/pas0000062

Lemmens, J. S., Valkenburg, P., & Peter, J. (2009). Development and validation of a game addiction scale for adolescents. *Media Psychology, 12,* 77-95. doi:10.1080/15213260802669458.

Levenson, J. C., Shensa, A., Sidani, J. E., Colditz, J. B., & Primack, B. A. (2016). The association between social media use and sleep disturbance among young adults. *Preventive medicine, 85,* 36-41. doi: 10.1016/j.ypmed.2016.01.001

Li, D., Zhang, W., Li, X., Zhou, Y., Zhao, L., & Wang, Y. (2016). Stressful life events and adolescent internet addiction: The mediating role of psychological needs satisfaction and the moderating role of coping style. *Computers in Human Behavior, 63,* 408-415. doi:10.1016/j.chb.2016.05.070.

Lin, L. yi, Sidani, J. E., Shensa, A., Radovic, A., Miller, E., Colditz, J. B., … Primack, B. A. (2016). Association between social media use and depression among U.S. young adults. *Depression and Anxiety, 33*(4), 323-331. doi:10.1002/da.22466

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research, 36*(4), 611-637. doi: 10.1207/S15327906MBR3604_06

Mahmood, S., & Farooq, U. (2014). Facebook addiction: A study of big-five factors and academic performance amongst students of IUB. *Global Journal of Management and Business Research, 14*(5), 55-71.

Milani, L., Osualdella, D., & Di Blasio, P. (2009). Quality of interpersonal relationships and problematic internet use in adolescence. *CyberPsychology & Behavior, 12,* 681-684. doi:10.1089/cpb.2009.0071.

Morahan-Martin, J,, & Schumacher, P. (2000). Incidence and correlates of pathological internet use among college students. *Computers in Human Behavior, 16*(1), 13-29.

Moreno, M. A., Jelenchick, L. A., Egan, K. G., Cox, E., Young, H., Gannon, K. E., & Becker, T. (2011). Feeling bad on Facebook: Depression disclosures by college students on a social networking site. *Depress Anxiety, 28*(6), 447-455. doi: 10.1002/da.20805.

Morin-Major, J. K., Marin, M-F., Durand, N., Wan, N., Juster, R-P., Lupien, S. J. (2016). Facebook behaviors associated with diurnal cortisol in adolescents: Is befriending stressful? *Psychoneuroendocrinology, 63,* 238-246. doi:10.1016/j.psyneuen.2015.10.005

Mulaik, S. A. (2010). *The foundations of factor analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Muller, K. W., Beutel, M. E., Egloff, B., & Wolfling, K. (2014). Investigating risk factors for internet gaming disorder: A comparison of patients with addictive gaming, pathological gamblers and healthy controls regarding the big five personality traits. *European Addiction Research, 20*(3), 129-136. doi:10.1159/000355832

Muller, K. W., Koch, A., Dickenhorst, U., Beutel, M. E., Duven, E., & Wolfling K. (2013). Addressing the question of disorder-specific risk factors of internet addiction: A comparison of personality traits in patients with addictive behaviors and comorbid internet addiction. *Biomed Research International, 2013*, 1-7. doi:10.1155/2013/546342

Nabi, R. L., Prestin, A., & So, J. (2013). Facebook friends with (health) benefits? Exploring social network site use and perceptions of social support, stress, and well-being. *Cyberpsychology, Behavior, and Social Networking, 16*(10), 721-727. doi: 10.1089/cyber.2012.0521

Nunnally, J. C. (1974). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

O'Keeffe, G. S., & Clarke-Pearson, K. (2011). The impact of social media on children, adolescents, and families. *Pediatrics, 127*(4), 800-804. doi: 10.1542/peds.2011-0054

Otrar, M. & Argın, F. S. (2014). Öğrencilerin sosyal medyaya ilişkin tutumlarının kullanım alışkanlıkları bağlamında incelenmesi. *Journal of Research in Education and Teaching, 3*(3), 1-13.

Otu, A. A. (2015). *Social media addiction among students of the University of Ghana* (Master dissertation). Retrieved from http://hdl.handle.net/123456789/21223

Owusu-Acheaw, M., & Larson, A. G. (2015). Use of social media and its impact on academic performance of tertiary institution students: A study of students of Koforidua Polytechnic, Ghana. *Journal of Education and Practice, 6*(6), 94-101. Retrieved from http://files.eric.ed.gov/fulltext/EJ1083595.pdf

Özdamar, K. (2013). *Paket programları ile istatiksel veri analizi cilt 2* (9. Baskı). Ankara: Nisan.

Özdamar, K. (2002). *Paket programlarla istatistiksel veri analizi-1*. Eskişehir: Kaan.

Pantic, I. (2014). Online social networking and mental health. *Cyberpsychology, Behavior, and Social Networking, 17*(10), 652-657. doi: 10.1089/cyber.2014.0070

Pawlikowski, M., Altstötter-Gleich, C., & Brand, M. (2013). Validation and psychometric properties of a short version of Young's Internet Addiction Test. *Computers in Human Behavior, 29*(3), 1212-1223. doi: 10.1016/j.chb.2012.10.014

Robbins, T. W., Gillan, C. M., Smith, D. G., de Wit, S., & Ersche, K. D. (2012). Neurocognitive endophenotypes of impulsivity and compulsivity: Towards dimensional psychiatry. *Trends in Cognitive Sciences, 16*(1), 81-91. doi: 10.1016/j.tics.2011.11.009.

Rosenfield, M. (2016). Computer vision syndrome (a.k.a. digital eye strain). *Optometry in Practice, 17*(1), 1-10.

Rus, H. M., & Tiemensma, J. (2017). Social media under the skin: Facebook use after acute stress impairs cortisol recovery. *Frontiers in Psychology, 8*, 1609. doi: 10.3389/fpsyg.2017.01609

Ryan, T., Chester, A., Reece, J., & Xenos, S. (2014). The uses and abuses of Facebook: A review of Facebook addiction. *Journal of Behavioral Addictions, 3*(3), 133-148. doi: 10.1556/JBA.3.2014.

Saied, S. M., Elsabagh, H. M., & El-Afandy, A. M. (2016). Internet and facebook addiction among Egyptian and Malaysian medical students: A comparative study, Tanta University, Egypt. *International Journal of Community Medicine and Public Health, 3*(5), 1288-1297. doi:10.18203/2394-6040.ijcmph20161400

Sallis, M. (2013). *An exploratory study of internet and social media addiction in Millennials.* (Honors Projects: 674). Retrieved from https://repository.tcu.edu/handle/116099117/7308

Saraji, J. H., & Fini, A. A. S. (2017). Examine the relationship between internet addiction with academic achievement and mental health of high school students in Bandar Abbas. *South Journal of Educational Psychology and Counseling, 3*(1), 97-101.

Sarıçam, H., Tarhan, D. & Soyuçok, E. (2015, September). *The examination of the relations between social anxiety, loneliness, Facebook® addiction and depression with multiple regression analysis*. Paper presented at the Third International Instructional Technologies & Teacher Education Symposium, Trabzon, Turkey. doi: 10.13140/RG.2.1.1024.8561

Sarıçam, H., Yaman, E., & Çelik, İ. (2016). The mediator effect of loneliness between percieved social competence and cyber bullying in Turkish adolescents. *International Journal of Progressive Education (IJPE), 12*(1), 99-107.

Scealy, M., Phillips, J. G., & Stevenson, R. (2002). Shyness and anxiety as predictors of patterns of internet usage. *Cyber Psychology & Behavior, 5*(6), 507-515. doi: 10.1089/109493102321018141

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.

Seabrook, E. M., Kern, M. L., & Rickard, N. S. (2016). Social networking sites, depression, and anxiety: A systematic review. *JMIR Mental Health, 3*(4), e50. doi:10.2196/mental.5842

Shaw, M., & Black, D. (2008) Internet addiction: Definition, assessment, epidemiology and clinical management. *CNS Drugs, 22*(5), 353-365.

Shaw, A. M., Timpano, K. R., Tran, T. B., & Joormann, J. (2015). Correlates of Facebook usage patterns: The relationship between passive Facebook use, social anxiety symptoms, and brooding. *Computers in Human Behavior, 48*, 575-580. doi: 10.1016/j.chb.2015.02.003

Shensa, A., Escobar-Viera, C. G., Sidani, J. E., Bowman, N. D., Marshal, M. P., & Primack, B. A. (2017). Problematic social media use and depressive symptoms among U.S. young adults: A nationally-representative study. *Social Science & Medicine, 182*, 150-157. doi: 10.1016/j.socscimed.2017.03.061

Sioni, S. R., Burleson, M. H., & Bekerian, D. A. (2017). Internet gaming disorder: Social phobia and identifying with your virtual self. *Computers in Human Behavior, 71*, 11-15. doi:10.1016/j.chb.2017.01.044

Soğuksu, Y. B. & Alıcı, D. (2016). Eşdeğer yarılar güvenirliğinin farklı homojenlik düzeylerindeki örneklem büyüklüklerinde, test uzunluğuna, yarıya bölme yöntemlerine ve güvenirlik kestirme tekniklerine göre incelenmesi. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 12*(1), 237-252.

Soyer, R. B., Rovenpor, J. L., Kopelman, R. E., Mullins, L. S., & Watson, P. J. (2001). Further assesment of the construct validity of four measures of narcissism: Replication and extension. *Journal of Psychology, 135*(3), 245-258. doi:10.1080/00223980109603695

Starcevic, V., & Aboujaoude, E. (2017). Internet addiction: Reappraisal of an increasingly inadequate concept. *CNS Spectrums, 22*(1), 7-13. doi: 10.1017/S1092852915000863

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences.* New Jersey: Lawrance Erlbaum Association, Inc.

Şahin, C., & Yağcı, M. (2017). Sosyal Medya Bağımlılığı Ölçeği-Yetişkin Formu: Geçerlilik ve güvenirlik çalışması. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi (KEFAD), 18*(1), 523-538.

Şencan, H. (2005). *Sosyal ve davranışsal ölçümlerde güvenilirlik ve geçerlilik.* Ankara. Seçkin.

Şimşek, E., & Balaban Sali, J. (2014). The role of internet addiction and social media membership on university students' psychological capital. *Contemporary Educatıonal Technology, 5*(3), 239-256.

Şişman Eren, E. (2014). Sosyal Medya Kullanım Amaçları Ölçeğinin geliştirilmesi ve bazı kişisel değişkenlere göre incelenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education), 29*(4), 230-243.

Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics* (6th ed.). USA: Pearson Education Limited.

Tandoc, E. C., Ferrucci, P., & Duffy, M. (2015). Facebook use, envy, and depression among college students: Is Facebooking depressing? *Computers in Human Behavior, 43,* 139-146. doi: 10.1016/j.chb.2014.10.053

Tavernier, R., & Willoughby, T. (2014). Sleep problems: Predictor or outcome of media use among emerging adults at university? *Journal of Sleep Research, 23*(4), 389-396. doi:10.1111/jsr.12132

Tavşancıl, E. (2014). *Tutumların ölçülmesi ve SPSS ile veri analizi* (5.Baskı). Ankara: Nobel Akademik.

Tekeş, B., & Hasta, D. (2015). Özgecilik Ölçeği: Geçerlik ve güvenirlik çalışması. *Nesne Psikoloji Dergisi (NPD), 3*(6), 55-75. doi: 10.7816/nesne-03-06-03

Thompson, B. (2004). *Exploratory and confirmatory factor analysis understanding concepts and applications.* Washington, DC: APA.

Torrente, E., Piqueras, J. A., Orgilés, M., & Espada, J. P. (2014). Association of internet addiction with social anxiety and lack of social skills in Spanish adolescents. *Terapia Psicológica, 32,* 175-184. doi:10.4067/S0718-48082014000300001

Traub, R. E. (1994). *Reliability for the social sciences.* London: Sage.

Tuğut, N., & Gölbaşı, Z. (2010). Cinsel Yaşam Kalitesi Ölçeği-Kadın Türkçe versiyonunun geçerlik ve güvenirlik çalışması. *Cumhuriyet Tıp Dergisi, 32,* 172-180.

Türkyılmaz, M. (2015). Facebook Bağımlılığı Ölçeğinin Türkçeleştirilmesi ve Facebook bağımlılığının okuma becerisine etkisi. *International Journal of Social Science, 36,* 265-280. doi: 10.9761/JASSS2942

Van Rooij, A. J., Schoenmakers, T. M., van den Eijnden, R. J. J. M., & Van de Mheen, D. (2010). Compulsive internet use: The role of online gaming and other internet applications. *The Journal of Adolescent Health, 47*(1), 51-57. doi: 10.1016/j.jadohealth.2009.12.021.

van den Eijnden, R. J. J. M., Lemmens, J. S., & Valkenburg, P. M. (2016). The Social Media Disorder Scale. *Computers in Human Behavior, 61*(2016), 478-487. doi: 10.1016/j.chb.2016.03.038

Vannucci, A., Flannery, K. M., & Ohannessian, C. M. (2017). Social media use and anxiety in emerging adults. *Journal of Affective Disorders, 207,* 163-166. doi:10.1016/j.jad.2016.08.040

Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods, 3*(2), 231-251. doi: 10.1037/1082-989X.3.2.231

Veneziano, L., & Hooper, J. (1997). A method for quantifying content validity of health-related questionnaires. *American Journal of Health Behavior, 21*(1), 67-70.

Veronica, S. A., & Samuel, A. U. (2015). Social media addiction among adolescents with special reference to Facebook addiction. *IOSR Journal Of Humanities And Social Science (IOSR-JHSS), 4,* 72-76. Retrieved from http://www.iosrjournals.org/iosr-jhss/papers/Conf.17004/Volume-4/15.%2072-76.pdf

Vieira, A. L. (2011). *Preparation of the analysis. Interactive LISREL in practice*. London: Springer

Vishwanath, A. (2015). Habitual Facebook use and its impact on getting deceived on social media. *Journal of Computer-Mediated Communication, 20*(1), 83-98. doi: 10.1111/jcc4.12100

Walker, D. A. (2006). A comparison of the Spearman-Brown and Flanagan-Rulon formulas for split half reliability under variance parameter conditions. *Journal of Modern Applied Statistical Methods, 5*(2), 443-451.

Wang, M., & Qi, W. (2017). Harsh parenting and problematic internet use in Chinese adolescents: Child emotional dysregulation as mediator and child forgiveness as moderator. *Computers in Human Behavior, 77,* 211-219. doi: 10.1016/j.chb.2017.09.005

Wegmann, E., Stodt, B., & Brand, M. (2015). Addictive use of social networking sites can be explained by the interaction of internet use expectancies, internet literacy, and psychopathological symptoms. *Journal of Behavioral Addictions, 4*(3), 155-162. doi: 10.1556/2006.4.2015.021

Yen, J. Y., Ko, C. H., Yen, C. F., Wu, H. Y., & Yang, M. J. (2007). The comorbid psychiatric symptoms of internet addiction: Attention deficit and hyperactivity disorder (ADHD), depression, social phobia, and hostility. *Journal of Adolescent Health, 41*(1), 93-98.

Young, K. S. (1998a). Internet addiction: The emergence of a new clinical disorder. *CyberPsychology & Behavior, 1*(3), 237-244.

Young, K. S. (1998b). *Caught in the net: How to recognize the signs of internet addiction and a winning strategy for recovery.* New York: John Wiley & Sons.

Yuan, K., Qin, W., Wang, G., Zeng, F., Zhao, L., Yang, X. … Tian, J. (2011). Microstructure abnormalities in adolescents with internet addiction disorder. *PLoS ONE, 6*(6), e20708. doi: 10.1371/journal.pone.0020708

EXTENDED ABSTRACT

Introduction

There is current evidence that social media disorder is a growing problem, especially among children and adolescents. However, there is no comprehensive instrument for measuring social media addiction in Turkey. The current study, therefore, aimed to examine the psychometric properties of Turkish culture and to test the reliability and validity of a short and easy to administer Social Media Disorder (SMD-9) Scale that contains a clear diagnostic cutoff point to discriminate between disordered (i.e. addicted) and high-engaging non-disordered social media users.

Method

Data was obtained from 586 (202+204+180) adolescents who volunteered to take part in this study. They were selected via convenience sampling technique. The participants were all aged between 13 and 18. There were 113 girls and 89 boys in the first study group, 108 girls and 96 boys in the second study group, 66 girls and 114 boys in the third study group. 549 of the adolescents have at least one account on Facebook, WhatsApp, Instagram, Twitter, etc. social media sites.

The Social Media Disorder Scale (SMDS-9): The original form of the SMDC-9 was developed by Van den Eijnden, Lemmens, and Valkenburg (2016). In sample 1, the unconstrained first-order structural 9-item model using Mean- and Variance-adjusted Weighted Least Square (WLSMV) estimators yielded a good fit, c^2 (27, n = 724) = 24.846, p = 0.58, CFI = 1.000, RMSEA = 0.000 (90% CI:0.000-0.026). This short SMD scale was strongly correlated with the 27-item SMD scale (r= 0.89, p< 0.001) and showed good reliability with a Cronbach's alpha of 0.81 (M = 1.22, SD = 1.87). The long (27-item) and short (9-item) versions of the SMD scale both showed large positive correlations with compulsive Internet use (r > 0.50) and medium to large correlations with self-declared social media addiction, (r> 0.48), indicating satisfactory convergent validity. Test-retest reliability of the 9-item short SMD scale was assessed among the 238 adolescents who participated in both the first and the second online survey (with an interval of 2 months between these two surveys). A moderate degree of reliability was found between the first and second SMD scales. The Pearson correlation between both scales was 0.50, p< 0.001.

Young's Internet Addiction Test-Short Form: The original of the Internet Addiction Scale was developed by Young (1998a; 1998b) and its short version was made by Pawlikowski Altstötter-Gleich, and Brand (2013). Kutlu et al. (2016) undertook the Turkish adaptation of this scale. The scale consists of 12 items, and each item was presented on a 5-point, Likert-type scale (1=never, through to 5=always).

Smart Phone Addiction Scale-Short Form for Adolescent: The original of the Smart Phone Addiction Scale was developed by Kwon, Kim, Cho, and Yang (2013) and its Turkish adaptation was made by Akın et al. (2014). The scale consists of 10 items, and each item was presented on a 5-point, Likert-type scale (1= largely untrue, 5= largely true).

With Regina Van den EIJNDEN, one of the developers for the adaptation study of the SMDC, was contacted via e-mail and permission for the adaptation to Turkish was obtained (Date 28 February 2017). The psychometric properties of the Turkish SMDC-9 were examined via language content validity, structural validity (EFA and CFA), criterion-related validity, discrimination validity, internal consistency, split half reliability, and item analysis. Generally, for GFI, CFI, NFI, RFI, and IFI indices, values greater than .90, for RMSEA and RMR, values less than .05 are taken as criterion. p<.01 is based as the level of significance. For the validity and reliability analysis of the SMDC-9, a statistical computer program package was used.

Results and Discussion

According to Lawshe technique, language Content Validity Index (CVI) values were calculated between 0.78 and +1.00 and it was found that the index is statistically significant. Moreover,

according to Davis technique, content validity index (CVI) scores were calculated between 0.89 and 1.00.

As a result of the explanatory factor analysis applied to data from the study conducted on study group 1, Kaiser- Meyer-Olkin (KMO) measure of sampling adequacy was .84 and there was a significant result on Bartlett's test of Sphericity $\chi 2=450.74$ ($df=36$, $p< .001$). One-factor structure explained 48.11% of the total variance, and the factor loadings of the items ranged between .28 and .81.

The results of confirmatory factor analysis indicated that the model had good fit $\chi^2/df=1.87$, RMSEA=.066, CFI=.98, GFI=.98, IFI=.98, NFI=.96, RFI=.97, and SRMR=.039 (Sample 2). Factor loadings ranged from .35 to .76. In the concurrent validity, the Social Media Disorder Scale-9 had significant relationships with Young's Internet Addiction Test-Short Form and Smart Phone Addiction Scale-Short Form for Adolescent (r= .64, .66, respectively).

In canonical discriminant analysis, independent variable contributing most to accurate classification of social media addiction was "smart phone addiction". It was followed by "internet addiction". Cronbach alpha internal consistency coefficient was found as $\alpha=.75$ (Sample 3).

Guttman split-half reliability coefficient was found as .64. Corrected item-total correlations ranged from .29 to .73. According to t-test results concerning the significance of the difference between the upper and lower 27% of the total mean scores, there is a significant difference in favor of the upper group. Overall results demonstrated that Social Media Disorder Scale-9 can generate reliable and valid measurements to be used for assessing social media addictions of adolescents in Turkey.

Ekler

Sosyal Medya Kullanım Bozukluğu Ölçeği-9

Son 1 yıl içerisinde

Aşağıdaki 9 tane ifadeyi ne sıklıkla yaşadığınız karşısında bulunan "**0-7" ("0=Hiçbir zaman", "1=Günde bir kereden daha az", "2=Günde 1-2 defa", "3=Günde 3-5 defa", "4=Günde 6-10 defa", "5= Günde 11-20 defa", "6= Günde 21-40 defa", "7= Günde 40 kereden fazla"** anlamına gelmektedir) arasındaki rakamları (X) işareti koyarak belirtiniz. Lütfen sadece tek rakamı işaretleyiniz ve boş bırakmayınız.

1.Sosyal medyayı tekrar kullanabileceğin zamana kadar kendini sürekli olarak başka hiçbir şey düşünemez halde buldun mu?	0	1	2	3	4	5	6	7
2.Sosyal medyada daha fazla zaman harcamak istediğin için kendini sürekli memnuniyetsiz (tatmin olmamış) hissettin mi?	0	1	2	3	4	5	6	7
3.Sosyal medyayı kullanmadığında kendini sıklıkla kötü hissettin mi?	0	1	2	3	4	5	6	7
4.Sosyal medyada daha az zaman harcamaya çalıştın ama başaramadın mı?	0	1	2	3	4	5	6	7
5.Olumsuz duygulardan kaçmak için sosyal medyayı sıklıkla kullandın mı?	0	1	2	3	4	5	6	7
6.Sosyal medya kullanımı yüzünden başkalarıyla sürekli olarak tartışma yaşadın mı?	0	1	2	3	4	5	6	7
7.Anne babana veya arkadaşlarına sosyal medyada geçirdiğin süreyle ilgili sürekli yalan söyledin mi?	0	1	2	3	4	5	6	7
8.Sosyal medya kullanmak istediğin için diğer aktiviteleri (örn. hobiler, spor) sürekli olarak ihmal ettin mi?	0	1	2	3	4	5	6	7
9.Anne baban ve kardeşlerinle sosyal medya kullanımı yüzünden ciddi çatışmalar yaşadın mı?	0	1	2	3	4	5	6	7

Not: Kaynak gösterilmesi koşuluyla ölçek maddeleri kullanılabilir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(2);136-149

Differential Item and Differential Distractor Functioning Analyses on Turkish High School Entrance Exam*

Seviye Belirleme Sınavında Değişen Madde ve Değişen Çeldirici Fonksiyonu Analizleri

Ragip TERZİ** Levent YAKAR***

Abstract
Test fairness is one of the most critical elements in creating assessments. Differential item functioning (DIF) and differential distractor functioning (DDF) analyses play complementary roles in justifying test fairness. This study aims to investigate the correct (i.e., DIF) and incorrect (i.e., DDF) response choices of students based on gender in a standardized high-stakes test administered in Turkey. Given the purpose of this study, the Math section of 2011 Turkish High School Entrance Exam was investigated. For DIF analyses, Mantel Haenszel and Logistic Regression methods were used. For DDF analyses, two odds ratio approaches under the two-parameter logistic-nested logit model and nominal response model were used. According to the findings, in 500 and 1,000 sample sizes, DIF was not detected, however, only a 2,000 sample size indicated significant DIF results. Five DIF items were observed among 20 items, where three out of those five DIF items also showed DDF.

Key Words: DIF, DDF, SBS

Öz
Testin adil olması başarı testi hazırlamadaki vazgeçilmez unsurlardan biridir. Değişen madde fonksiyonu (DMF) ve değişen çeldirici fonksiyonu (DÇF) analizleri test adilliğini değerlendirmede birbirlerini tamamlayan rollere sahiptirler. Bu çalışma Türkiye'de yapılan geniş ölçekli bir teste katılan öğrencilerin cinsiyetlerine göre doğru ve yanlış yanıtlarını ve çeldirici seçimlerini DMF ve DÇF yöntemleri ile incelemeyi amaçlamaktadır. Bu amaçla 2011 yılı Seviye Belirleme Sınavı Matematik bölümü DMF ve DÇF açısından analiz edilmiştir. DMF analizleri için, Mantel-Haenszel ve Lojistik Regresyon metotları kullanılmıştır. DÇF analizleri için, iki parametreli lojistik kümelenmiş logit ve nominal yanıt modellerinin altındaki iki olasılık oranı yaklaşımları kullanılmıştır. Sonuçlara göre 500 ve 1,000 örneklem büyüklüğünde DMF bulunmazken sadece 2,000 örneklem büyüklüğünde anlamlı DMF sonuçları bulunmuştur. 20 maddelik testin beş maddesinde DMF ve bu beş maddenin üçünde de DÇF görülmüştür.

Anahtar Kelimeler: DMF, DÇF, SBS

INTRODUCTION

The Turkish Ministry of National Education initiated a high school entrance exam (SBS) for sixth- and seventh-graders in 2008 and eighth-graders in 2009. The purpose of administering the SBS each year was to closely assess whether students of those grades have met the requirements of each academic year. However, note that the SBS administration is now carried out only at the end of 8th grades, called another name, passing to high school from elementary (TEOG) until 2017. Both exams mainly have the same purpose of high school registration, ordering students based on their test scores. The SBS was a nation-wide, high stakes test administered to more than one million

* An early draft of this paper was presented at the annual meeting of Northeastern Educational Research Association, Trumbull, CT., USA. (2016, October).
** Ph.D., Harran University, School of Education, Sanliurfa, Turkey, email: terziragip@gmail.com, ORCID ID: 0000-0003-3976-5054.
*** Ph.D., Kahramanmaras Sutcu Imam University, School of Education, Kahramanmaras, Turkey, email: l_yakar@hotmail.com, ORCID ID: 0000-0001-7856-6926.

DOI: 10.21031/epod.368081
Retrieved: 18.12.2017
Accepted: 26.03.2018

students in Turkey every year (Ministry of National Education, 2011). The results from SBS were used for summative assessment purposes.

Given the high number of students who have taken such a high-stakes test should be prepared carefully. Haladyna and Downing (2004) classified construct-irrelevant variance into 21 potential sources of systematic errors, including test development, item format, and item quality. Thus, it is crucial but difficult to thoroughly investigate each measurement component of the SBS. There is still much left to explore in order to have valid and reliable measures of their achievement in the SBS that complies with minimal construct-irrelevant variance, even after test was already examined. Investigating problematic and misused parts of previous tests can help us to create better tests in future.

Writing objective items requires a lot of effort due to many reasons. One aspect of writing objective items in a test form requires that members of different examinee groups at the same level of ability be not negatively affected by the test. Furthermore, test analysis, particularly item analysis, can help evaluate how the objective items have served to assessment purposes. In other words, item analysis is a crucial way to get feedback about the quality of items. Moreover, item analysis should be routinely carried out by psychometricians or measurement experts while especially developing an item pool.

In general, designing a test includes three steps: writing test items, assembling and administering a test, and analyzing test items. Starting from the beginning towards the end of the test development, the whole process must be carefully carried out. Test fairness is an inevitable component in designing not only achievement tests but also any type of assessments. The investigation of measurement invariance plays a key role for the fairness of the test scores across the observed groups at the same ability level (Camilli, 2006). Otherwise, the item is considered as differentially functioning if the assumption of the measurement invariance is violated (Dorans & Holland, 1993). At this point, differential item functioning (DIF) and differential distractor functioning (DDF) are important analyses to be conducted for the justification of test fairness.

DIF and DDF can give useful information about the measurement invariance of a test, which refers to the degree whether the test behaves in the same way for groups given the same ability level (Dorans & Holland, 1993; Zumbo, 2007). In this context, DIF is used to describe the situation in which one group answered an item correctly more often than the other group at the same ability level (Zumbo, 2007). In other words, DIF can be used to check the stability of item performance among equally knowledgeable groups (DeMars & Lau, 2011). In addition to DIF, this study also provides additional analyses for differential distractor functioning (DDF). The concept of DDF was extended from the notion of DIF by Green, Crone, and Folk (1989). Similar to DIF conditional on the ability level, DDF can occur if a group is disproportionately attracted to a distractor due to a biasing factor in the distractor (Suh & Bolt, 2011). In DIF analyses, all correct answers are compared against all incorrect answers; whereas, only incorrect responses are examined in DDF analyses (Green et al., 1989). Because all distractors are incorrect, the choice of a distractor does not have any impact on test scores, however, DDF analyses can be informative for different subgroups (Abedi, Leon, & Kao, 2007). Therefore, DDF should be investigated along with DIF because DDF can cause DIF in correct responses (e.g., Penfield, 2008; Suh & Bolt, 2011; Terzi & Suh, 2015).

Examining DIF and DDF based on gender is important because Marshall (1983) found a significant interaction between gender and a choice of distractor in a large majority of items administered to 6^{th} grade students. In analyzing multiple-choice items, DIF and DDF analyses are very important because they are typically used for a variety of purposes such as refining existing test items, developing new scales, and validating test score inferences (Zumbo, 2007). Item bias may exist where there is DIF, which is not relevant to item quality and thus test purposes (Zumbo, 1999). Moreover, the presence of item bias is a matter of concern for test fairness, especially for high-stakes tests (Camilli, 2006).

Purpose of the Study

It is quite imperative to provide all students with conceptually and psychometrically sound assessments regardless of their differences except for ability. As explained earlier, the SBS was a nationally administrated exam, conducted by the Turkish Ministry of National Education. More than one million each sixth-, seventh- and eighth-grade students in Turkey took this exam every year. The SBS is a high-stakes test that has lasting consequences on students' futures. Even though many studies have examined measurement issues related to DIF and DDF, the impact of DIF along with DDF specifically on the assessment of the SBS has not been explored. Results from analyses of this test may give a clue about TEOG. The purpose of this study, therefore, is to conduct the DIF and DDF analyses of the SBS in which equity and test fairness are paramount to more than one million students.

METHODOLOGY

Data Source

In our study, 27,952 students without missing responses were randomly chosen from more than one million eight-grade students who took the test in 2011. Only eighth-graders' responses were used because the new version of the exam is only available for the 8^{th}-grade students. Among 27,952 students, 55% of them are male with a mean of raw scores of 8.82 and standard deviation of 6.10 and 45% of them are female with a mean of raw scores of 9.60 and standard deviation of 6.44. Because the data include a large sample size, we sampled it based on three different sizes: 500, 1,000, and 2,000. To be able to generalize results to the whole population, 100 replications were implemented. As a cut-off point, significant 50 results out of 100 replications were considered DIF or DDF. In addition, for the impact of sample sizes, effect size results were considered in the DIF analysis.

Statistical Procedures

Two methods were used for DIF analyses. The first method was the Mantel Haenszel (MH), which is a chi-square based technique. The second method was the Logistic Regression (LR), which is a regression model with observed test scores and group membership (Swaminathan & Rogers, 1990). Nagelkerke's R^2 (Nagelkerke, 1991), which is the proportion of explained variance in a model, can be taken as a measure of effect size. Two definitions of DDF have been discussed in the literature (Suh & Talley, 2015): DDF in a "divide-by-total" framework (Thissen & Steinberg, 1986) and DDF in a "divide-by-distractors" framework (Suh & Bolt, 2011). For these frameworks, DDF was analyzed using two nonparametric odds ratio approaches under the nested logit model (ORA-NLM; Terzi & Suh, 2015) for divide-by-distractors and under the nominal response model (ORA-NRM; Penfield, 2008) for divide-by-total. The reasons we chose these nonparametric approaches are because (1) it is easy for practitioners and teachers to calculate with available user-friendly software (e.g., SPSS, Excel), which could be obtained upon request; (2) these two approaches can help us understand whether DIF can have a considerable impact on DDF by including or isolating the correct option in the analyses; and (3) these approaches require a relatively smaller sample size compared to parametric approaches.

Mantel Haenszel (MH) for DIF Analysis

The Mantel Haenszel (MH; Mantel & Haenszel, 1959) is a chi-square based technique based on a contingency table developed by Holland and Thayer (1986) for DIF detection purposes. This contingency table shown in Table 1 includes two columns and two rows; columns present the number of correct and incorrect answers, rows present the focal (F) and reference (R) group sizes.

Table 1. Contingency Table by Groups for an Item

Groups	Correct	Incorrect	Total
Focal Group	A	B	A+B
Reference Group	C	D	C+D
Total	A+C	B+D	T

$$MH = \frac{AD}{BC} \quad (1)$$

$$\Delta_{MH} = -\frac{4}{1.7}\ln MH = -2.35 \ln MH \quad (2)$$

Equation 1 shows the odds ratio of DIF for a particular item, which ranges from 0 to ∞. The expected optimum result is 1. That is, if the MH value exceeds 1, the item favors focal group, and if the value is smaller than 1, item favors reference group. For more interpretable results, Equation 1 was modified into Equation 2 by taking log of the MH statistic (Holland & Thayer, 1986). Equation 2 identifies the level of DIF. If $|\Delta_{MH}| < 1$, DIF level is A (ignorable); if $1 < |\Delta_{MH}| < 1.5$, DIF level is B (medium); and if $|\Delta_{MH}| \geq 1.5$, DIF level is C (high) (Zieky, 1993). One of the disadvantages of the MH method is that it cannot distinguish between uniform and non-uniform DIF.

Logistic Regression (LR) for DIF Analysis

The Logistic Regression (LR) method was used for DIF detection purposes based on the Logistic Regression analysis (Swaminathan & Rogers, 1990). Briefly, the LR model is established by using group membership as a dummy independent variable. If the group variable has an important role for the model, the item is flagged as DIF. While responses to a particular item become the dependent variable, intercept, the total score (*X*), group variable (*g*), and an interaction term between total scores and group membership are independent variables for the LR model.

$$\ln\frac{P}{1-P} = \beta_0 + \beta_1 x + \beta_2 g + \beta_3 xg.$$

If $\beta_3$is significant for the model, the item shows non-uniform DIF. If it is not significant, β_3 is excluded from the model. Then, if β_2 is significant for the new model, the item shows uniform DIF, otherwise the item does not show any DIF. Likelihood ratio statistic was used in the study to compute DIF statistic.

Differences between explained variance (R^2) by models can be considered as an effect size. Gierl, Khaliq, and Boughton (1999) and Zumbo and Thomas (1997) proposed two different cut-off points for effect size measures. According to Table 2, Zumbo and Thomas's criteria are more conservative than Gierl, Khaliq and Boughton's criteria. If a calculated effect size is 0.1 for Gierl, Khaliq, and Boughton, it is large, but it is negligible for Zumbo and Thomas.

Table 2. Effect Size Criterion for LR DIF Detection

Level of DIF	*Gierl, Khaliq and Boughton*	*Zumbo and Thomas*	*Meanings*
A	$R^2<0.035$	$R^2<0.13$	*Negligible effect*
B	$0.035<R^2<0.07$	$0.13<R^2<0.26$	*Moderate effect*
C	$0.07<R^2$	$0.26<R^2$	*Large effect*

2PL-NLM and 2PL-NRM

A two-parameter logistic-nested logit model (2PL-NLM; Suh & Bolt, 2010) was designed as an alternative to the nominal response model (NRM; Bock, 1972). The probability of an examinee at ability θ_j choosing the correct response under the 2PL-NLM can be modeled as a 2PL model:

$$P(u_{ij}=1|\theta_j)=\left[\frac{\exp(\beta_i+\alpha_i\theta_j)}{1+\exp(\beta_i+\alpha_i\theta_j)}\right],$$

where β_i is an intercept parameter and α_i is a slope parameter for item *i*.

Given an incorrect response, the probability of choosing a distractor *v* as the product of the probability of an incorrect response and the probability of selecting distractor *v* can be modeled as follows:

$$P(u_{ij}=1,d_{ijv}=1|\theta_j)=P(u_{ij}=0|\theta_j)P(d_{ijv}=1|u_{ij}=0,\theta_j)=$$

$$=\left[1-\frac{\exp(\beta_i+\alpha_i\theta_j)}{1+\exp(\beta_i+\alpha_i\theta_j)}\right]\left[\frac{\exp(z_{iv}(\theta_j))}{\sum_{k=1}^{m-1}\exp(z_{ik}(\theta_j))}\right], \quad (3)$$

where $z_{iv}\left(\theta_j\right)=\zeta_{iv}+\lambda_{iv}\left(\theta_j\right)$, which is multinomial logit for the propensity for each distractor; u_{ij} is the item response for item *i* by examinee *j*, where if an examinee *j* answers item *i* correctly, u_{ij} = 1, and 0 otherwise; and d_{ijv} represents an item by an examinee by a distractor array such that if an examinee *j* selects a distractor *v* (*v* = 1, 2, …, *m*–1) of item *i*, d_{ijv} = 1, and 0 otherwise (Suh & Bolt, 2010).

The NRM (Bock, 1972) has the same form as the second bracket term in Equation 3. An arbitrary linear restriction is imposed on distractor parameters as $\sum_{v=1}^{m-1}\lambda_{iv}$ and $\sum_{v=1}^{m-1}\zeta_{iv}$ in the NLM, whereas, this constraint is applied to all response categories, including the correct response in the NRM.

Odds Ratio Approach under the NRM (ORA-NRM) for DDF Analysis

Following the notion in Bock (1972) under the NRM, Penfield (2008) proposed an odds-ratio approach (ORA) based on the MH common odds ratio estimator (Mantel & Haenszel, 1959). In this approach, ability is divided into *k* ability levels, which are based on total raw scores. The conditional odds ratio for distractor *v* across *k* ability levels is shown as follows:

$$\hat{\alpha}_v=\frac{\sum_{k=1}^{K}R_{0k}F_{vk}/T_k}{\sum_{k=1}^{K}R_{vk}F_{0k}/T_k},$$

where R_{0k} is the number of examinees in the reference (R) group at the k^{th} ability level who have responded to the item correctly; R_{vk} is the number of examinees in the R group who have chosen distractor *v*; F_{0k} and F_{vk} represent the counterparts in the focal (F) group; and the summation of those numbers is represented by T_k (Mantel & Haenszel, 1959):

$$T_k = R_{0k} + R_{vk} + F_{0k} + F_{vk}.$$

The natural logarithm of $\hat{\alpha}_v$ is,

$$\hat{\lambda}_v = \ln(\hat{\alpha}_v). \quad (4)$$

Then, DDF in distractor *v* can be obtained by dividing $\hat{\lambda}_v$ by its standard error (SE). The SE of $\hat{\lambda}_v$ is as follows:

$$SE(\hat{\lambda}_v) = \sqrt{\frac{\sum_{k=1}^{K} T_k^{-2}(R_{0k}F_{vk} + \hat{\alpha}_v R_{vk}F_{0k})(R_{0k} + F_{vk} + \hat{\alpha}_v R_{vk} + \hat{\alpha}_v F_{0k})}{2\left(\sum_{k=1}^{K} \frac{R_{0k}F_{vk}}{T_k}\right)^2}}. \quad (5)$$

$z(\hat{\lambda}_v) = \hat{\lambda}_v / SE(\hat{\lambda}_v)$, which is calculated based on Equations 4 and 5, is approximately distributed as the standard normal (Hauck, 1979).

Odds Ratio Approach under the NLM (ORA-NLM) for DDF Analysis

The 2PL-NLM evaluates items with DDF independent of DIF because it separates distractor parameters from correct response parameters in multiple-choice (Terzi & Suh, 2015). Suh and Bolt (2011) made a noticeable investigation of whether distractors can be considered responsible for DIF. Briefly, in contrast to Penfield (2008)'s approach, the correct response is excluded in evaluating distractor *v* under the ORA-NLM. For more details, refer to Penfield (2008) for the ORA-NRM, and Terzi and Suh (2015) for ORA-NLM.

RESULTS

The number of DIF detection results based on gender was reported in Table 3. When *N* = 500 and 1,000, none of the items showed DIF more than 50% out of 100 replications under both MH and LR approaches. Increasing the sample size caused increases in the percentage of DIF items for almost all items as expected. The MH and LR analyses showed similar results, as reported in other studies (Doğan & Öğretmen, 2008; Gierl, Khaliq, & Boughton 1999; Mazor, Kanjee, & Clauser, 1995). The MH cannot distinguish non-uniform DIF, but the LR can, researchers used both uniform and non-uniform DIF detection, so the LR can detect more DIF than the MH (Rogers & Swaminathan, 1993; Hidalgo & LóPez-Pina, 2004). Similarly, the LR approach detected DIF in items 1, 8, and 18, which were not detected by the MH when *N*=2,000. These items indicated non-uniform DIF, so the LR was able to detect these DIF items.

When *N* = 2,000, items 1 and 8 showed DIF only under the LR method. Items 13 and 17 displayed DIF in both analyses; whereas, item 5 showed DIF only under the MH. However, DIF levels of all items had a negligible effect based on the LR, even for Gierl, Khaliq, and Boughton's liberal criteria. For the MH, item 5 was advantaged for male, which showed DIF 53 times in 100 replications, and 11 of 53 DIF detections had moderate effect sizes. Item 13 was advantaged for female, which showed DIF 74 times in 100 replications, and 13 of 74 DIF detections had moderate, two of 74 had large effect sizes. Moreover, item 17 was advantaged for male, which showed DIF 73 times in 100 replications, and 20 of 73 DIF detections had moderate effect sizes.

Table 3. Number of DIF Detection

Item	*N* = 500		*N* = 1,000		*N* = 2,000	
	LR	MH	LR	MH	LR	MH
1	22	5	37	13	64*	15
2	11	3	10	9	17	8
3	4	3	6	4	3	5
4	13	15	20	23	46	47
5	12	10	27	26	47	53*
6	13	7	17	12	19	13
7	2	3	8	5	3	3
8	13	3	31	9	62*	9
9	4	6	8	4	8	13
10	11	11	6	5	10	7
11	2	2	6	5	6	7
12	4	3	9	7	8	6
13	23	18	39	45	77*	74*
14	3	4	6	3	13	5
15	8	6	8	7	15	14
16	2	2	1	4	4	6
17	19	16	30	39	63*	73*
18	11	3	16	11	47	21
19	4	2	8	3	7	3
20	4	4	12	3	19	9

Note. * indicates percentage of observing significant DIF based on 50% cut-off criterion

Among responses to 20 items, in addition to five DIF items, item 4 showed DIF close to the cut off, 46% and 47%, using the LR and MH, respectively. Thus, six DIF items were further investigated for DDF. The other 14 DIF-free items were considered as anchor items where 15 ability levels were obtained, ranging from zero to 14. For convenience, the fourth option among the four response options was set as the correct response. The significance level was set to .05 with the Bonferroni correction based on the number of response options considered.

Results are reported as the average of significant DDF detections across 100 sampled data sets for the each of the three-sample sizes, as shown in Table 4. DDF was not observed when *N* = 500 under the two approaches. However, the distractor *c* showed 50% DDF for item 13 when *N* = 1,000 under the ORA-NRM. When *N* = 2,000, the distractor *b* in item 1 showed 54% DDF under the ORA-NLM, whereas, it was 53% under the ORA-NRM. As expected with a larger sample size, the distractor *c* showed 88% DDF for item 13, in addition to the distractor *b*, which displayed 72% DDF for item 17 under the ORA-NRM.

For reference, the items with DDF can be seen in appendices. Note that these distractors remained as is while the recoding was implemented because the correct responses were *a*, *a*, and *c* for items 1, 13, and 17, respectively, which were recoded into *d*. Then, the *d* option for these items was recoded into *a*, *a*, and *c*, respectively. The purpose of DDF investigation along with DIF was to check whether DIF in the correct response could be explained through DDF. The findings suggest that the distractors *c* and *b* in items 13 and 17, respectively, can be a considerable reason for DIF in the correct response due to the nature of the ORA-NRM. However, the distractor *b* in item 1 can be

ignored because not only it was detected with a large sample size but also DIF in item 1 had a negligible effect size.

Table 4. Number of DDF Detection

		ORA-NLM			ORA-NRM		
N	Items	a	b	c	a	b	c
500	1	5	16	5	5	17	0
	4	1	1	3	1	2	3
	5	1	3	3	4	2	7
	8	5	3	1	1	1	2
	13	1	8	6	6	2	20
	17	3	4	1	3	7	5
1,000	1	9	26	6	3	27	3
	4	2	6	3	4	6	10
	5	1	5	8	4	5	16
	8	10	2	6	2	5	10
	13	3	17	9	5	1	50*
	17	4	15	4	3	36	2
2,000	1	8	54*	18	7	53*	2
	4	3	2	2	10	1	23
	5	0	4	8	12	10	33
	8	15	2	10	1	5	10
	13	1	40	29	9	3	88*
	17	12	36	3	4	72*	16

Note. * indicates percentage of observing significant DDF based on 50% cut-off criterion.

SUMMARY and DISCUSSION

This study investigated one of the nationwide high-stakes tests (SBS) in Turkey, which is a previous version of a recent high school entrance exam (TEOG). The high school entrance exam was one of the most important high-stakes tests in Turkey taken by over one million students. For valid and fair results, these types of tests must be prepared and applied carefully. To evaluate this exam thoroughly in terms of DIF and DDF, results were obtained based on 100 replications for each condition with small sample sizes because single analyses with large sample sizes could give misleading results. Increase in sample sizes caused an increase in significant DIF rates. Various sample sizes are included as a condition similar to other DIF studies (Mazor, Clauser, & Hambleton, 1992; Scott et al, 2009; Zwick, 2012). If none of the items is flagged as DIF in small sample sizes, we cannot ensure that none of the items has DIF. Similarly, if most of the items are flagged as DIF in large sample sizes, it does not guarantee that those items have DIF. Therefore, to overcome this conflict, we used moderate sample sizes and reported effect sizes, as used in many studies.

When N = 500 and 1,000, none of the items showed DIF more than 50%. Because they were lower than the cut-off point, we accept that none of the items showed DIF, and therefore, we did not investigate effect sizes. When N = 2,000, the LR method detected more DIF items than the MH method because the MH method was not originally proposed for non-uniform DIF detection (Rogers & Swaminathan, 1993). But, all DIF detections had negligible effect sizes based on the LR, even for a liberal criterion (Gierl, Khaliq, & Boughton, 1999). The MH approach detected three DIF items, which were more than 50% over the replications. Because some of these DIF detections had moderate or large effect sizes, items 5, 13, and 17 are candidate for a possible bias investigation. While items 5 and 17 favored male, item 13 favored female. For reference, these items were demonstrated in appendices. If the approaches could detect DIF in small sample sizes, it would strengthen suspicion. Because sample size effects in DIF detection or differences between the approaches were not the main purpose of this research, we did not confront them.

While there are several DIF studies on SBS math exam with the MH approach, some of these studies also investigated item bias. In terms of DIF detection, Kelecioğlu, Karabay, and Karabay (2014) detected one math item with a moderate effect size in the 8th grade SBS exam administered in 2009. Another study used the same test detected another DIF item with a moderate effect (Arıkan, Uğurlu, & Atar, 2016). Results are not the same due to using different samples. Our dataset was used by Karakaya (2012), with N = 9,374, two items (2 and 6) had DIF with moderate effect sizes, none of them had DIF in our results. Another research, which used the same dataset with N = 121,137, detected items 5 and 7 as DIF with moderate effect sizes (Kan, Sünbül, & Ömür, 2013). Item 5 had DIF in our result, but not item 7. These results show that DIF detection can vary for different samples, and even for the same samples with different sample sizes. Therefore, by using different sample sizes and the number of replications, DIF information can be enriched.

Up until now, DDF analyses have not been investigated for the SBS exam. Therefore, this study aimed to carry out additional DDF analyses for the exam. Even though the choice of a distractor does not affect test scores, DDF analyses can give information about different subgroups (Abedi, Leon, & Kao, 2007). It is important to examine DIF and DDF together, because a significant interaction between gender and a choice of distractor was found (Marshall, 1983). It is also crucial that DIF and DDF analyses can be generally applied for different goals such as refining existing test items, developing new scales, and validating test score inferences (Zumbo, 2007).

Having said that, in this study DDF was investigated along with DIF to understand whether DDF can cause DIF in correct responses. Using 14 DIF-free items as anchor items, six items (i.e., 1, 4, 5, 8, 13, and 17) were further examined for DDF. When N = 500, the two approaches did not detect any DDF. However, when N = 1,000 under the ORA-NRM, the distractor *c* in item 13 showed DDF. Increasing the sample size to 2,000, the distractor *b* in item 1 showed DDF under both ORA-NLM and ORA-NRM. Again, when N = 2,000, DDF was detected for the distractor *c* in item 13 and *b* in item 17 under the ORA-NRM. Specifically, we could state that the distractor *c* in item 13 can be a serious cause for DIF in the correct response because it was detected twice with two sample sizes under the ORA-NRM. This distractor favors male. The distractor *b* in item 17 can also be considered a reason for DIF, which favors female. However, the distractor *b* in item 1 requires a special investigation because DIF in this item had a negligible effect size. Therefore, this distractor may or may not cause DIF because it was detected by both approaches, where the ORA-NLM isolated the distractor *b* from the correct response. In summary, based on the findings, we could conclude that those DIF items detected with DDF need to be revised and evaluated with caution for valid test inferences. Note that similar to effect size results in DIF analyses, large sample sizes showed DDF more than small sample sizes.

With high-stakes tests continuing to be a central part of the educational systems, the analyses of DIF and DDF thus remain important in obtaining measurement invariance for test items across different subgroups at the same ability level. We further suggest that DIF and DDF analyses of the remaining parts of the SBS and/or TEOG such as science and social studies should be investigated in order to obtain valid test results. There is also a limitation of this study that those items detected with DIF and DDF should be investigated in the context of gender, test bias, and mathematics. A deeper discussion of why these items exhibit DIF and DDF should be a topic of future studies from the perspective of mathematics educators.

REFERENCES

Abedi, J., Leon, S., & Kao, J. (2007). *Examining differential distractor functioning in reading assessments for students with disabilities*. Minneapolis, MN: University of Minnesota, Partnership for Accessible Reading Assessment.

Arikan, Ç. A., Uğurlu, S., & Atar, B. (2016) Mimic, Sibtest, Lojistik Regresyon ve Mantel-Haenszel yöntemleriyle gerçekleştirilen DMF ve yanlılık çalışması. *Hacettepe Eğitim Fakültesi Dergisi*, *31*: 34-52. doi:10.16986/huje.2015014226

Camilli, G. (2006). Test fairness. *Educational Measurement, 4*, 221-256.

DeMars, C. E., & Lau, A. (2011). Differential item functioning detection with latent classes: How accurately can we detect who is responding differentially? *Educational and Psychological Measurement, 71*, 597–616. doi.org/10.1177/0013164411404221

Doğan, N., & Öğretmen, T. (2010). Değişen madde fonksiyonunu belirlemede Mantel-Haenszel, ki-kare ve lojistik regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim*, *33*(148), 100-112.

Doornik, J. A. (2002). *Object-Oriented matrix programming using ox* (3rd ed.). London: Timberlake Consultants Press and Oxford: www.nuff.ox.ac.uk/Users/Doornik.

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

Gierl, M., Khaliq, S. N., & Bougthon, K. (1999, June). *Gender differential item functioning in mathematics and science: Prevalence and policy implications*. Paper presented at the symposium entitled "Improving large–scale assessment in education" at the Annual Meeting of the Canadian Society for the Study of Education, Canada.

Green, B. F., Crone, C. R., & Folk, V. G. (1989). A method for studying differential distractor functioning. *Journal of Educational Measurement, 26*, 147–160.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*, 17–27.

Hidalgo, M. D., & LÓPez-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*, 903-915.

Holland, P. W., & Thayer, D. T. (1986). *Differential item functioning and the Mantel-Haenszel procedure*. ETS Research Report No. 86-31. Princeton, NJ.

Kan, A., Sünbül, Ö., & Ömür, S. (2013). 6.-8. Sınıf seviye belirleme sınavları alt testlerinin çeşitli yöntemlere göre değişen madde fonksiyonlarının incelenmesi. *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, *9*, 207-222.

Karakaya, İ. (2012). Seviye belirleme sınavındaki fen ve teknoloji ile matematik alt testlerinin madde yanlılığı açısından incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri, 12*, 222-229.

Kelecioğlu, H., Karabay, B., & Karabay, E. (2014). Seviye belirleme sınavı'nın madde yanlılığı açısından incelenmesi. *İlköğretim Online*, *13*, 934-953.

Nagelkerke, N. J. D. (1991). A note on the general definition of the coefficient of determination. *Biometrika*, *78*, 691-692.

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*, 847-862.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719-748.

Marshall, S. P. (1983). Sex differences in mathematical errors: An analysis of distractor choices. *Journal for Research in Mathematics Educations*, *14*, 325–336.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement*, *52*, 443-451.

Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement*, *32*, 131–144.

Ministry of National Education. (2011). *2011 yılı 8. sınıflar seviye belirleme sınavı sayısal bilgiler*. Retrieved from http://www.meb.gov.tr/duyurular/duyurular2011/EGITEK/sbs2011BasinBulteni/01_2011SBS_8SayisalBilgiler.pdf

Penfield, R. D. (2008). An odds ratio approach for assessing differential distractor functioning effects under the nominal response model. *Journal of Educational Measurement*, *45*, 247–269.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*, 105-116.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., & EORTC Quality of Life Group. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of Clinical Epidemiology*, *62*, 288-295.

Suh, Y., & Bolt, D. M. (2011). A nested logit approach for investigating distractors as causes of differential item functioning. *Journal of Educational Measurement, 48*, 188–205. doi.org/10.1111/j.1745-3984.2011.00139.x

Suh, Y., & Talley, A. E. (2015). An empirical comparison of DDF detection methods for understanding the causes of DIF in multiple-choice items. *Applied Measurement in Education, 28*, 48–67. doi.org/10.1080/08957347.2014.973560

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

Terzi, R., & Suh, Y. (2015). An odds ratio approach for detecting DDF under the nested logit modeling framework. *Journal of Educational Measurement, 52*, 376–398. doi.org/10.1111/jedm.12091

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

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Lawrence Erlbaum Associates.

Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF*. Prince George, Canada: University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores.* Ottawa: National Defense Headquarters.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language assessment quarterly*, *4*, 223-233.

GENİŞ ÖZET

Giriş

Bu araştırma Türkiye'de gerçekleştirilen geniş ölçekli bir sınavın değişen madde fonksiyonu (DMF) ve değişen çeldirici fonksiyonu (DÇF) açısından incelemeyi amaçlamaktadır. Bu amaçla 2011 yılı Seviye Belirleme Sınavı (SBS) analize tabi tutulmuştur. DMF ve DÇF çalışmaları adil sınav imkanının tüm gruplara eşit olarak sunulup sunulmadığını araştırmasından dolayı çok büyük bir önem arz etmektedir.

SBS, Orta Öğretim Kurumları Öğrenci Seçme ve Yerleştirme Sınavının (OKS) yerine getirilen ilköğretimden orta öğretime geçişte öğrencileri seçme amacı taşıyan bir sınavdı. 2013-2014 eğitim öğretim yılında yerini Temel Eğitimden Ortaöğretime Geçiş (TEOG) sınavına bırakmıştır. Bu sınavların her birine yaklaşık 1 milyon öğrenci katılmaktadır. Sınava katılan öğrenci sayısı ve sınavın amacı sınavın önemini ortaya koymaktadır.

Objektif bir test geliştirmenin en önemli aşamalarından biri de kaliteli madde yazımıdır. Yazılan maddeler sınava katılan tüm alt grup bireyler için aynı işleve sahip olmalıdır. Bunun içinde soru yazımında azami gayret gösterilmelidir. Ancak yine de göz önünde bulundurulmayan bazı faktörler maddeleri alt gruplar için farklı anlam taşır hale getirebilmektedir. DMF araştırmaları ile maddelerin farklı alt gruptaki eşit yeteneğe sahip bireyler için aynı zorluğa sahip olup olmadığı sorgulanmaktadır.

SBS örneğinde olduğu gibi ülkemizde ve dünyanın pek çok yerinde kullanılan geniş katılımlı sınavlarda özellikle objektif değerlendirme olanağından faydalanmak adına çoktan seçmeli sorular kullanılmaktadır. Kolay ve adil değerlendirme çoktan seçmeli testlerde avantaj olarak ele alınırken madde hazırlamanın zorluğu dezavantaj olarak karşımıza çıkmaktadır. Madde içerik ve köklerinde bulunan kimi hatalar DMF'ye neden olabilmektedir. Çoktan seçmeli testlerde madde yazımı kaliteli ve objektif çeldirici yazımını da gerektirmektedir. Farklı gruplar için farklı anlam içeren çeldirici seçenekler, adil sınav ilkesini ihlal etmekte ve DÇF'ye neden olabilmektedir. DÇF'ler ise DMF sebep olabilmektedir.

DÇF araştırmaları DMF araştırmalarına göre sayıca daha azdır. Türkiye'de gerçekleştirilen bir sınav ile ilgili DÇF çalışması bulunmamaktadır. Araştırmanın Türkiye'de gerçekleştirilen bir sınav için ilk olması ve bu sınavın ismen sona ermiş olmasına rağmen bazı yapısal değişikliklerle devam eden geniş katılımlı önemli bir sınav olması araştırmanın önemini bir kat daha artırmaktadır.

Yöntem

Araştırmanın çalışma grubunu 2011 SBS'ye giren bir milyondan fazla öğrenciden, matematik testinin tüm sorularını yanıtlayan rasgele seçilmiş 27.952 öğrenci oluşturmaktadır. B kitapçığını alan öğrencilerin yanıtları, A kitapçığındaki sıraya göre düzenlenmiştir. DMF analizi için veriler cinsiyete göre 0 ve 1'e dönüştürülmüştür. Analizler cinsiyete göre farklılaşmayı araştırmak üzere gerçekleştirilmiştir.

DMF analizi için çok kullanılan yaklaşımlardan olan Mantel-Haenzsel (MH; Holland ve Thayer, 1986) ve Lojistik Regresyon (LR; Swaminathan ve Rogers, 1990) teknikleri seçilmiştir. DMF analizinde farklı yöntemler farklı sonuçlar sunabilmektedir. Bu nedenle eksik sonuçlar üretebilecek tek yönteme bağlı kalınmamış ve biri birini destekleyerek eksik yönlerini tamamlaması beklentisiyle iki yöntem kullanılmıştır.

DÇF analizleri de iki farklı yöntemle gerçekleştirilmiştir; iki parametreli lojistik iç-içe logit model (2PL-NLM) ve nominal yanıt modelinin (NRM) altındaki iki olasılık oranı yaklaşımı (ORAs) kullanılmıştır. DÇF'nin iki tanımı, "çeldiricilere bölünme" (Suh ve Bolt, 2011) çerçevesinde DÇF olmak üzere karşı "toplama bölünme" (Thissen ve Steinberg, 1986) çerçevesinde DÇP olmak üzere literatürde sunulmuştur (Suh ve Talley, 2015). Bu çerçevede DÇF, "toplama bölünme" için nominal yanıt modeli (ORA-NRM, Penfield, 2008) ve "çeldiricilere bölünme" için iç-içe geçmiş logit modeli (ORA-NLM; Terzi ve Suh, 2015) olmak üzere parametrik olmayan iki olasılık oranı yaklaşımları kullanılarak analiz edildi. Bu parametrik olmayan yaklaşımları seçmemizin sebepleri şunlardır: Birincisi, uygulayıcılar ve öğretmenler, mevcut kullanıcı dostu yazılım (örneğin, SPSS, Excel) ile hesaplamalarını kolayca yapabilirler; bunlar, talep üzerine elde edilebilir; İkincisi, bu iki yaklaşım, analizlerde doğru seçeneği ekleyerek veya izole ederek DMF'nin DÇF üzerinde önemli bir etkisi olup olmadığını anlamamıza yardımcı olabilir. Üçüncü olarak, bu yaklaşımlar parametrik yaklaşımlara kıyasla nispeten daha küçük bir örneklem boyutu gerektirir. Analizlerimizde kolaylık sağlaması için, doğru cevaplar D seçeneği olarak düzenlenmiştir; eğer sorunun doğru yanıtı D seçeneği ise, doğru yanıt olan D seçeneğinin yeri değiştirilmemiştir.

DMF analizlerinde etkin olan en önemli faktörlerden biri de örneklem büyüklüğüdür. Bu veri setinin tamamen kullanılması DMF gösteren madde sayısında çok büyük bir artışı getirebileceği göz önünde tutularak, 500, 1,000 ve 2,000 bireylik veri setleri üzerinden analiz yapılmasına karar verilmiştir. Tek bir örneklem yerine veri setinden 100 farklı örneklem seçerek, 100 tekrarla analiz gerçekleştirilmiştir. Böylece sonuçların daha genele hitap etmesi sağlanmıştır. Tekrarların en az %50'sinde DMF'nin anlamlı sonuç olarak görülmesi durumunda madde DMF'li olarak ele alınmıştır. DMF'li olarak görülen maddeler DÇF analizine tabi tutulmuştur. Böylece DMF'li maddelerin kaynağının DÇF olup olmadığı araştırılmıştır.

DMF analizleri R (R Core Team, 2015) yazılım uygulamasındaki difR (Magis, Beland, Tuerlinckx ve De Boeck, 2010) paketi aracığıyla gerçekleştirilmiştir. DÇF analizleri ise Ox (Doornik, 2002) yazılım programında yazılan kodlarla gerçekleştirilmiştir.

Bulgular ve Tartışma

Yapılan DMF analizleri sonucu 500 ve 1,000 bireylik örneklemlerde MH ve LR analizlerinin her ikisi de hiçbir maddede, tekrarların en az yarısında anlamlı sonuç üretmemiştir. Bu durum, bu örneklem büyüklükleri için DMF'ye rastlanılmadığı yönünde yorumlanmıştır. 2,000 bireylik örneklem büyülüğünde ise tekrarların en az yarısında, LR ve MH yöntemleri 13. ve 17. maddeleri DMF'li, LR yöntemi 1. ve 8. maddeleri, MH yöntemi ise 5. maddeyi DMF'li bulmuştur. Ancak tekrarların hiçbirinde, LR tarafından DMF'li olarak işaretlenen dört maddenin ihmal edilebilir düzeyin üstünde etki büyüklüğüne sahip olmadığı görülmüştür. MH tarafından tespit edilen DMF'lerde ise bazı tekrarlarda orta ve üst düzey DMF görülmüştür.

DÇF analizleri ise potansiyel DMF taşıyan yukarıda belirtilen beş madde, ve MH ve LR analizlerinde sınır olan 50'ye çok yakın DMF gösteren 4. madde için gerçekleştirilmiştir. Sonuçlara

göre ORA-NLM yöntemi 2,000 örneklem büyüklüğünde 1. maddenin C çeldiricisinde DÇF tespit etmiştir. ORA-NRM yöntemi ise 13. maddenin C çeldiricisini 1,000 ve 2,000 örneklem büyüklüklerinde, ve 1. ve 17. maddelerinin B çeldiricisini 2,000 bireylik örneklem büyüklüğünde DÇF'li olarak işaretlemiştir.

Yapılan analizler sonucu DMF tespit edilen maddelerin üçünde DÇF tespit edilmiştir. Geniş katılımlı önemli ve ciddi sınavda 20 maddenin 5'inin analizlerce farklı fonksiyona sahip olduğunun ortaya konması önemli bir bulgudur. Test hazırlayan ve uygulayanların DMF'ye yol açabilecek şüpheli ifadelerden kaçınmasına dikkat edilmelidir. DMF'nin kaynaklarından biri olan DÇF'nin farklı fonksiyona sahip maddelere neden olabileceği görülmektedir. Sınav uzmanlarının çeldirici hazırlamada da daha dikkatli olması gerektiği araştırma sonucunda görülmektedir.

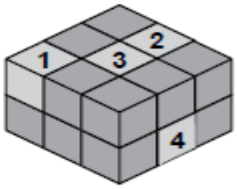
Bu çalışma diğer test alanlarında ve/veya diğer geniş ölçekli testlerde gerçekleştirilebilir. Yanlılık kaynağı araştırması için uzmanlar eşliğinde değerlendirmesi gelecek çalışmalar için önerilmektedir.

Appendices

1. $(-3)^{-2}$ sayısı aşağıdaki sayılardan hangisi ile çarpılırsa sonuç 3 olur?

A) 3^3 B) 3^{-1}
C) 3^2 D) $(-3)^{-3}$

13. Birim küplerden oluşan yandaki yapıda, numaralandırılmış küplerden hangisi çıkarıldığında yapının yüzey alanı <u>değişmez</u>?



A) 1 B) 2 C) 3 D) 4

17.

1. adım 2. adım 3. adım 4. adım

Yukarıda verilen örüntü, aynı kurala göre devam ettirildiğinde 19. adımdaki çemberin içine çizilen çokgenin kenar sayısı kaçtır?

A) 24 B) 33 C) 39 D) 42

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(2);150-160

Merkezsizleştirme Becerisini Değerlendirme: Yaşantılar Ölçeğinin Türkçe Formunun Psikometrik Özellikleri*

Measuring Decentering: Psychometric properties of the Turkish Version of Experiences Questionnaire*

Fatma Zehra ÜNLÜ KAYNAKÇI**

Öz
Bu çalışmada, bilinçli farkındalığa dayalı yaklaşımların ortak unsuru olarak görülen merkezsizleştirme becerisini ölçen Yaşantılar Ölçeğinin (Fresco ve diğ., 2007) Türkçeye uyarlamasının yapılması amaçlanmıştır. Katlımcılar, Ankara'da bir devlet üniversitesinde okumakta olan 363 üniversite öğrencisinden (251 kadın, 112 erkek) oluşmaktadır. Yaşantılar Ölçeğinin tek faktörlü yapısını test etmek için Doğrulayıcı Faktör Analizi (DFA) yapılmıştır. DFA bulguları, ölçeğin tek faktörlü yapısının doğrulandığını göstermiştir (Satorra-Bentler $\chi 2/df$ = 3.05 (p < .001), *RMSEA* = .07, *NNFI* = .93, *CFI* = .94, *SRMR* = .06, *NFI* = .92). Yaşantılar Ölçeğinin, merkezsizleştirme boyutunun Cronbach-α katsayısı .80 olarak bulunmuştur. Ayrıca ölçeğin ölçüt bağıntılı geçerliğini test etmek için Ruminasyon Ölçeği Kısa Formu (Treynor, Gonzealez ve Nolen-Hoeksama, 2003) ve Yaşantılar Ölçeği – Merkezsizleştirme arasındaki ilişki 620 üniversite öğrencisinin katılımıyla incelenmiştir. Ruminasyon ve merkezsizleştirme arasında negatif yönde anlamlı ilişki bulunmuştur. Sonuç olarak, üniversite öğrencilerinin katılımı ile gerçekleştirilen bu çalışmada Yaşantılar Ölçeği - Merkezsizleştirme Türkçe formunun geçerli ve güvenilir bir ölçme aracı olduğuna dair bulgular elde edilmiştir. Elde edilen bulgular, alan yazın ışığında tartışılmıştır.

Anahtar Kelimeler: Merkezsizleştirme, yaşantılar ölçeği, geçerlik, güvenirlik

Abstract
The aim of the current study was to adapt Experiences Questionnaire (EQ) (Fresco et.al, 2007) which measures decentering, a common element of all mindfulness-based therapies, into Turkish. The participants of the study were 363 undergraduate students (251 females, 112 males) enrolled at a state university in Ankara. Confirmatory Factor Analysis (CFA) was carried out to test the one factor structure of the Experiences Questionnaire. CFA results confirmed one factor structure of the questionnaire (Satorra-Bentler χ^2/df = 3.05 (p < .001), RMSEA = .07, NNFI = .93, *CFI* = .94, *SRMR* = .06 and *NFI* = .92). Cronbach alpha value for decentering factor of Experiences Questionnaire was found to be .80. Furthermore, in order to test the criterion-related validity of the EQ-Decentering, the relationship between EQ- Decentering and the Short form of the Ruminative Response Scale (Treynor, Gonzealez, & Nolen-Hoeksama, 2003) was investigated with the participation of 620 university students. Decentering was found to be significantly and negatively related to rumination. As a result, the current study conducted with the university students provided evidence for the reliability and validity of the Turkish form of EQ-Decentering. Findings were discussed in the light of the relevant literature.

Keywords: Decentering, experiences questionnaire, reliability, validity

GİRİŞ

Merkezsizleştirme, bireyin duygu ve düşünceleri ile olan ilişkisinde onlarla özdeşleşmesi ve onları tek bir gerçeklik olarak değerlendirmesi yerine duygu ve düşüncelerini zihninden geçen geçici olaylar

* Bu çalışmanın bir kısmı yazarın, Prof. Dr. Oya Yerin Güneri danışmanlığında 19 Haziran 2017 tarihinde tamamladığı "A model for psychological distress among university students: mindfulness, decentering, reframing, and indirect effect of emotion regulation difficulties" adlı doktora tezinden üretilmiştir. Ayrıca çalışma, 9-10 Nisan 2015 tarihlerinde Eskişehir'de düzenlenen VIII. Psikolojik Danışma ve Rehberlik Sempozyumun'da sözlü bildiri olarak sunulmuştur.

** Arş. Gör. Dr., Bülent Ecevit Üniversitesi, Eğitim Fakültesi, Ereğli-Zonguldak-Türkiye, e-posta: fzunlu@gmail.com, ORCID ID: orcid.org/0000-0002-9491-2750

DOI: 10.21031/epod.367560
Geliş Tarihi: 18.12.2017
Kabul Tarihi: 04.04.2018

olarak gözlemlemesi becerisi olarak tanımlanmaktadır (Fresco ve diğ., 2007; Safran ve Segal, 1990). Mennin ve Fresco (2014), merkezsizleştirme kavramına açıklık getirirken duygu ve düşünceleri büyük bir göle benzetip merkezsizleştirmenin gölün içine atlamak yerine onun yanında oturup onu izlemek olarak açıklamıştır. Duygu ve düşünceleri uzaktan izleyebilmenin, onları daha objektif bir şekilde algılamaya (Fresco ve diğ., 2007; Safran ve Segal, 1990) ve olumsuz duygu ve düşüncelerin azalmasına yardımcı olduğu ifade edilmektedir (Segal, Williams ve Teasdale, 2002). Ayrıca merkezsizleştirme, Nolen-Hoeksema (1998) tarafından bireyin pasif ve tekrarlayan bir şekilde sıkıntılarının nedenleri ve sonuçları üzerine düşünmesi olarak tanımlanan ruminasyonun bir nevi zıttı olarak ele alınmış ve merkezsizleştirme becerisinin ruminasyona bir çeşit çözüm olduğu ifade edilmiştir (Segal, Williams ve Teasdale, 2012).

Merkezsizleştirme teriminin kökeni bilişsel yaklaşımdan gelmektedir (Beck, Rush, Shaw ve Emery, 1979). Ancak bilişsel yaklaşımlarda "üçüncü kuşak" veya "üçüncü dalga" olarak ifade edilen bilinçli farkındalık ve kabul temelli yaklaşımlarla merkezsizleştirme becerisinin psikolojik sağlık üzerindeki olumlu etkisi daha çok ön plana çıkmıştır (Fresco ve diğ., 2007). Özellikle, bilinçli farkındalık temelli yaklaşımlardan biri olan ve depresyona tekrar girmeyi önlemek için geliştirilmiş "Bilinçli Farkındalık Temelli Bilişsel Yaklaşımın" temelini oluşturmaktadır (Baer, Walsh ve Lykins, 2009; Segal ve diğ., 2012). Bununla birlikte tüm farkındalık temelli yaklaşımların ortak unsuru olarak görülmektedir (Baer ve Huss, 2008). Psikolojik danışma sürecinde ise merkezsizleştirme ile ilgili farklı bakış açıları bulunmaktadır. Bilişsel yaklaşım bağlamında, bu beceri işlevsel olmayan düşünceleri değiştirmek için bir yol olarak değerlendirilmektedir (Segal ve diğ., 2012). Başka bir deyişle merkezsizleştirme, işlevsel olmayan düşünceyi değiştirebilmenin ilk basamağıdır çünkü bu yaklaşımda danışanlar düşüncelerini merkezsîzleştirme bakış açısı ile değerlendirmeyi öğrenerek düşüncelerinin doğruluğu ve faydalılığı hakkında tartışabilirler (Hayes, 2004; Herbert ve Forman, 2011). Öte yandan, bilinçli farkındalık temelli yaklaşımlarda düşüncenin içeriğini değiştirmek gerekli görülmez veya danışmada çok sınırlı bir işlevinin olduğu düşünülür (Hayes, 2004; Sauer ve Baer, 2010). Bilinçli farkındalık temelli yaklaşımları geleneksel bilişsel davranışçı yaklaşımla bütünleştirip duygu bilimi ile entegre eden "Duygu Düzenleme Terapisi" ise merkezsizleştirmeyi duygu düzenleme becerisi olarak ele alır ve hem merkezsizleştime hem de düşünce içeriğinin değiştirilmesi bu yaklaşımın temel bileşenlerindendir (Mennin ve Fresco, 2014). Merkezsizleştirme kavramının psikolojik danışmada ele alınışı ile ilgili kuramlar arasında farklı bakış açıları olsa bile yaklaşımlar bu becerinin psikolojik danışmada öneminin altını çizmektedirler.

Bilinçli farkındalık temelli yaklaşımlarla birlikte araştırmalar hem bilinçli farkındalık hem de merkezsizleştirme'nin ruh sağlığı üzerindeki etkisine odaklanmakta ve bu becerilerin ruh sağlığını olumlu yönde etkilediğine dair sonuçlar elde edilmektedir. Çalışmalar özelikle merkezsizleştirmenin depresyon üzerindeki etkisine vurgu yapmaktadır (örn; Fresco, Segal, Buis ve Kennedy, 2007; McCracken, Gutierrez-Martinez ve Smyth, 2012). Ancak ruh sağlığı ile ilgili diğer değişkenler ile ilgili çalışmalar da merkezsizleştirme becerisinin ruh sağlığındaki önemini ortaya çıkarmaktadır. Örneğin; merkezsizleştirme becerisi arttıkça sosyal kaygı (Hayes-Skelton ve Graham, 2012), duygu düzenleme güçlükleri (Lafferty, 2013), alkol ile ilgili problemler (Pearson, Brown, Bravo ve Witkiewitz, 2014) ve psikolojik sıkıntının (Morgan, 2015) azaldığı belirtilmektedir.

Merkezsizleştirme becerisi bilinçli farkındalık bağlamında ele alınarak; ruh sağlığı ve bilinçli farkındalık arasındaki ilişkide merkezsizleştirmenin rolü incelenmiştir (Corcoran, Farb, Anderson ve Segal, 2010; Hayes-Skelton ve Graham, 2012). Shapiro, Carlson, Astin ve Freedman (2006) bilinçli farkındalığın psikolojik sağlık üzerindeki etkisini merkezsizleştirme aracı rolü ile açıklayan bir farkındalık modeli sunmuştur. Buna göre, bilinçli farkındalık, merkezsizleştirme becerisinde önemli bir değişime neden olmakta ve bu değişim duygu-düşünce ve davranış esnekliği, öz düzenleme gibi mekanizmaları kolaylaştırmaktadır. Corcoran ve diğerleri (2010) da bilinçli farkındalık ve merkezsizleştirme ile ilgili bir model sunmuştur. Bu modele göre, bilinçli farkındalık duygu düzenlemeyi ve dikkat kapasitesini merkezsizleştirme aracı rolü ile artırmaktadır. Hölzel ve diğerleri (2011) tarafından geliştirilen diğer bir modele göre ise bilinçli farkındalık ile iyi oluş arasındaki ilişki; dikkat düzenleme, duygu düzenleme, beden farkındalığı ve kendilik üzerindeki algı değişimi aracı

rolleri ile açıklanmaktadır. Kendilik üzerindeki algı değişimi ile merkezsizleştirme bakış açısı üzerinde durulmuştur.

Alan yazında, merkezsizleştirme becerisinin psikolojik sağlıkla doğrudan ilişkili olduğu ve aynı zamanda bilinçli farkındalık ve psikolojik sağlık arasındaki ilişkide kritik rolü vurgulanmaktadır. Ancak ulusal alan yazında bilinçli farkındalık ile ilgili çalışmalar (örn., Albayrak, 2015; Özyeşil, 2011; Ülev, 2014) hızla artmasına rağmen bilinçli farkındalığın oluşturduğu olumlu etkide anahtar role sahip olduğu öne sürülen merkezsizleştirme ile ilgili yayınlanmış bir çalışmaya rastlanmamıştır. Ülkemizde merkezsizleştirme becerisini ölçmeye yönelik bir ölçme aracının alana kazandırılmasının bu konu ile ilgili araştırmalar yapılmasına olumlu katkılar sağlayacağı düşünülmektedir. Yaşantılar Ölçeği (Fresco ve diğ., 2007) ise merkezsizleştirme becerisini hem üniversite öğrencilerinde hem de klinik örneklemde ölçmek için kullanılan ve Almanya (Gecht ve diğ., 2014), Japonya (Kurihara, Hasegawa ve Nedate, 2011), İspanya (Soler ve diğ., 2014), Portekiz (Gregório, Pinto-Gouveia, Duarte ve Simoes, 2015), ve İran (Taherifar, Ferdowsi, Mootabi, Tehrani ve Fata, 2017) gibi farklı kültürlerde uyarlama çalışmaları yapılan bir ölçektir. Yapılan çalışmalar, ölçeğin merkezsizleştirme becerisini değerlendirmek için geçerli ve güvenilir olduğuna işaret etmektedir. Bu nedenle, Yaşantılar Ölçeği'nin Türkçeye uyarlama çalışmasının, bilinçli farkındalık temelli yaklaşımlarda ruh sağlığındaki kritik rolünün daha çok vurgulandığı merkezsizleştirme becerisini ölçmek için alana yeni bir ölçme aracı kazandırılması açısından önemli olduğu düşünülmektedir. Bu bağlamda, çalışmada Fresco ve diğerleri (2007) tarafından geliştirilen merkezsizleştirme becerisini ölçmek için kullanılan Yaşantılar Ölçeğinin Türkçeye uyarlanması amaçlanmaktadır.

YÖNTEM

Bu çalışma ölçek uyarlama çalışmasına uygun olarak tarama yöntemi ile gerçekleştirilmiştir. Tarama yöntemi belli bir popülasyonda belli bir değişkenin doğası veya sıklığı hakkında bilgi veren araştırma yöntemi olarak tanımlanmaktadır (Heppner, Wampold ve Kivlighan, 2008).

Örneklem

Çalışmaya, Ankara'da bir devlet üniversitesinde okumakta olan 394 üniversite öğrencisi katılmıştır. Çalışmanın örnekleminin belirlenmesinde uygun örnekleme (convenient sampling) yöntemi kullanılmıştır. Veri tarama işlemi sonucunda kayıp veri ve uç değerler nedeni ile örneklem sayısı 363 (251 kadın, 112 erkek) olarak belirlenmiştir. Katılımcıların yaş aralığı 18 ile 31 arasında değişmektedir ve yaş ortalaması 21.90'dır (*ss* = 2.27). Katılımcıların akademik ortalamaları 0.50 ve 4.00 arasında değişiklik göstermiştir. Fakülte dağılımına bakıldığında 210'u (%57.8) Eğitim Fakültesi, 83'ü (%22.9) Fen Edebiyat Fakültesi, 41'i (%11.3) Mühendislik Fakültesi, 22'si (%6.1) İktisadi ve İdari Bilimler Fakültesi ve 7'si (%1.9) Mimarlık Fakültesi öğrencileridir. Sınıf düzeyine göre katılımcıların dağılımı ise şu şekildedir: 53'ü (%14.6) birinci sınıf, 88'i (%24.2) ikinci sınıf, 97'si (%26.7) üçüncü sınıf, 124'ü (% 34.2) dördüncü sınıf öğrencisinden oluşmuş; 1 öğrenci (%0.3) sınıfını belirtmemiştir. Çalışmanın ikinci kısmında, ölçeğin ölçüt geçerliğini test etmek için 429 kadın (%69.2) ve 191 erkek (%30.8) olmak üzere 620 lisans öğrencisinden veri toplanmıştır. Katılımcıların yaş aralığı 18 ve 30 arasında değişmektedir ve yaş ortalaması 21.88'dir (*ss* = 1.68). Fakülte dağılımına bakıldığında, 367'sinin (%59.2) Eğitim Fakültesi, 131'inin (%21.1) Mühendislik Fakültesi, 72'sinin (%11.6) Fen Edebiyat Fakültesi, 40'ının (% 6.5) İktisadi ve İdari Bilimler Fakültesi ve 10'unun (%1.6) Mimarlık Fakültesi öğrencileri olduğu görülmektedir. Sınıf düzeyinde ise 144'ü (%23.2) birinci sınıf, 93'ü (%15.0) ikinci sınıf, 177'si (%28.5) üçüncü sınıf, 199'u (% 32.1) dördüncü sınıf öğrencisinden oluşmuş; 7 öğrenci (%1.1) sınıfını belirtmemiştir.

Veri Toplama Araçları

Yaşantılar Ölçeği (Fresco ve diğ., 2007)

Yaşantılar ölçeği merkezsizleştirme ve ruminasyon olmak üzere iki alt ölçekten oluşmaktadır. Ölçek ilk oluşturulduğunda 14 madde merkezsizleştirme (örn; *"Düşünce ve duygularımı kendimden ayrı tutabilirim"*) ve 6 madde ruminasyon (örn; "*Başkalarının bana söyledikleri hakkında tekrar tekrar düşünürüm.*") olmak üzere 20 maddeden oluşmaktadır. Merkezsizleştirme boyutundaki maddeler kişinin olumsuz deneyimlerine alışkanlıkla tepki göstermeme**,** kişinin kendini düşünceleri ile aynı görmeme ve öz şefkati ölçmeyi amaçlayan maddelerden oluşmaktadır. Ruminasyon ile ilgili maddeler ise tepki yanlılığını kontrol etmek amacıyla kullanılmıştır. Fresco ve diğerleri (2007) tarafından yapılan analizler sonrasında iki faktörlü yapı doğrulanmamıştır ve sonrasında analiz sadece merkezsizleştirme maddelerini içeren tek faktörlü yapı ile tekrar edilmiştir. Bu sonuçlara göre ise yaşantılar ölçeğinin 2, 5 ve 8. maddeleri teorik ve istatistiksel değerlendirmeler sonucunda ölçekten çıkarılarak 11 maddelik (YÖ3, YÖ6, YÖ9, YÖ10, YÖ12, YÖ14, YÖ15, YÖ16, YÖ17, YÖ18, YÖ20) merkezsizleştirme alt boyutu geçerli ve güvenilir bulunmuştur. Sonuç olarak, Yaşantılar Ölçeği - Merkezsizleştirme 11 maddeden oluşan 5'li likert tipinde bir ölçektir (1 = Hiçbir zaman; 5 = Her zaman). Ölçekten alınabilecek en düşük puan 11 en yüksek puan 55'tir. Ölçekteki tüm maddeler olumlu yöndedir. Ölçekten alınan yüksek puanlar merkezsizleştirme becerisinin yüksek olduğunu göstermektedir. İç tutarlılık katsayısı Yaşantılar Ölçeği - Merkezsizleştirme için üniversite öğrencileri örnekleminde .83 ve klinik örneklemde .90 olarak bulunmuştur (Fresco ve diğ., 2007). Çalışma kapsamında araştırmacı tarafından ölçeğin Türkçeye uyarlama çalışması yapılmıştır. Yazarın talebi ile ölçeğin 20 maddelik orijinal hali Türkçeye çevrilmiştir.

Ruminasyon Ölçeği Kısa Formu (Treynor ve diğ., 2003)

Ölçek ruminatif düşünceleri ölçmek için geliştirilmiştir. Saplantılı düşünme ve derin düşünme olmak üzere iki alt boyuttan oluşmaktadır. Saplantılı düşünce 5 madde ve derin düşünme 5 madde olmak üzere toplam 10 maddelik ve 4' lü likert tipi bir ölçektir (1 = Hiçbir zaman; 5 = Her zaman). Ölçek Türkçeye Erdur-Baker ve Bugay (2012) tarafından uyarlanmıştır. Orijinal formda Cronbach Alfa katsayısı derin düşünme .72 ve saplantılı düşünce .77 bulunmuştur.(Treynor ve diğ., 2003). Ölçeğin Türkçe formunda ise iç tutarlılık katsayısı tüm ölçek için .85, derin düşünme için .72 ve saplantılı düşünce için ise .77 olarak hesaplanmıştır (Baker ve Bugay, 2010).

İşlem

Ölçeğin uyarlama çalışması yapılmadan önce ölçeği geliştiren yazarlardan biri olan Fresco ile iletişime geçilerek ölçeğin Türkçeye uyarlama çalışması için izin alınmıştır. Sonrasında, ODTÜ İnsan Araştırmaları Etik Kurulundan izin alınarak ölçek uyarlama çalışması yapılmıştır. Ölçek uyarlama çalışması, Hambleton (2001) Uluslararası Test Komisyonu (International Test Commission; ITC) çeviri ve uyarlama rehberinde belirtilen beş aşama dikkate alınarak gerçekleştirilmiştir. Takip edilen aşamalar şunlardır: orijinal dilden hedef dile çeviri, uzmanlar tarafından ölçek maddeleri ile ilgili görüş alınması, bilişsel sorgulama yapılması ve hedeflenen örneklemde ölçeğin psikometrik özelliklerinin test edilmesidir. Bu doğrultuda ölçek İngilizce ve Türkçeyi iyi derecede bilen Rehberlik ve Psikolojik Danışma Bölümünde doktora yapan beş kişi tarafından bağımsız olarak Türkçeye çevrilmiştir. Çeviriler Rehberlik ve Psikolojik Danışma Bölümünde bir öğretim üyesi ve yine aynı bölümde doktora yapan bir kişi tarafından değerlendirilerek en doğru çevrilen maddeler seçilmiştir. Belirlenen ölçek maddeleri Eğitim Bilimlerinde öğretim üyesi olan iki kişiden görüş alınarak düzenlenmiştir. Ölçek izni kapsamında Fresco tarafından ölçeğin Türkçe formunun İngilizce' ye geri çevirisi istendiği için Rehberlik ve Psikolojik Danışma Bölümünde doktora yapan bir İngilizce öğretmeni tarafından ölçeğin İngilizceye geri çevirisi yapılmış ve yazardan ölçeğin uygunluğuna yönelik onay alınmıştır. Tüm bu aşamalar sonrasında bilişsel sorgulama kapsamında, eğitim fakültesinde okuyan 21 üniversite öğrencisine ölçek uygulanarak geri bildirim alınmıştır. Bu geri bildirimler şunlardır: Üstünlük belirteçleri içeren maddeler anlaşılır bulunmamıştır ve madde 17 "Düşüncelerimden ibaret olmadığımı görebiliyorum" maddesi çok soyut olarak değerlendirilmiştir. Geri bildirimler doğrultusunda üstünlük belirteçleri içeren maddeler ölçeğin tıpkı Almanca çevirisinde (Gecht ve diğ., 2014) olduğu gibi tekrar düzenlenmiştir. Örneğin; madde 3 "Kendimi olduğum gibi

kabul etmede daha iyiyim", "Kendimi olduğum gibi kabul edebilirim" olarak değiştirilmiştir. Madde 17 "Düşüncelerimden ibaret olmadığımı görebiliyorum" ile ilgili olarak araştırmacı tarafından öğrencilere bu maddeden ne anladıkları sorulmuştur. Açıklamalar ölçeğin kapsamıyla tutarlı olduğu için bu maddede herhangi bir değişiklik yapılmamıştır. Tüm bu aşamalardan sonra ölçeğin Türkçe formu oluşturulmuştur.

Verilerin Analizi

Ölçeğin geçerliğini test etmek için LISREL 8.80 programı kullanılarak Doğrulayıcı Faktör Analizi (DFA) yapılmış ve güvenirliğini belirlemek için Cronbach alfa iç tutarlık katsayısı hesaplanmıştır.

Gerekli analizlerin yapılmasından önce minimum ve maksimum değerlere bakılarak yanlış veri girişi olup olmadığı kontrol edilmiş ve kayıp değerler, uç değerler, doğrusallık ve normallik sayıltıları test edilmiştir. Kayıp değerler incelenmiş ve sadece 15 katılımcıda kayıp değer bulunduğu için liste bazında veri silme yöntemi kullanılarak bu kayıp değerler analizden çıkarılmıştır. Çalışmadaki 16 uç değer de analizden çıkarılmıştır. Dolayısıyla sonraki analizler 251'i kadın ve 112'si erkek olmak üzere 363 üniversite öğrencisi ile gerçekleştirilmiştir. Örneklem büyüklüğünün uygunluğu LISREL 8.80 kullanılarak Hoelter's kritik *N* (Critical N-CN) değeri ile 145.06 olarak hesaplanmıştır. Bu değer çalışmada kullanılan 363 birimlik örneklem büyüklüğünün doğrulayıcı faktör analizi için yeterli olduğunu göstermektedir. Normal dağılım varsayımına bakıldığında basıklık ve çarpıklık değerlerinin +3 ile -3 arasında olduğu görülmüştür (Tabachnick ve Fidell, 2007). Ancak çok değişkenli normal dağılım sayıltısı sağlanamadığı için Maksimum olasılık yerine Güçlü maksimum olasılık (Robust maximum likelihood) metodu kullanılmıştır.

Çalışmanın Doğrulayıcı Faktör Analizi sonuçlarının değerlendirilmesinde bazı uyum iyiliği indeksleri ve kabul edilebilir düzeyde uyumu işaret eden değerler dikkate alınmıştır. Bunlar; Ki-kare'nin serbestlik derecesine oranının (χ^2 / *sd*) 5'den küçük olması (Wheaton ve diğ., 1977), NFI, NNFI ve CFI değerlerinin .95'e yakın olması, SRMR değerinin .08 veya .08'den küçük olması (Hu ve Bentler, 1999), RMSEA değerinin de .08'den küçük olması (Browne ve Cudeck, 1993)'dır.

BULGULAR

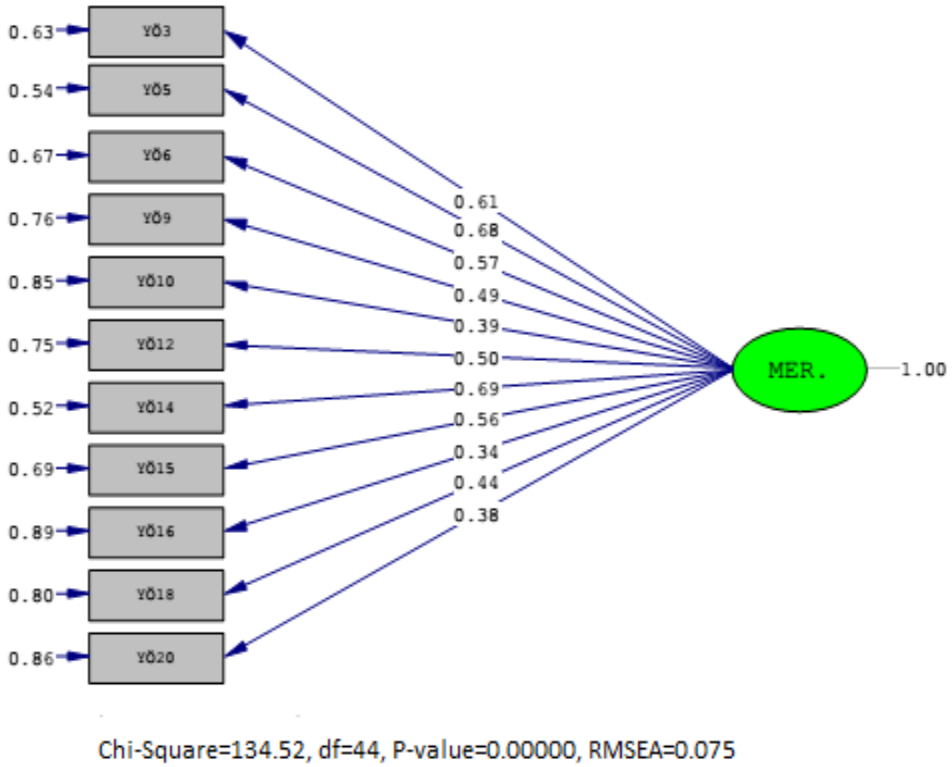
DFA sonuçlarına göre, Yaşantılar Ölçeği'nin iki faktörlü yapısı madde 2'nin anlamsız *t* değerinden ve madde 4, 11, 17 ve 8'in bu sırayla düşük faktör yükü ve yüksek hata varyansı nedeni ile analizden çıkarılması ile doğrulanmıştır (Satorra-Bentler $\chi 2/df$ = 2.76 ($p < .001$), *CFI* = .91, *GFI* = .93, *SRMR* = 0.06, *RMSEA* = 0.07 ve *NNFI* = 0.90). İç tutarlılık katsayıları, merkezsizleştirme alt ölçeği için .80 ancak ruminasyon alt ölçeği için .53 olarak bulunmuştur. Hair, Black, Babin ve Anderson (2010)'a göre güvenirlik katsayısının .70 üzerinde olması ölçeğin güvenilir olduğunu göstermektedir. Fraenkel ve Wallen (2006)'ya göre de güvenilir olmayan bir ölçeğin geçerli olması beklenmemelidir. Bu çalışmada da ruminasyon alt ölçeğinin yeterli güvenirliğe sahip olmamasından hareketle ve ölçeğin orijinal hali ve diğer uyarlama çalışmaları (örn; Fresco ve diğ., 2007; Soler ve diğ., 2014) dikkate alınarak sadece merkezsizleştirme maddeleri ile analiz tekrar edilmiştir.

Sadece 14 maddelik Yaşantılar Ölçeği-Merkezsizleştirme ölçeğine uygulanan doğrulayıcı faktör analizi sonucuna göre istatistiksel olarak anlamsız sonuç veren maddeler sırayla modelden çıkarılmıştır. İlk olarak YÖ2 anlamsız *t* değerinden (< .1.96) dolayı, sonrasında sırayla YÖ17 ve YÖ8 düşük faktör yükü (< .32; Tabachnick ve Fidell, 2013) ve yüksek hata varyansı nedeni ile analizden çıkarılmıştır. Bu maddelerin çıkarılmasıyla gerçekleştirilen analiz sonrasında, Şekil 1'de de belirtildiği üzere ölçeğin tek faktörlü yapısının doğrulandığı görülmüştür (Satorra-Bentler χ^2/df = 3.05, *RMSEA* = .07). Ölçeğin CFI değeri .94, NNFI değeri .93, ve son olarak SRMR değeri .06 olarak elde edilmiştir (Tablo 1). Tablo 2' de analize ilişkin standartlaştırılmış yükler, *t* – değeri ve R^2 sonuçları verilmiştir. Standartlaştırılmış yükler ve R^2 sonuçları merkezsizleştirmenin en çok madde YÖ14 (*Kendime şefkatli davranabilirim*) tarafından açıklandığını göstermektedir. İç tutarlılık katsayısı ise Yaşantılar Ölçeği-Merkezsizleştirme için .80 olarak hesaplanmıştır.

Tablo 1. Yaşantılar Ölçeği – Merkezsizleştirme için Uyum Ölçütleri

	χ^2	*df*	χ^2/df	*CFI*	*NNFI*	*NFI*	*RMSEA*	*SRMR*
Model	134.52***	44	3.05	.94	.93	.92	.07	.06

*** $p < .001$



Şekil 1. Yaşantılar Ölçeği Merkezsizleştirme için Tek Faktörlü Model

Not. MER. = Merkezsizleştirme

Tablo 2. Yaşantılar Ölçeği – Merkezsizleştirme için Doğrulayıcı Faktör Analizi Sonuçları (N=363)

Madde	Standardize Edilmiş Faktör Yükleri	Standardize Edilmemiş Faktör Yükleri	*t* - değeri	R^2
YÖ3	.61	.54	12.82	.37
YÖ5	.68	.62	13.09	.46
YÖ6	.57	.52	10.07	.33
YÖ9	.49	.43	8.10	.24
YÖ10	.39	.37	6.11	.15
YÖ12	.50	.43	8.16	.25
YÖ14	.69	.61	13.73	.48
YÖ15	.56	.49	11.07	.31
YÖ16	.34	.27	6.07	.11
YÖ18	.44	.35	8.14	.20
YÖ20	.38	.29	7.19	.14

Ölçüt Bağıntılı Geçerlik

Teorik olarak ruminasyon merkezsizleştirme becerisinin bir nevi zıttı olarak değerlendirildiği için Yaşantılar Ölçeği – Merkezsizleştirme'nin ölçüt bağıntılı geçerliğini değerlendirmek üzere 620

üniversite öğrencisine Ruminasyon Ölçeği Kısa Formu (Treynor ve diğ., 2003) ve Yaşantılar Ölçeği – Merkezsizleştirme ölçekleri uygulamıştır. Ruminasyon ve Merkezsizleştirme arasındaki ilişki Pearson Çarpım Momentler Korelasyon Analizi ile incelenmiştir. Buna göre merkezsizleştirme ve ruminasyon arasında negatif yönde anlamlı bir ilişki bulunmuştur (r = -.29, p< .001). Ayrıca, ruminasyon ölçeğinin alt boyutları olan saplantılı düşünme (r = -.36, p< .001) ve derin düşünme (r = -.16, p< .001) ile merkezsizleştirme arasında negatif yönde anlamlı ilişkiler olduğu görülmüştür.

SONUÇLAR ve TARTIŞMA

Bu çalışmanın amacı, merkezsizleştirme becerisini değerlendirmek amacıyla Yaşantılar ölçeğini Türkçeye uyarlamaktır. Bu bağlamda, ölçeğin geçerlik ve güvenirliği bir grup üniversite öğrencisinde incelenmiştir. Yaşantılar Ölçeğinin iki faktörlü yapısından ruminasyon alt ölçeği güvenilir sonuçlar vermediği için tıpkı ölçeğin orijinal halinde olduğu gibi analizden çıkarılmıştır. Ruminasyon alt ölçeğinin analizden çıkarılmasının ardından Fresco ve diğerleri (2007) tarafından ölçeğin geliştirme aşamasında merkezsizleştirmeyi ölçtüğü varsayılan 14 madde ile analiz gerçekleştirilmiştir. DFA sonucuna göre, YÖ2 anlamsız t değerinden, YÖ17 ve YÖ8 düşük faktör yükü nedeniyle ölçekten sırayla çıkarılmış ve Yaşantılar Ölçeği-Merkezsizleştirme'nin Türkçe formunda ölçeğin orijinalinde olduğu gibi 11 madde yer almıştır. Sonuç olarak ölçeğin tek faktörlü yapısı doğrulayıcı faktör analizi ile doğrulanmıştır. Bu çalışma bulgusu ile paralel olarak araştırmacıların ölçeğin merkezsizleştirmeyi tek faktörlü olarak kabul ettiği görülmektedir (Fresco ve diğ., 2007; Gregoria ve diğ., 2015; Soler ve diğ., 2014).

Sonuçlara bakıldığında, ölçeğin orijinal hali oluşturulurken istatistiksel ve teorik olarak anlamsız bulunan madde 2 ve 8 bu çalışmada da benzer sonuçlar göstermiş ve o maddeler ölçekten çıkarılmıştır. Ölçeğin orijinal formu ile Türkçe formu arasında sadece iki maddede fark bulunmuştur. Bunlar; Madde 17 *("Aslında düşüncelerimden ibaret olmadığımı görebiliyorum.")* çalışmanın örnekleminde anlamlı sonuçlar vermezken orijinal formda anlamlı sonuçlar vermiştir ve Madde 5 ("*Bir şeyler yanlış gittiği zaman kendime nazik davranırım*.") orijinal formda anlamsız sonuçlar verirken bu örneklemde anlamlı sonuçlar vermiştir. Madde 17 ile ilgili olarak, ölçek uyarlama aşamasında ölçeğin bir grup üniversite öğrencisine uygulanması sonrası alınan geribildirimlerde madde 17'nin açık ve anlaşılır olarak bulunmaması ile bağdaşır bir şekilde bu madde istatistiksel olarak da anlamlı sonuç vermemiştir.

Yaşantılar Ölçeği – Merkezsizleştirme'nin güvenirlik çalışması kapsamında, ölçeğin iç tutarlılık katsayısının Türkçe formu için yeterli düzeyde olduğu (.80) görülmektedir. Bu bulgunun ölçeğin Portekizce .82 (Gregório ve diğ., 2015), İngilizce .83 (Fresco ve diğerleri, 2007), İspanyolca .89 (Soler ve diğ., 2014), ve Farsça versiyonunda .77 (Taherif ve diğ., 2017) elde edilen verilerle hemen hemen benzer olduğunu göstermektedir.

Ölçeğin ölçüt bağıntılı geçerliği kapsamında, merkezsizleştirme ve ruminasyon arasındaki ilişki incelenmiştir. Elde edilen verilere göre merkezsizleştirme ve ruminasyon arasında negatif yönde anlamlı bir ilişki olduğu görülmüştür. Bu bulgu, merkezsizleştirme ile ruminasyon arasında negatif yönde ilişki olduğunu gösteren Fresco ve diğerleri (2007) ve Gregório ve diğerleri (2015) tarafından gerçekleştirilen çalışmalar ile tutarlıdır.

Tüm bu sonuçlara bakıldığında, Yaşantılar Ölçeği -Merkezsizleştirme'nin çalışmanın örnekleminde psikometrik açıdan kabul edilebilir düzeyde güvenilir ve geçerli olduğu ve hem doğrulayıcı faktör analizi sonuçlarının hem de iç tutarlılık katsayısının orijinal çalışmada elde edilen bulgularla benzer olduğu görülmüştür. Yaşantılar Ölçeğinin üniversite örnekleminde Türkçeye uyarlamasının üniversite öğrencileri ile yapılacak ruh sağlığı çalışmalarına önemli katkılar sağlayacağı düşünülmektedir. Son olarak, çalışma kapsamında Türkçeye uyarlanan Yaşantılar Ölçeğinin, Farkındalık Temelli Bilişsel Terapi (Teasdale, Segal ve Williams, 1995) ve Duygu Düzenleme Terapisi (Mennin ve Fresco, 2009) gibi merkezsizleştirme becerisini temel ögelerden biri olarak alan terapilerin etkililiğini bu çalışmanın katılımcılarına benzer özellikteki kişilerde değerlendirmek için uygulayıcılar ve araştırmacılar tarafından kullanılabileceği düşünülmektedir.

Bu önemli bulguların yanı sıra çalışma bazı sınırlılıklar barındırmaktadır. Çalışma uygun örnekleme yöntemi ile Ankara'daki tek bir devlet üniversitesinden veri toplanarak gerçekleştirilmiştir. Ayrıca örnekleminin büyük bir çoğunluğu kadın ve eğitim fakültesinde okuyan üniversite öğrencilerinden oluşmaktadır. Bu nedenle, üniversite öğrencilerini daha iyi temsil eden bir örneklem ile başlangıç niteliğindeki bu ölçek uyarlama çalışmasının yinelenmesinin faydalı olacağı düşünülmektedir. Özellikle depresyonun azalmasında etkili olduğu ifade edilen merkezsizleştirme becerisinin klinik örneklemde ölçülmesinin önemi vurgulanmaktadır. Bu nedenle, Yaşantılar Ölçeği – Merkezsizleştirme geçerlik ve güvenirlik çalışmasının Türkiye'de klinik örneklem ile de test edilmesi önerilmektedir. En son olarak bu çalışma tarama yöntemi kullanılarak gerçekleştirildiği ve katılımcılara anket uygulaması yapıldığı için katılımcıların kendilerine verilen veri toplama araçlarını içtenlikle ve nesnel olarak yanıtladıkları sayıltısına dayanmaktadır.

KAYNAKÇA

Albayrak. B. (2015).*Üniversite öğrencilerinin bağlanma biçimleri, bilinçli farkındalık düzeyleri ve psikolojik belirtileri arasındaki ilişkiler* (Yayımlanmamış Yüksek Lisans Tezi, Hasan Kalyoncu Üniversitesi). http://tez2.yok.gov.tr/ adresinden edinilmiştir.

Baer, R. A., & Huss, D. B. (2008). Mindfulness and acceptance-based therapy. In J. L. Lebow (Ed.), *Twenty-first century psychotherapies: Contemporary approaches to theory and practice* (pp. 123-166). United States of America: John Wiley & Sons.

Baer, R. A., Walsh, E., & Lykins, E. L. (2009). Assessment of mindfulness. In F. Didonna (Ed.), *Clinical handbook of mindfulness* (pp. 153-168). New York, NY: Springer.

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression.* New York, NY: Guilford Press.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Ed.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Corcoran, K. M., Farb, N., Anderson, A., & Segal, Z. V. (2010). Mindfulness and emotion regulation. In A. M. Kring & D. M. Sloan (Eds.), *Emotion regulation and psychopathology: A transdiagnostic approach to etiology and treatment* (pp. 339-355). United States of America: Guilford Press.

Erdur-Baker, O., & Bugay, A. (2012). The Turkish version of the Ruminative Response Scale: An examination of its reliability and validity. *The International Journal of Educational and Psychological Assessment, 10*(2), 1-16.

Fraenkel, J. R., & Wallen, N. E. (2006). *How to design and to evaluate research in education* (3rd ed.). United State of America: Mc Graw Hill.

Fresco, D. M., Moore, M. T., van Dulmen, M. H., Segal, Z. V., Ma, S. H., Teasdale, J. D., & Williams, J. M. G. (2007). Initial psychometric properties of the Experiences Questionnaire: Validation of a self-report measure of decentering. *Behavior Therapy, 38*(3), 234-246. http://doi.org/10.1016/j.beth.2006.08.003

Fresco, D. M., Segal, Z. V., Buis, T., & Kennedy, S. (2007). Relationship of posttreatment decentering and cognitive reactivity to relapse in major depression. *Journal of Consulting and Clinical Psychology, 75*(3), 447. doi:10.1037/0022-006X.75.3.447

Gecht, J., Kessel, R., Mainz, V., Gauggel, S., Drueke, B., Scherer, A., & Forkmann, T. (2014). Measuring decentering in self-reports: Psychometric properties of the Experiences Questionnaire in a German sample. *Psychotherapy Research, 24*(1), 67-79. http://dx.doi.org/10.10 80/10503307.2013.821635

Gregório, S., Pinto-Gouveia, J., Duarte, C., & Simões, L. (2015). Expanding research on decentering as measured by the portuguese version of the experiences questionnaire. *The Spanish Journal of Psychology, 18*, 1-14. https://doi.org/10.1017/sjp.2015.18

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis.* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17*(3), 164-172. http://dx.doi.org/10.1027//1015-5759.17.3.164

Hayes, S. C. (2004). Acceptance and commitment therapy and the new behavior therapies: Mindfulness, acceptance, and relationship. In S. C. Hayes, V. M. Follette & M. M. Linehan (Eds.), *Mindfulness and acceptance: Expanding the cognitive-behavioral tradition* (pp. 1-29). New York, NY: Guilford Press.

Hayes-Skelton, S., & Graham, J. (2012). Decentering as a common link among mindfulness, cognitive reappraisal, and social anxiety. *Behavioral and Cognitive Psychotherapy, 1*(1), 1-12. https://doi.org/10.1017 /S1352465812 000902

Heppner, P. P., Kivlighan, D. M., & Wampold, B. E. (2008). *Research design in counseling.* United State of America: Brooks/Cole

Herbert, J. D., & Forman, E. M. (2011). The evolution of cognitive behavior therapy: The rise of psychological acceptance and mindfulness. In J. D. Herbert, & E. M. Forman (Eds.), *Acceptance and mindfulness in cognitive behavior therapy: Understanding and applying the new therapies* (pp. 3-25). Hoboken, NJ: John Wiley & Sons.

Hölzel, B. K., Lazar, S. W., Gard, T., Schuman-Olivier, Z., Vago, D. R., & Ott, U. (2011). How does mindfulness meditation work? Proposing mechanisms of action from a conceptual and neural perspective. *Perspectives on Psychological Science*, *6*(6), 537-559. doi:10.1177/1745691611419671

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. doi:10.1080/107055199 09540118

Kurihara, A., Hasegawa, A., & Nedate, K. (2011). Development of the Japanese version of the Experiences Questionnaire and examination of its reliability and validity. *Japanese Journal of Personality*, *19*(2), 174-177. doi:10.4236/psych.2015.65059

Lafferty, M. (2013). *Mediators of the relationship between mindfulness and alcohol use* (Master thesis, Eastern Illinois University). Retrieved from http://thekeep.eiu.edu/cgi/viewcontent.cgi?article=2164 &context=theses

McCracken, L. M., Gutierrez-Martinez, O., & Smyth, C. (2012). "Decentering" reflects psychological flexibility in people with chronic pain and correlates with their quality of functioning, *Journal of Healthy Psychology*, *32*(7) 1-4. doi:10.1037/a0028093

Mennin, D. S., & Fresco, D. M. (2009). Emotion regulation as an integrative framework for understanding and treating psychopathology. In A. M. Kring & D. M. Sloan (Eds.), *Emotion regulation and psychopathology: A transdiagnostic approach to etiology and treatment* (pp. 356-379). New York, NY: Guilford Press.

Mennin, D. S., & Fresco, D. M. (2014). Emotion regulation therapy. In J. J. Gross (Ed.), *Handbook of emotion regulation* (2nd ed., pp. 469-490). New York, NY: Guilford Press.

Morgan, L. P. (2015). *In the face of adversity: Valued living and decentering as buffering factors in the relations among social disadvantage, psychological distress, drinking to cope and problem drinking* (Doctoral dissertation). Available from Proquest Dissertation and Theses database. (UMI No. 3706474.)

Nolen-Hoeksema, S. (1998). The other end of the continuum: The costs of rumination. *Psychological Inquiry*, *9*(3), 216-219. http://dx.doi.org/10.1207/s15327965pli0903_5

Özyeşil, Z. (2011). *Üniversite öğrencilerinin öz-anlayış düzeylerinin bilinçli farkındalık kişilik özellikleri ve bazı değişkenler açısından incelenmesi* (Doktara Tezi, Selçuk Üniversitesi). http://tez2.yok.gov.tr/ adresinden edinilmiştir.

Pearson, M. R., Brown, D. B., Bravo, A. J., & Witkiewitz, K. (2015). Staying in the moment and finding purpose: The associations of trait mindfulness, decentering, and purpose in life with depressive symptoms, anxiety symptoms, and alcohol-related problems. *Mindfulness*, *6*(3), 645-653. doi:10.1007/s12671-014-0300-8

Safran, J. D., & Segal, Z. V. (1990). *Cognitive therapy: An interpersonal process perspective*. New York, NY: Basic.

Sauer, S., & Baer, R. A. (2010). Mindfulness and decentering as mechanisms of change in mindfulness and acceptance-based interventions. In R. A. Baer (Ed.), *Assessing mindfulness and acceptance processes in clients: Illuminating the theory and practice of change* (pp. 25-50). United States of America: Context Press.

Segal, Z. V., Williams, J. M. G., & Teasdale, J. D. (2002). *Mindfulness-based cognitive therapy for depression: A new approach to relapse prevention.* New York, NY: Guilford Press.

Segal, Z. V., Williams, J. M. G., & Teasdale, J. D. (2012). *Mindfulness-based cognitive therapy for depression.* New York, NY: Guilford Press.

Shapiro, S. L., Carlson, L. E., Astin, J. A., & Freedman, B. (2006). Mechanisms of mindfulness. *Journal of Clinical Psychology*, *62*(3), 373-386. doi:10.1002/jclp.20237

Soler, J., Franquesa, A., Feliu-Soler, A., Cebolla, A., García-Campayo, J., Tejedor, R., ... Portella, M. J. (2014). Assessing decentering: Validation, psychometric properties, and clinical usefulness of the experiences questionnaire in a Spanish sample. *Behavior Therapy*, *45*(6), 863-871. http://doi.org/10.1016/j.beth.2014 .05.004

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson.

Taherifar, Z., Ferdowsi, S., Mootabi, M., Tehrani, M. A., & Fata, L. (2017). Assessing "decentering": Validity, reliability and factor structure of experiences questionnaire in university students. *Knowledge & Research in Applied Psychology*, *18*(1), 46-56. http://journals.khuisf.ac.ir/jsr-p/article-1-2039-en.html

Teasdale, J. D., Segal, Z., & Williams, J. M. G. (1995). How does cognitive therapy prevent depressive relapse and why should attentional control (mindfulness) training help? *Behavior Research and Therapy, 33*(1), 25-39. https://doi.org/10.1016/0005-7967(94)E0011-7

Treynor, W., Gonzalez, R., & Nolen-Hoeksema, S. (2003). Rumination reconsidered: A psychometric analysis. *Cognitive Therapy and Research*, *27*(3), 247-259. https://doi.org/10.1023/A:1023910315561

Ülev, E. (2014). *Üniversite öğrencilerinde bilinçli farkındalık düzeyi ile stresle başa çıkma tarzının depresyon, kaygı ve stres belirtileriyle ilişkisi* (Yayımlanmamış Yüksek Lisans Tezi, Hacettepe Üniversitesi). http://tez2.yok.gov.tr/ adresinden edinilmiştir.

Wheaton, B., Muthen, B., Alwin, D. F., & Summers, G. F. (1977). Assessing reliability and stability in panel models. *Sociological Methodology, 8,* 84-136.

EXTENDED ABSTRACT

Introduction

Decentering has been described as being able to observe or recognize thoughts and feelings as objective and temporary events in the mind rather than as absolute truths (Fresco et al., 2007). Decentering is a term that comes from cognitive therapy (Beck, Rush, Shaw, & Emery, 1979), and it has gained greater importance with mindfulness and acceptance-based therapies (Fresco et al., 2007), and has been considered a common element of all mindfulness-based therapies (Baer & Huss, 2008).

The role of decentering in several factors has been examined in a number of studies. For instance; studies have indicated that decentering is negatively correlated with depression (e.g., Fresco, Segal, Buis & Kennedy, 2007), psychological distress (e.g., Morgan, 2015), social anxiety (e.g., Hayes-Skelton & Graham, 2013), and emotion regulation difficulties (Lafferty, 2013). Moreover, decentering plays a key role in beneficial outcomes of mindfulness (e.g., Corcoran, Farb, Anderson & Segal, 2010).

Recently, mindfulness has become a popular research topic in Turkey (e.g., Albayrak, 2015; Özyeşil, 2011; Ülev, 2014), but there is a lack of research on decentering. Experiences Questionnaire has been used to measure decentering with both undergraduate students and clinical sample (Fresco et al., 2007). Moreover, Experiences Questionnaire (EQ) has been conducted on different cultural groups such as German (Gecht et al., 2014), Japanese (Kurihara, Hasegawa, & Nedate, 2011) and Spanish (Soler et al., 2014). Thus, the aim of the study is to adapt EQ into Turkish. Adapting the EQ into Turkish language can assist both practitioners and researchers to gain more insight about decentering in Turkey. The existing Turkish literature on decentering also can be advanced.

Method

In the present study, the Experience Questionnaire developed to measure both rumination and decentering was adapted into Turkish. The adapted scale included both EQ-Decentering (14 items) and EQ-Rumination (6 items) subscales.

Through the study, the permission to use EQ was obtained from Fresco who is one of the developers of the scale. The researcher also obtained approval from the Middle East Technical University Human Subjects Ethics Committee prior to data collection. Five bilingual persons translated the questionnaire into Turkish. Following the translation process, the best fitted translations of items were selected by the researcher. Secondly, three English language experts from School of Foreign Languages and Faculty of Education identified and worked on the inadequate expressions in translation of the items as well as any discrepancies between the original form and the translated one. Back translation was a required permission agreement offered by Fresco. Thus, back translation of the EQ was conducted by an English language instructor. The researcher applied the questionnaire to ten undergraduate students to get feedback on the comprehensibility of the EQ items.

The participants of this study were 394 undergraduate students from a state university in Ankara. After the data screening process, 31 cases were excluded from the data because of missing values (15 cases),

and univariate and multivariate outliers (16 cases), and the sample size decreased to 363 undergraduate students. There were 251 females (69%) and 112 males (31%). The participants' ages ranged from 18 to 31, and the mean age of the sample was 21.90 (*SD* = 2.27). In order to test the criterion-related validity of the EQ-Decentering, the Short form of the Ruminative Response Scale (Treynor, Gonzealez, & Nolen-Hoeksama, 2003) was used, and the data were collected from 620 undergraduate students (429 females, 191 male).

Results and Discussion

Confirmatory factor analysis indicated an acceptable model fit for the two factor model, but EQ-Rumination subscale had not an adequate reliability (.53). Therefore, EQ-Rumination factor was removed from the data file and the model was re-analyzed using only the decentering items. According to results, item 2 was eliminated because of non-significant *t* value (< .1.96) as found in the original version of the scale, and item 17 and 8 (low standardized loading < .32; Tabacnick & Fidell, 2013) were eliminated from the model. Therefore, confirmatory factor analysis indicated an acceptable model fit; Satorra-Bentler χ^2/df = 3.05 (p < .001), *CFI* = .94, *SRMR* = 0.06 and *NFI* = .92, and Cronbach alpha for EQ-Decentering was calculated as .80.

The Turkish version of EQ-Decentering has strong evidence for construct validity and reliability. There were two differences between the Turkish version of EQ-Decentering and the original measure (Fresco et al., 2007). The first difference was EQ17 ("I can actually see that I am not my thoughts.") loaded significantly in the original measure, but in the Turkish version, EQ17 was removed from the model because of its low standardized loading. The second difference was EQ5 ("I am kinder to myself when things go wrong.") did not load significantly in the original measure; EQ5 loaded significantly in the present study. Except for these two differences, the questionnaire confirmed as having one-factor structure with acceptable fit indexes, and showed similar psychometric properties to the original measure (Fresco et al., 2007).

In order to test the criterion-related validity of the EQ-Decentering, the relationship between EQ-Decentering and the Short form of the Ruminative Response Scale (Treynor et al., 2003) was investigated with the participation of 620 university students. As expected, results revealed that there was a significant and negative relationship between decentering and rumination. This result was consistent with earlier studies (e.g., Fresco et al., 2007; Gregório et al., 2015).

In conclusion, the psychometric properties of the Experiences Questionnaire (EQ; Fresco et al., 2007) were examined in Turkey. Findings revealed its validity and reliability evidences in the current study conducted with undergraduate students. Measuring decentering accurately is important to evaluate the efficacy of theories such as Mindfulness-Based Cognitive Therapy (MBCT; Teasdale, Segal, & Williams, 1995), and Emotion Regulation Therapy (ERT; Mennin & Fresco, 2009) because decentering is one of the basic constructs of those theories (Mennin & Fresco, 2014; Sauer & Baer, 2010). Therefore, Experiences Questionnaire could be used to measure decentering by practitioners and researchers in Turkey.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(2);161-177

Bireyselleştirilmiş Bilgisayarlı Sınıflama Testi Kriterlerinin Test Etkililiği ve Ölçme Kesinliği Açısından Karşılaştırılması*

A Comparison of Computerized Adaptive Classification Test Criteria in Terms of Test Efficiency and Measurement Precision

Ceylan GÜNDEĞER** Nuri DOĞAN***

Öz

Bu çalışmada Bireyselleştirilmiş Bilgisayarlı Sınıflama Testleri'nin (BBST) etkililiğinin sınıflama kriterlerine, madde seçme ve yetenek kestirim yöntemlerine göre nasıl değiştiğinin belirlenmesi amaçlanmıştır. Bu amaçla 3 Parametreli Lojistik Model temel alınmış; belirlenen kesme noktası ve etrafında yüksek bilgi verecek şekilde 500 maddelik bir havuz oluşturulmuş; birey yetenekleri (N(0,1)) 3000 kişi üzerinden türetilmiş ve bireylerin madde cevap örüntüleri R yazılımda rasgele türetilmiştir. Sınıflama kriterlerinden Ardışık Olasılık Oran Testi (AOOT), Genelleştirilmiş Olabilirlik Oranı (GOO) ve Güven Aralığı (GA) yöntemleri; yetenek kestirim yöntemlerinden Beklenen Sonsal Dağılım (BSD) ve Ağırlıklandırılmış Olabilirlik Kestirimi (AOK) yöntemleri; madde seçme yöntemlerinden ise kesme noktasında (KN) ve kestirilen yetenek (KY) temelinde Maksimum Fisher Bilgisi (MFB) ve Kullback-Leibler Bilgisi (KLB) yöntemleri çaprazlanarak 48 koşul oluşturulmuştur. R yazılımında yürütülen BBST simülasyonu sonunda, ortalama test uzunluğu (OTU), ortalama sınıflama doğruluğu (OSD), bireylerin gerçek yetenek düzeyleri ile kestirilen yetenek düzeyleri arasındaki korelasyon (r), yanlılık, RMSE ve ortalama mutlak hata (OMH) değerlerinin 25 tekrara ait ortalamaları hesaplanmıştır. Araştırma sonuçlarına göre test etkililiği bakımından GOO ve GA yöntemlerinin; ölçme kesinliği bakımından ise AOOT'nin daha iyi performans gösterdiği; sınıflama kriterlerinin farksızlık bölgesi genişledikçe veya hata düzeyi değeri küçüldükçe test etkililiğinin arttığı; sınıflama kriterlerinin tümünün her koşulda oldukça yüksek düzeyde sınıflama doğruluğuna sahip olduğu belirlenmiştir. Bireylerin gerçek yetenek düzeyleri ile kestirilen yetenek düzeyleri arasındaki korelasyon bakımından BSD ve AOK yetenek kestirim yöntemlerinin her ikisinin de başarılı kestirimlerde bulundukları ancak ölçme kesinliği bakımından BSD'nin daha iyi performans sergilediği; madde seçme yöntemlerinin ise tümünün birbirine benzer çalıştığı ancak MFB-KY'nin tüm bağımlı değişkenler açısından tüm koşullarda daha iyi performans gösterdiği görülmüştür.

Anahtar Kelimeler: bireyselleştirilmiş bilgisayarlı sınıflama testi, sınıflama kriteri, yetenek kestirimi, madde seçme yöntemi, ölçme kesinliği

Abstract

In this study, it was aimed to determine how the efficiency of the Computerized Adaptive Classification Testing (CACT) changes according to classification criteria, item selection and ability estimation methods. For this purpose, a pool of 500 items, which is based on 3 PLM and informs at the arbitrary cut-point and around, has been generated; individual abilities have been generated using normal distribution (N(0,1)) for 3000

*Bu çalışma ilk yazarın, ikinci yazar danışmanlığında tamamladığı "Bireyselleştirilmiş Bilgisayarlı Sınıflama Testi Kriterlerinin Sınıflama Doğruluğu ve Test Uzunluğu Açısından Karşılaştırılması" isimli doktora tezinden üretilmiştir.

**Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, e-posta: cgundeger@gmail.com , ORCID ID: https://orcid.org/0000-0003-3572-1708
***Prof. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, e-posta: nurid@hacettepe.edu.tr , ORCID ID: orcid.org/0000-0001-6274-2016

DOI: 10.21031/epod.401077
Geliş Tarihi: 03.03.2018
Kabul Tarihi: 22.05.2018

individuals and the item response patterns have been generated randomly in R software with the Monte Carlo simulation. As classification criteria, Sequential Probability Ratio Test (SPRT), Generalized Likelihood Ratio (GLR) and Confidence Interval (CI) methods; as ability estimation methods, Expected a Posteriori (EAP) and Weighted Likelihood Estimation (WLE) methods; and as item selection methods, Maximum Fisher Information (MFI) and Kullback-Leibler Information (KLI) methods on the basis of cut-point (CP) and estimated ability (EA) have been crossed and 48 conditions have been investigated. At the end of the CACT simulations in R, the mean values of Average Test Length (ATL), Average Classification Accuracy (ACA), correlation between the true thetas and estimated thetas (r), bias, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for 25 replications have been calculated. According to the results of the study, it has been observed that the GLR and the CI classification criteria perform better in terms of test efficiency, however the SPRT works better in terms of the measurement precision; test efficiency increases as the indifference region of classification criteria expands or the error value decreases; all classification criteria have considerably high level of the classification accuracy in all conditions. It has been concluded that both ability estimation methods have successful estimation results in terms of the correlation between true and estimated thetas (r); whereas the EAP relatively performs better in terms of the measurement precision; and all of the item selection methods work similarly to each other however the MFI-EA performs better for all conditions in terms of all dependent variables.

Keywords: computerized adaptive classification testing, classification criteria, ability estimation, item selection method, measurement precision

GİRİŞ

Bilgi ve iletişim teknolojilerinde yaşanan gelişmeler, bilgiye ulaşmada ve eğitim uygulamalarında sıklıkla kendini göstermektedir. Bu gelişmeler sayesinde bireylerin öğrenme sürecinde, yeteneklerinin-becerilerinin ölçülmesinde ve değerlendirilmesinde birçok değişiklik meydana gelmektedir. Bu değişikliklerden biri Bireyselleştirilmiş Bilgisayarlı Test (BBT; Computerized Adaptive Testing: CAT) uygulamalarıdır. BBT'de iki temel özellikten bahsedilebilir. Bunlardan ilki bireyin bilgisayar ekranında gördüğü maddeyi cevaplaması iken; ikincisi, testin bireyin yetenek düzeyine göre ayarlanmış olmasıdır (McBride, 1985).

BBT uygulamaları ile Madde Tepki Kuramı'nın (MTK) avantajları sayesinde bireylere yetenek düzeylerine uygun maddeler sunulabilmekte; bireyin aldığı test bireyin yetenek düzeyine göre ayarlanarak bireyselleştirilebilmektedir. Böylece BBT ile geleneksel testlere kıyasla daha kısa zamanda, daha az sayıda maddeyle ve yüksek güvenirlik düzeyinde yetenek kestirimi elde edilebilmektedir (Wainer, 2000). Ayrıca BBT ile hızlı puanlama yapılabilmekte ve bireyler sınav sonucunu uygulama sonunda öğrenebilmektedir. Bu nedenlerle özellikle yurt dışında, GRE (Graduate Record Examination), GMAT (Graduate Management Admission Test) gibi sınavlarda BBT'nin tercih edildiği görülmektedir.

Bireylerin yeteneklerini test etme süreci zaman zaman, belirli bir kesme noktasına (ya da birden fazla sayıda kesme noktasına) dayalı olarak bireyleri başarılı-başarısız, geçti-kaldı (veya düşük-orta-yüksek yetenek düzeyi) vb. sınıflara ayırmayı da hedeflemektedir. BBT'nin bir alt dalı olan Bireyselleştirilmiş Bilgisayarlı Sınıflama Testleri (BBST; Computerized Adaptive Classification Testing: CACT) bireyleri iki ya da daha çok kategoriye ayırmayı amaçlar (Weiss, 1982). Aşağıda bu çalışmanın temelini oluşturan BBST hakkında detaylı bilgiye yer verilmiştir.

Bireyselleştirilmiş Bilgisayarlı Sınıflama Testleri (BBST)

BBT uygulamaları genel olarak, *(i)* Tepki modeli; *(ii)* Madde havuzu; *(iii)* Başlama kuralı; *(iv)* Madde seçme yöntemi; *(v)* Yetenek kestirim yöntemi ve *(vi)* Sonlandırma kuralı olmak üzere altı ana bileşenden oluşmaktadır (Weiss ve Kingsbury, 1984). BBST'de ise ilk beş bileşen aynı olmakla beraber sonlandırma kuralı yerine sınıflama kriterleri kullanılmakta ve bu kriterler aslında BBST'nin odak noktasını oluşturmaktadır. Sınıflama kriterleri sayesinde bireylerin başarılı-başarısız vb. şekilde sınıflara ayrılması söz konusu olmaktadır.

BBST uygulamalarındaki ilk aşama hangi modelin kullanılacağıdır. MTK kapsamında model, çoklu puanlanan maddelere dayanan Ardışık Tepki Modeli (Graded Response Model), Kısmi Kredi Modeli (Partial Credit Model) vb. olabildiği gibi ikili puanlanan maddeleri temel alan 1 PLM, 2 PLM veya 3 PLM olabilmektedir. Bu çalışmada 3 PLM temel alınmıştır. Bu modelde maddelerin ayırt edicilik (a) ve güçlük (b) parametreleri değişkenlik gösterdiği gibi maddelere ait şans parametresi (c) de söz konusudur (Hambleton ve Swaminathan, 1985):

İkinci aşamada madde havuzu yer almaktadır. Genellikle başarı testlerinde kullanılan madde havuzu orta güçlükteki maddelerin yanında çok zor ve çok kolay maddeler içerir. Madde güçlükleri ise tekbiçimli (uniform) dağılıma sahiptir. Ölçüt referanslı testlerde ise madde havuzundaki maddelerin kesme noktası etrafında en yüksek bilgiyi verebilecek madde güçlük değerlerine sahip olması beklenir (Boyd, 2003). Bu çalışmada madde havuzu BBST amacına uygun olarak belirlenen kesme noktası ve etrafında yüksek bilgi veren ve testi alan bireylerin yetenek ranjını kapsayacak şekilde oluşturulmuştur.

BBST'nin üçüncü aşaması başlama kuralının, bir başka deyişle BBST'nin nasıl bir maddeyle başlayacağının belirlenmesidir. Eğer test tekrarlı olarak alınabiliyorsa, testi ikinci kez alanların başlama noktası, bir önceki testten kestirilen yetenek düzeyleri olabilir. Bunun dışında ise genellikle popülasyonun ortalaması atanabilir (Thompson, 2007b). Bu çalışmada başlama noktası, tüm veri setleri ve tüm koşullar için $\theta = 0$ olarak belirlenmiştir.

BBST'nin dördüncü aşaması, madde seçme yönteminin belirlenmesidir. BBT'de, Maksimum Fisher Bilgisi (MFB; Maximum Fisher Information: MFI), Kullback-Leibler Bilgisi (KLB; Kullback-Leibler Information: KLI), a-tabakalama (a-stratified) vb. birçok yöntemin tanımlanmış ve çalışılmış olduğu görülmektedir. Bu çalışmada MFB ve KLB incelenmiştir. MFB bilginin tek bir noktada maksimize edilmesini sağlarken (Embretson ve Reise, 2000); KLB θ_0'dan θ_1'e kadar olan bölgedeki bilgiyi değerlendirir (Eggen, 1999; Akt: Thompson, 2007b). Klasik BBT uygulamalarında madde seçilirken, bireyin kestirilen yetenek (KY) düzeyinde en yüksek bilgiyi veren maddenin seçimi söz konusu iken; BBST uygulamalarında bireyin kestirilen yetenek düzeyinde ve bununla birlikte BBT'den farklı olarak kesme noktasında (KN) en yüksek bilgiyi veren maddenin seçimi gündeme gelmektedir. KY ve KN temelli madde seçimi zeki madde seçim yöntemleri (intelligent item selection) olarak adlandırılmıştır (Thompson, 2007b). KN temelli yöntemlerde kesme noktasında en yüksek bilgiyi sağlayan madde seçilirken; KY temelli yöntemlerde kesme puanı dikkate alınmaksızın bireyin kestirilen geçici yetenek düzeyinde maksimum bilgiyi veren madde seçilmektedir. Bu çalışmada, MFB ve KLB madde seçme yöntemleri KY ve KN temelli madde seçimleriyle çaprazlanmış ve kestirilen yetenekte MFB (MFB-KY), kesme noktasında MFB (MFB-KN), kestirilen yetenekte KLB (KLB-KY) ve kesme noktasında KLB (KLB-KN) olmak üzere dört madde seçme yöntemi incelenmiştir.

BBST'nin beşinci bileşeni yeteneğin kestirilmesidir ve bu bileşen son sınıflama kararlarının etkililiği ve uygunluğu bakımından oldukça önemli bir değişkendir (Yang, Poggio ve Glasnapp, 2006). Wang ve Wang'a (2001) göre bu değişken, raporlanan son yetenek kestirimini etkilediği gibi madde seçimi ve test sonlanmasını da etkilemektedir. Maksimum Olabilirlik Kestirimi (MOK; Maximum Likelihood Estimation: MLE), Beklenen Sonsal Dağılım yöntemi (BSD; Expected a Posteriori: EAP), Maksimum Sonsal Dağılım yöntemi (MSD; Maximum a Posteriori: MAP), Owen'ın Bayesci Kestirim yöntemi gibi alanyazında birçok yetenek kestirim yöntemi yer almaktadır (Wang ve Wang, 2001). Bunların dışında MOK'un geliştirilmiş bir versiyonu olan Ağırlıklandırılmış Olabilirlik Kestirimi (AOK; Weighted Likelihood Estimation: WLE) nadiren de olsa çalışmalarda yer almıştır. Alanyazın incelendiğinde BBST araştırmalarında yetenek kestirim yöntemlerinin pek çalışılmadığı; değişken uzunluklu bu testlerde yöntemlerin birbirlerine kıyasla nasıl performans gösterdiklerinin henüz fazla bilinmediği görülmektedir. Bu sebeple çalışmada ortalama madde sayısını azaltması ve hızlı kestirim kestirimler yapabilmesi bakımından BSD yetenek kestirim yöntemi ile MOK'un yanlılığını azaltmak amacıyla Warm (1989) tarafından geliştirilmiş olan ve yetenek kestiriminde olabilirlik fonksiyonunun modunun yerine ortalamasının dikkate alınmasını sağlayan AOK yetenek kestirim yöntemi incelenmiştir.

BBST'nin BBT'ten farkı ve odak noktasını sınıflama kriteri oluşturmaktadır. Geleneksel BBT'deki sonlandırma kurallarından farklı olarak sınıflama kriterleri temelde bir hipotez testi sürecine dayanmaktadır. Hipotezin kabulüne veya reddine karar verme, bireyi sınıflama çabasının sonuç vermesi anlamına gelmektedir. Sınıflama kriterlerine, Wald (1947) tarafından önerilen Ardışık Olasılık Oran Testi (AOOT; Sequential Probability Ratio Test: SPRT), Weiss ve Kingsbury (1984) tarafından önerilen Bireyselleştirilmiş Uzmanlık Testi (BUT; Adaptive Mastery Testing: AMT), van der Linden (1990) tarafından önerilen Bayesci Karar Kuramı (BKK; Bayesian Decision Theory: BDT), AOOT'nin daha genel bir hali olan Genelleştirilmiş Olabilirlik Oranı (GOO; Generalized Likelihood Ratio: GLR) ve Güven Aralığı (GA; Confidence Interval: CI) yöntemleri örnek olarak gösterilebilir. Bu çalışmada AOOT, GOO ve GA sınıflama kriterlerinin etkililiği incelenmiştir.

AOOT'nin altındaki temel felsefe, iki alternatif hipotez altında gözlenen cevap dağılımının olabilirliğini belirleyerek iki hipotezden birinin doğruluğuna karar verilmesidir. Eğer hipotezlerden birinin olabilirliği diğerinden oldukça büyükse bu hipotez kabul edilirken; iki hipotezin olabilirlikleri benzerse birey yeni bir madde alır ve süreç bu şekilde devam eder (Reckase, 1983). GOO, AOOT'nin modifiye edilmiş daha genel bir halidir. AOOT'de kestirilen yetenek ile gruplara atamada kullanılan yetenek düzeyleri arasındaki eşitlik temel alınırken; GOO'da bu değişklenler arasındaki eşitsizlik durumu da dikkate alınmaktadır. Bu iki sınıflama kriterinde de farksızlık bölgesi (indifference region) ismiyle anılan ve hipotezler yazılırken bireyleri gruplara atamada kullanılacak olan başarılı ve başarısız bölgesine koyulan sabit gündeme gelmektedir. Bu çalışmada AOOT ve GOO sınıflama kriterleri için Nydick'in (2013) çalışması göz önünde bulundurularak tolere edilebilir hatalar için farksızlık bölgesi 0,05 ve 0,10 değerleri dikkate alınmıştır. GA sınıflama kriteri ise sınıflama amacını istatistiksel bir kestirim problemi gibi formüle etmektedir (Eggen ve Straetmans, 2000). GA, ölçmenin koşullu standart hatasını dikkate alarak bireyin kestirilen yetenek düzeyi için kestirimin belirlenen güven aralığına göre kesme noktasının hangi tarafına düştüğünü belirleyen bir yöntemdir. Eğer aralık tam olarak kesme puanının üstündeyse birey geçti-başarılı şeklinde; aralık kesme puanının tam olarak altındaysa kaldı-başarısız olarak sınıflanmaktadır. Aralığın kesme puanını içermesi durumunda ise bireye yeni bir madde sunulmaktadır (Thompson, 2007b). Bu çalışmada GA sınıflama kriteri için Eggen ve Straetmans'in (2000) çalışmalarında incelemiş olduğu %70 ve %90 güven aralığı değerleri ele alınmıştır.

Araştırmanın Amacı ve Önemi:

Cheng ve Liou'ya (2000) göre başarılı bir BBT veya BBST uygulamasında *i)* yetenek kestirim yönteminin uygunluğu ve *ii)* madde seçme yönteminin etkililiği oldukça önemlidir. Bu iki bileşenin farklı sınıflama kriterleriyle birlikte nasıl performans gösterdiğinin belirlenmesi; bir başka deyişle sınıflama kriterlerinin, madde seçme ve yetenek kestirim yöntemlerinin farklı koşullar altında sınıflama doğruluğu, test uzunluğu ve ölçme kesinliği bakımından incelenmesi bu çalışmanın temel amacını oluşturmaktadır.

Çalışmanın odak noktası olan sınıflama kriterleri, 0,05 ve 0,10 farksızlık bölgesi değerleri ile Ardışık Olabilirlik Oranı Testi (AOOT) ve Genelleştirilmiş Olabilirlik Oranı (GOO); %70 ve %90 güven aralığı düzeylerini içeren Güven Aralığı (GA) yöntemleridir. Çalışmada incelenen yetenek kestirim yöntemleri Beklenen Sonsal Dağılım (BSD) ve Ağırlıklandırılmış Olabilirlik Kestirim (AOK) yöntemleri; madde seçme yöntemleri ise kestirilen yetenek temelli MFB (MFB-KY), kesme noktası temelli MFB (MFB-KN), kestirilen yetenek temelli KLB (KLB-KY) ve kesme noktası temelli KLB (KLB-KN) şeklinde belirlenmiştir. Buna göre çalışmanın alt problemleri aşağıdaki gibidir.

BBST simülasyonu sonunda:

1. Yetenek kestirim yöntemi BSD olduğunda AOOT sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GOO sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GA sınıflama kriterinin %70 ile %90 güven düzeyleri için sınıflama doğruluğu, test uzunluğu ve ölçme kesinliği madde seçme yöntemlerinden MFB-KY, MFB-KN, KLB-KY ve KLB-KN'ye göre nasıl değişmektedir?

2. Yetenek kestirim yöntemi AOK olduğunda AOOT sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GOO sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GA sınıflama kriterinin %70 ile %90 güven düzeyleri için sınıflama doğruluğu, test uzunluğu ve ölçme kesinliği madde seçme yöntemlerinden MFB-KY, MFB-KN, KLB-KY ve KLB-KN'ye göre nasıl değişmektedir?
3. Sınıflama kriterlerine, madde seçme yöntemlerine ve yetenek kestirim yöntemlerine göre sınıflama doğruluğu, test uzunluğu ve ölçme kesinliği (r, yanlılık, RMSE ve OMH) değerleri nasıl değişmektedir?

Alanyazındaki BBST çalışmaları incelendiğinde, konunun özellikle yurt dışı alanyazında çalışılmış olduğu ve Türkiye'de çalışılmamış olduğu görülmektedir. Yurt dışında yapılan çalışmalar incelendiğinde ise 1980'lerden bugüne BBST ile ilgili oldukça fazla sayıda çalışmaya rastlanmaktadır. Çalışmalar incelendiğinde sadece sınıflama kriterlerinin karşılaştırılmış olduğu araştırmaların yanında (Huebner, 2012; Jiao ve Lau, 2003; Kingsbury ve Weiss, 1980; Nydick, Nozawa ve Zhu, 2012; Nydick, 2013; Reckase, 1983; Spray ve Reckase, 1996; Thompson ve Ro, 2007; Thompson, 2011; Wouda ve Eggen, 2009) sınıflama kriterlerinin madde seçme yöntemleriyle çaprazlanarak ele alındığı çalışmaların da olduğu göze çarpmaktadır (Eggen, 1999; Eggen ve Straetmans, 2000; Lau ve Wang, 1998, 1999; Lin ve Spray, 2000; Spray ve Reckase, 1994; Thompson, 2007a, 2009).

Yurt dışındaki çalışmalar incelendiğinde sınıflama kriterlerinin madde seçme yöntemleri ve yetenek kestirim yöntemleriyle çaprazlandığı ve sonuçların ölçme kesinliği, test uzunluğu ve sınıflama doğruluğu açısından karşılaştırıldığı herhangi bir çalışmaya rastlanmamıştır. Bu açıdan çalışmanın alanyazına katkı sağlayacağı düşünülebilir. Ayrıca teknolojinin gelişmesi ve eğitimin çağa ayak uydurma çabasının bir sonucu olarak ülkemizde de bilgisayarlı sınavlara doğru bir yönelim olduğu görülmektedir. Buna göre yakın zamanda BBST uygulamalarına geçilebilir. Bu noktada çalışmanın uygulayıcılara, sınıflama kriterleri, madde seçme yöntemleri ve yetenek kestirim yöntemleri hakkında bilgi sağlaması beklenmektedir.

YÖNTEM

Bu çalışma, *"... olsa ne olurdu?"* sorusuna cevap arayan bir Monte Carlo simülasyon çalışmasıdır (Dooley, 2002). Çalışmada hem bireylere ait yetenek parametreleri hem de oluşturulan madde havuzlarının parametreleri R ortamında araştırmacı tarafından türetilmiştir (R Core Team, 2013).

Veri Üretimi

Bu çalışmada bireylerin yetenek parametreleri, (-3,+3) yetenek düzeyleri aralığında, ortalaması 0,0 ve standart sapması 1,0 olacak şekilde normal dağılım yardımıyla toplam 3000 kişi üzerinden random türetilmiştir. Çalışmada birey parametreleri gibi madde parametreleri de simülatif veriden oluşmaktadır. Madde havuzu oluşturmada Thompson'ın (2011) araştırması dikkate alınarak madde havuzunun 3 PLM temelinde 500 maddeden oluşması sağlanmıştır. Araştırmada hem kestirilen yetenek (KY) hem de kesme noktası (KN) temelli madde seçme yöntemleri karşılaştırılacağından madde havuzunun belirlenen kesme noktası olan 1,0 ve etrafında yüksek bilgi verecek; -3,+3 yetenek düzeyleri aralığını kapsayacak şekilde oluşturulmasına dikkat edilmiştir. Bu sebeple havuzdaki maddeler, a parametresinin orta ve yüksek değerlerde olabilmesi adına tekbiçimli dağılımdan [0,5; 2,0] aralığından; b parametresinin, Warm'ın (1989) da çalışmasında belirttiği gibi, gerçek uygulamadaki değerlere yakın olabilmesi adına normal dağılımdan ortalaması 1,0 ve standart sapması 1,5 olmak üzere; c parametresi ise yine gerçek bir uygulama düşünülerek normal dağılımdan ortalaması 0,15 ve standart sapması 0,05 olacak şekilde türetilmiştir. Birey parametrelerinin türetilmesi ve madde havuzunun oluşturulmasının ardından bireylerin madde cevap örüntüsü R ortamında rasgele türetilmiş ve BBST simülasyonuna geçilmiştir.

İşlem

Yetenek parametrelerinin türetilmesi ve madde havuzunun oluşturulması aşamalarından sonra, BBST simülasyonu ***6 sınıflama kriteri x 4 madde seçme yöntemi x 2 yetenek kestirimi yöntemi = 48 koşul*** için yazılan döngülerle 25 tekrarla R'da tamamlanmıştır (R Core Team, 2013). Çalışmanın odak noktası olan 6 sınıflama kriteri; 0,05 ve 0,10 farksızlık bölgesi değerleri ile Ardışık Olabilirlik Oranı Testi (AOOT) ve Genelleştirilmiş Olabilirlik Oranı (GOO) ile %70 ve %90 güven aralığı düzeylerini içeren Güven Aralığı (GA) yöntemleridir. Çalışmada incelenen madde seçme yöntemleri, kestirilen yetenek temelli MFB (MFB-KY), kesme noktası temelli MFB (MFB-KN), kestirilen yetenek temelli KLB (KLB-KY) ve kesme noktası temelli KLB (KLB-KN); yetenek kestirim yöntemleri ise Beklenen Sonsal Dağılım (BSD) ve Ağırlıklandırılmış Olabilirlik Kestirim (AOK) yöntemleri şeklinde belirlenmiştir. BBST simülasyonunda Nydick (2014) tarafından yazılan *catirt* paketinden yararlanılmıştır. Simülasyonda tüm koşullarda başlama noktası, yetenek düzeyi 0 olarak belirlenmiş ve her koşul için bu değer sabit tutulmuş; madde seçme yöntemleri, yetenek kestirim yöntemleri ve sınıflama kriterleri çalışmanın amacına uygun şekilde manipüle edilmiştir. Tüm koşullar için 25 tekrar yapılmıştır.

Verilerin Analizi

Veri analizinde bağımlı değişkenler olan ortalama test uzunluğu (OTU), ortalama sınıflama doğruluğu (OSD), ölçme kesinliğine ilişkin değerler (gerçek yetenekler ile kestirilen yetenekler arasındaki ilişki (r) için Pearson korelasyon katsayısı, yanlılık, RMSE, ortalama mutlak hata: OMH) 25 tekrarın ortalaması olacak şekilde araştırmacı tarafından yazılan fonksiyonlarla R'dan çekilmiştir.

Simülasyon sonuçlarından ortalama test uzunluğu *$test_length* koduyla; bireylerin simülasyon sonunda atandıkları sınıflar ise *$cat_categ* koduyla çekilmiştir. Ortalama sınıflama doğruluğunu hesaplayabilmek amacıyla, simülasyondan çekilen sınıflarla bireylerin türetilen gerçek sınıfları arasındaki uyuma Cohen'in Kappa istatistiğiyle bakılmıştır. Cohen (1960) tarafından geliştirilen Kappa istatistiği, iki veya daha fazla gözlemcinin yaptığı değerlendirmeler arasındaki uyuşmayı belirlemek için kullanılır. Bu uyum -1 ile +1 arasında değer alır. Sıfır değeri tesadüfî uyuşmayı, negatif değerler tesadüfî olmaktan daha kötü bir uyuşmayı ve +1 değeri ise mükemmel uyuşmayı temsil eder (Akt: Şencan, 2005).

Yanlılık, BBST simülasyonu sonucu her birey için kestirilen son yetenek düzeyi ile ($\widehat{\theta\imath}$) bireyin gerçek yetenek düzeyi (θ_i) arasındaki ortalama farklılıktır. Yanlılığın formülü aşağıdaki gibidir (Miller ve Miller, 2004):

$$\text{Yanlılık} = \frac{\sum_{i=1}^{n}(\widehat{\theta\imath} - \theta_i)}{n} \quad (1)$$

RMSE, yanlılığa benzer şekilde tüm koşullar için hataların karesinin ortalamasının karekökü şeklinde hesaplanmıştır:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{\theta\imath} - \theta_i)^2}{n}} \quad (2)$$

Ortalama mutlak hata (OMH) ise, yanlılık formülünde de ele alınan bireyin kestirilen son yetenek düzeyinin gerçek yetenek düzeyinden farkının mutlak değer içerisinde verilmesidir:

$$\text{OMH} = \frac{\sum_{i=1}^{n}|\widehat{\theta\imath} - \theta_i|}{n} \quad (3)$$

BULGULAR

Araştırmanın birinci alt probleminde Monte Carlo simülasyonu BBST uygulamasında, yetenek kestirim yöntemi BSD olduğunda, AOOT sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GOO sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GA sınıflama kriterinin %70 ile %90

güven düzeyleri için sınıflama doğruluğu, test uzunluğu ve ölçme kesinliğinin madde seçme yöntemlerinden MFB-KY, MFB-KN, KLB-KY ve KLB-KN'ye göre nasıl değiştiği incelenmiştir. Aşağıda Tablo 1'de bu alt problemde belirtilen koşulları karşılaştırmada kullanılan bağımlı değişkenler olan ortalama test uzunluğu (OTU), ortalama sınıflama doğruluğu (OSD), bireylerin gerçek yetenek düzeyleri ile kestirilen son yetenek düzeyleri arasındaki korelasyon (r), yanlılık, RMSE ve ortalama mutlak hataya (OMH) ait değerler yer almaktadır. Değerlerin tümü her koşul için 25 tekrarın ortalaması alınarak hesaplanmıştır.

Tablo 1'de madde seçme yöntemi fark etmeksizin ortalama test uzunluğu (OTU) bakımından testi sonlandırmada, bir başka deyişle bireyleri sınıflamada, en az madde gerektiren sınıflama kriterinin GA yönteminin %70 düzeyi olduğu; bunu takiben sırasıyla GA yönteminin %90 güven düzeyinin; GOO FB: 0,10 yönteminin; GOO FB: 0,05 yönteminin ve AOOT FB: 0,10 yönteminin geldiği; en fazla madde gerektiren sınıflama kriterinin ise AOOT FB: 0,05 yönteminin olduğu görülmektedir. Bu bulguya göre, madde seçme yöntemi ile sınıflama kriterlerinin farksızlık bölgesi değerleri ya da güven düzeyleri fark etmeksizin, OTU bakımından en iyi performansı GA sınıflama kriteri göstermiş; bunu sırasıyla GOO ve AOOT sınıflama kriterleri takip etmiştir.

Tablo 1'de sınıflama kriterlerinden GOO ve AOOT yöntemlerinin farksızlık bölgesi değerleri küçüldükçe ve GA yönteminin güven düzeyi yükseldikçe bireyleri sınıflamada gerekli ortalama madde sayısı olan OTU değerlerinin arttığı görülmektedir. Araştırmanın bu bulgusu Reckase (1983), Lau ve Wang (1999), Thompson ve Ro (2007) ve Thompson'ın (2011) araştırma sonuçlarıyla örtüşmektedir. Madde seçme yöntemi fark etmeksizin, ortalama sınıflama doğruluğu (OSD) bakımından GA yöntemi dışındaki sınıflama kriterlerinin tümünün bireyleri geçti-kaldı kategorilerine sınıflamada benzer performans gösterdiği ve sınıflama doğruluğunun oldukça yüksek (0,95 ile 0,97 aralığında) olduğu görülmektedir. GA yönteminin %70 güven düzeyinin (0,95) ve %90 güven düzeyinin (0,96) diğer yöntemlere kıyasla daha düşük ancak yine de yüksek bir sınıflama katsayısına sahip olduğu görülmektedir.

Tablo 1'de bireylerin simülasyon öncesi türetilen gerçek yetenek düzeyleri ile BBST simülasyonu sonucu kestirilen son yetenek düzeyleri arasındaki korelasyon (r) bakımından en yüksek ilişkinin hesaplandığı yöntemin, madde seçme yöntemi fark etmeksizin, AOOT FB: 0,05 yöntemi olduğu; bunu takiben sırasıyla AOOT FB: 0,10 yönteminin, GOO yöntemlerinin ve GA %90 yönteminin geldiği; en düşük ilişkinin ise GA %70 yöntemi ile hesaplandığı görülmektedir. Bu korelasyonlar madde seçme yöntemlerine göre ayrı ayrı incelendiğinde en iyi performansı MFB-KY ve KLB-KY yöntemlerinin verdiği görülmektedir. Bu iki yöntem için hesaplanan korelasyonlar 0,86 ile 0,98 değerleri arasında değişmekte; bu durum da gerçek yeteneklerle kestirilen yetenekler arasındaki korelasyonların bu iki madde seçme yöntemi kullanıldığında oldukça yüksek olduğunu göstermektedir. Diğer madde seçme yöntemlerine ait korelasyon değerleri ise MFB-KN için 0,75 ile 0,92 aralığında ve KLB-KN için 0,75 ile 0,91 aralığında değişmektedir. Bu bulguya dayanarak, bireylerin gerçek yetenek düzeyleri ile kestirilen son yetenek düzeyleri arasındaki korelasyon bakımından, kestirilen yetenek (KY) temelli madde seçme yöntemlerinin kesme noktası (KN) temelli madde seçme yöntemlerine kıyasla daha iyi performans gösterdiği yorumu yapılabilir.

Tablo 1'de sınıflama kriterlerinin oldukça düşük yanlılık değerlerine sahip olduğu ve performanslarının madde seçme yöntemlerine göre farklılık gösterdiği görülmektedir. Madde seçme yöntemi MFB-KY olduğunda sınıflama kriterlerinin neredeyse tümünün yansız performans gösterdikleri; sadece GOO FB: 0,10 ve GA %90 yöntemlerinde düşük bir yanlılık değeri hesaplandığı görülmektedir. Buna göre yanlılık bakımından sınıflama kriterlerinin MFB-KY madde seçme yöntemi ile birlikte başarılı bir performans gösterdikleri yorumu yapılabilir. Madde seçme yöntemi MFB-KN, KLB-KY veya KLB-KN olduğunda ise sınıflama kriterlerinini az da olsa yanlı kestirimler yaptığı görülmektedir.

Tablo 1. Yetenek Kestirim Yöntemi BSD için Koşullara Ait OTU, OSD, r, Yanlılık, RMSE ve OMH Değerleri

Koşullar		Bağımlı Değişkenler					
Madde Seçme Yöntemi	**Sınıflama Kriteri**	**OTU**	**OSD**	**r**	**Yanlılık**	**RMSE**	**OMH**
MFB-KY	**AOOT FB: 0,05**	46,52	0,97	0,98	0,000	0,211	0,170
	AOOT FB: 0,10	31,92	0,97	0,96	0,000	0,254	0,204
	GOO FB: 0,05	16,75	0,97	0,90	0,000	0,403	0,322
	GOO FB: 0,10	15,65	0,97	0,90	0,002	0,412	0,330
	GA %70 güven düzeyi	8,14	0,95	0,86	0,000	0,477	0,387
	GA %90 güven düzeyi	12,68	0,97	0,88	0,002	0,443	0,354
MFB-KN	**AOOT FB: 0,05**	47,02	0,97	0,92	0,001	0,346	0,287
	AOOT FB: 0,10	31,98	0,97	0,88	0,002	0,409	0,342
	GOO FB: 0,05	17,16	0,97	0,82	0,002	0,498	0,423
	GOO FB: 0,10	15,85	0,97	0,81	0,002	0,503	0,430
	GA %70 güven düzeyi	7,67	0,95	0,75	0,002	0,589	0,504
	GA %90 güven düzeyi	13,36	0,97	0,79	-0,002	0,537	0,458
KLB-KY	**AOOT FB: 0,05**	46,42	0,97	0,98	0,000	0,212	0,171
	AOOT FB: 0,10	31,81	0,97	0,96	0,001	0,255	0,205
	GOO FB: 0,05	16,71	0,97	0,90	0,000	0,402	0,323
	GOO FB: 0,10	15,66	0,97	0,90	-0,001	0,412	0,330
	GA %70 güven düzeyi	8,12	0,95	0,86	-0,002	0,480	0,389
	GA %90 güven düzeyi	12,53	0,96	0,88	0,000	0,446	0,357
KLB-KN	**AOOT FB: 0,05**	47,24	0,97	0,91	0,001	0,349	0,290
	AOOT FB: 0,10	32,03	0,97	0,88	0,000	0,410	0,343
	GOO FB: 0,05	17,04	0,97	0,82	0,000	0,499	0,425
	GOO FB: 0,10	15,84	0,97	0,81	0,003	0,504	0,430
	GA %70 güven düzeyi	7,63	0,95	0,75	0,000	0,592	0,507
	GA %90 güven düzeyi	13,30	0,97	0,79	-0,001	0,537	0,459

Tablo 1'e göre, yanlılık ile birlikte kestirimin standart hatasını da dikkate alan RMSE ve ortalama mutlak hata (OMH) değerlerine göre, madde seçme yöntemi fark etmeksizin, en iyi performans gösteren sınıflama kriteri AOOT FB: 0,05 yöntemidir. Bunu takiben sırasıyla AOOT FB: 0,10; GOO FB: 0,05; GOO FB: 0,10; GA %90 ve GA %70 yöntemlerinin geldiği görülmektedir. Farksızlık bölgesi ve güven düzeyleri fark etmeksizin RMSE ve OMH bakımından en iyi performansı AOOT yöntemi göstermiştir. AOOT yönteminin ardından GOO yöntemi gelmiş ve göreli olarak en kötü performansı ise GA yöntemi göstermiştir.

Tablo 1'de görülüğü üzere, Monte Carlo simülasyonu BBST uygulamasında yetenek kestirim yöntemi BSD olduğunda, madde seçme yöntemi fark etmeksizin, test etkililiği (ortalama test uzunluğu ve ortalama sınıflama doğruluğu) bakımından GA %70 yönteminin diğerlerine kıyasla oldukça başarılı bir performans gösterdiği ortaya çıkmıştır. Bunu sırasıyla GA %90; GOO FB: 0,10; GOO FB: 0,05; AOOT FB: 0,10 ve AOOT FB: 0,05 sınıflama kriterleri izlemektedir. Bireyler geçti-kaldı kategorilerine GA %70 sınıflama kriteri ile ortalama 8 madde ve ortalama 0,95 sınıflama doğruluğuyla sınıflanabilirken; diğer yöntemlerin tümünde ortalama sınıflama doğruluğu 0,97 olmak üzere GA %90 sınıflama kriteri ile ortalama 14 madde; GOO FB: 0,10 sınıflama kriteri ile ortalama 16 madde; GOO FB: 0,05 sınıflama kriteri ile ortalama 17 madde; AOOT FB: 0,10 sınıflama kriteri ile ortalama 32 madde ve AOOT FB: 0,05 sınıflama kriteri ile de ortalama 47 maddeyle testin sonlanabildiği ve bireylerin kategorilere yerleştirilebildiği görülmüştür. Test etkililiği açısından bakıldığında GA ve GOO sınıflama kriterlerinin AOOT'ye kıyasla başarılı performans gösterdikleri görülmektedir. Bununla birlikte kestirilen ve gerçek yetenek düzeyleri arasındaki ilişki (r), yanlılık, RMSE ve OMH değerleri bakımından AOOT sınıflama kriterinin göreli olarak diğer yöntemlerden daha iyi performans gösterdiği; ancak tüm koşullar içinden bu göreli değerlerin en düşük olduğu MFB-KY madde seçme yönteminin ve AOOT FB: 0,05 sınıflama kriterinin birlikte ele alındığı koşulda, GA %70 sınıflama kriterinin neredeyse 6 katı maddede testin sonlandığı-bireylerin sınıflanabildiği dikkat çekmektedir. Bu noktada dikkat edilmesi gereken bir bulgu olarak, ortalama test uzunluğunun azalmasının mutlak hatayı artırdığı; bir başka deyişle BBST'de daha az sayıda madde kullanıldığında mutlak hata değerinin yükseldiği görülmektedir.

İkinci Alt Probleme İlişkin Bulgular

Araştırmanın ikinci alt probleminde Monte Carlo simülasyonu BBST uygulamasında, yetenek kestirim yöntemi AOK olduğunda, AOOT sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GOO sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GA sınıflama kriterinin %70 ile %90 güven düzeyleri için sınıflama doğruluğu, test uzunluğu ve ölçme kesinliğinin madde seçme yöntemlerinden MFB-KY, MFB-KN, KLB-KY ve KLB-KN'ye göre nasıl değiştiği incelenmiştir. Aşağıda Tablo 2'de bu alt problemde belirtilen koşulları karşılaştırmada kullanılan bağımlı değişkenler olan ortalama test uzunluğu (OTU), ortalama sınıflama doğruluğu (OSD), bireylerin gerçek yetenek düzeyleri ile kestirilen son yetenek düzeyleri arasındaki korelasyon (r), yanlılık, RMSE ve ortalama mutlak hataya (OMH) ait değerler yer almaktadır. Değerlerin tümü her koşul için 25 tekrarın ortalaması alınarak hesaplanmıştır.

Tablo 2'de madde seçme yöntemlerinden MFB-KY ve KLB-KY kullanıldığında ortalama test uzunluğu (OTU) bakımından testi sonlandırmada, bir başka deyişle bireyleri sınıflamada, en az madde gerektiren sınıflama kriterinin GA %70 yönteminin olduğu; bunu takiben sırasıyla GA %90; GOO FB: 0,10; GOO FB: 0,05 ve AOOT FB: 0,10 yöntemlerinin geldiği; en fazla madde gerektiren sınıflama kriterinin ise AOOT FB: 0,05 yönteminin olduğu görülmektedir. Madde seçme yöntemlerinden MFB-KN ve KLB-KN kullanıldığında ise OTU bakımından iyi performans gösteren sınıflama kriterlerinin GA %70; GOO FB: 0,10; GOO FB: 0,05; GA %90; AOOT FB: 0,10 ve AOOT FB: 0,05 yöntemleri şeklinde sıralandığı görülmektedir. Bu bulguya göre OTU bakımından en iyi performansı KY temelli madde seçme yöntemleri kullanıldığında GA sınıflama kriteri göstermiş, bunu sırasıyla GOO ve AOOT sınıflama kriterleri takip etmiş iken; KN temelli madde seçme yöntemlerinde GA yönteminin %90 güven düzeyinin GOO yöntemine kıyasla daha kötü

performans gösterdiği ve yine bireyleri sınıflamada en fazla sayıda madde gerektiren sınıflama kriterinin AOOT olduğu görülmüştür.

Tablo 2. Yetenek Kestirim Yöntemi AOK için Koşullara Ait OTU, OSD, r, Yanlılık, RMSE ve OMH Değerleri

Koşullar		Bağımlı Değişkenler					
Madde Seçme Yöntemi	**Sınıflama Kriteri**	**OTU**	**OSD**	**r**	**Yanlılık**	**RMSE**	**OMH**
MFB-KY	**AOOT FB: 0,05**	46,78	0,97	0,98	-0,004	0,217	0,174
	AOOT FB: 0,10	32,30	0,97	0,96	-0,013	0,270	0,214
	GOO FB: 0,05	16,73	0,97	0,89	0,019	0,444	0,344
	GOO FB: 0,10	15,71	0,97	0,89	0,025	0,455	0,353
	GA %70 güven düzeyi	8,45	0,95	0,86	0,033	0,514	0,409
	GA %90 güven düzeyi	12,57	0,96	0,87	0,024	0,487	0,380
MFB-KN	**AOOT FB: 0,05**	47,01	0,97	0,91	0,123	0,353	0,297
	AOOT FB: 0,10	31,98	0,97	0,87	0,197	0,425	0,364
	GOO FB: 0,05	17,04	0,97	0,81	0,286	0,533	0,469
	GOO FB: 0,10	15,98	0,97	0,80	0,295	0,542	0,477
	GA %70 güven düzeyi	9,38	0,96	0,76	0,355	0,615	0,545
	GA %90 güven düzeyi	17,84	0,97	0,83	0,262	0,514	0,443
KLB-KY	**AOOT FB: 0,05**	46,74	0,97	0,96	-0,001	0,219	0,175
	AOOT FB: 0,10	32,20	0,97	0,96	-0,013	0,270	0,214
	GOO FB: 0,05	16,80	0,97	0,89	0,028	0,450	0,346
	GOO FB: 0,10	15,77	0,97	0,89	0,025	0,463	0,355
	GA %70 güven düzeyi	8,26	0,95	0,85	0,044	0,525	0,416
	GA %90 güven düzeyi	12,58	0,96	0,87	0,026	0,488	0,380
KLB-KN	**AOOT FB: 0,05**	47,22	0,97	0,91	0,127	0,356	0,299
	AOOT FB: 0,10	32,03	0,97	0,87	0,197	0,427	0,366
	GOO FB: 0,05	17,06	0,97	0,81	0,287	0,534	0,470
	GOO FB: 0,10	15,94	0,96	0,80	0,296	0,545	0,478
	GA %70 güven düzeyi	9,32	0,96	0,76	0,353	0,616	0,546
	GA %90 güven düzeyi	17,99	0,97	0,83	0,259	0,512	0,442

Tablo 2'de sınıflama kriterlerinden GOO ve AOOT yöntemlerinin farksızlık bölgesi değerleri küçüldükçe ve GA yönteminin güven düzeyi yükseldikçe bireyleri sınıflamada gerekli ortalama madde sayısı olan OTU değerlerinin arttığı görülmektedir. Araştırmanın bu bulgusu Reckase (1983), Lau ve Wang (1999), Thompson ve Ro (2007) ve Thompson'ın (2011) araştırma sonuçlarıyla örtüşmektedir. Madde seçme yöntemi fark etmeksizin ortalama sınıflama doğruluğu (OSD) bakımından sınıflama kriterlerinin tümünün bireyleri geçti-kaldı kategorilerine sınıflamada benzer performans gösterdiği ve sınıflama doğruluğunun 0,95 ile 0,97 aralığında oldukça yüksek olduğu görülmektedir. Madde seçme yöntemlerinden MFB-KY veya KLB-KY kullanıldığında sınıflama kriterlerinden GA %70 ve GA %90 yöntemlerinin; MFB-KN kullanıldığında GA sınıflama kriterinin %70 güven düzeyinin; KLB-KN kullanıldığında ise GA %70 yöntemi ile GOO FB: 0,10 yönteminin diğer sınıflama kriterlerine kıyasla düşük OSD değerlerine sahip olsalar da sınıflama doğruluklarının yüksek olduğu görülmüştür.

Tablo 2'de gerçek yetenek düzeyleri ile kestirilen yetenek düzeyleri arasındaki korelasyon (r) bakımından en yüksek ilişkinin hesaplandığı yöntemin MFB-KY yöntemi kullanıldığında AOOT FB: 0,05 yöntemi olduğu; bunu takiben sırasıyla AOOT FB: 0,10 yönteminin, GOO yöntemlerinin ve GA %90 yönteminin geldiği; en düşük ilişkinin ise GA %70 yöntemi ile hesaplandığı görülmektedir. Tüm sınıflama kriterlerine ait r değerleri KY temelli yöntemler için 0,85 ile 0,98 aralığında ve KN temelli yöntemler için de 0,76 ile 0,91 aralığındadır. Bu bulguya göre r bakımından KY temelli madde seçme yöntemlerinin KN temelli madde seçme yöntemlerine kıyasla daha iyi performans gösterdiği ortaya koyulmuştur.

Tablo 2'de sınıflama kriterlerinin Tablo 1'e kıyasla daha yüksek yanlılık değerlerine sahip olduğu; bir başka deyişle yetenek kestirim yöntemi olarak BSD yerine AOK kullanıldığında yanlılık değerlerinin yükseldiği ve yöntemlerin performanslarının madde seçme yöntemlerine göre farklılık gösterdiği görülmektedir. Madde seçme yöntemi fark etmeksizin AOOT sınıflama kriterinin en düşük yanlılık değerlerine sahip olduğu görülmektedir. Bununla birlikte GA yönteminin %90 güven düzeyinin, madde seçme yöntemi MFB-KY veya KLB-KY olduğunda GOO FB: 0,10 yönteminden daha düşük yanlılık değerine; madde seçme yöntemi MFB-KN veya KLB-KN olduğunda ise GOO yönteminin her iki farksızlık bölgesine kıyasla daha az yanlılığa sahip olduğu görülmektedir. Buna dayanarak GA %90 sınıflama kriterinin madde seçme yöntemleriyle birlikte yanlılık bakımından iyi sonuçlar verdiği ve özellikle KN temelli madde seçme yöntemleriyle ele alındığında oldukça düşük yanlılık değeri verdiği söylenebilir.

Tablo 2'de yanlılık ile birlikte kestirimin standart hatasını da dikkate alan RMSE ve ortalama mutlak hata (OMH) değerlerine göre madde seçme yöntemi fark etmeksizin, en iyi performans gösteren sınıflama kriteri AOOT FB: 0,05 yöntemidir. Bununla birlikte GA yönteminin %90 güven düzeyinin, madde seçme yöntemi MFB-KN veya KLB-KN olduğunda, GOO yönteminin her iki farksızlık bölgesine kıyasla daha düşük RMSE ve OMH değerine sahip olduğu görülmektedir. Buna dayanarak GA %90 sınıflama kriterinin özellikle KN temelli madde seçme yöntemleriyle ele alındığında oldukça başarılı performans gösterdiği söylenebilir.

Tablo 2'de görülüğü üzere BBST simülasyonunda yetenek kestirim yöntemi AOK olduğunda, madde seçme yöntemi fark etmeksizin test etkililiği (ortalama test uzunluğu ve ortalama sınıflama doğruluğu) bakımından GA %70 yönteminin diğerlerine kıyasla oldukça başarılı bir performans gösterdiği görülmektedir. Bunu sırasıyla KY temelli madde seçme yöntemlerinde GA %90; GOO FB: 0,10; GOO FB: 0,05; AOOT FB: 0,10 ve AOOT FB: 0,05 ve KN temelli madde seçme yöntemlerinde ise GOO FB: 0,10; GOO FB: 0,05; GA %90; AOOT FB: 0,10 ve AOOT FB: 0,05 sınıflama kriterleri izlemektedir.

Tablo 2'de bireyler geçti-kaldı kategorilerine GA %70 sınıflama kriteri ile ortalama 9 madde ve ortalama 0,96 sınıflama doğruluğuyla sınıflanabilirken; diğer yöntemlerin tümünde ortalama sınıflama doğruluğu 0,96 ya da 0,97 olmak üzere GA %90 sınıflama kriteri ile ortalama 16 madde; GOO FB: 0,10 sınıflama kriteri ile ortalama 16 madde; GOO FB: 0,05 sınıflama kriteri ile ortalama 17 madde; AOOT FB: 0,10 sınıflama kriteri ile ortalama 32 madde ve AOOT FB: 0,05 sınıflama

kriteri ile de ortalama 47 maddeyle testin sonlanabildiği ve bireylerin kategorilere yerleştirilebildiği görülmüştür. Test etkililiği açısından bakıldığında GA ve GOO sınıflama kriterlerinin AOOT'ye kıyasla başarılı performans gösterdikleri söylenebilir. Bununla birlikte kestirilen ve gerçek yetenek düzeyleri arasındaki ilişki (r), yanlılık, RMSE ve OMH değerleri bakımından AOOT sınıflama kriterinin göreli olarak diğer yöntemlerden daha iyi performans gösterdiği; ancak tüm koşullar içinden bu göreli değerlerin en düşük olduğu KLB-KY madde seçme yönteminin ve AOOT FB: 0,05 sınıflama kriterinin birlikte ele alındığı koşulda, GA %70 yönteminin neredeyse 6 katı maddede testin sonlandığı-bireylerin sınıflanabildiği dikkat çekmektedir. Bu noktada dikkat edilmesi gereken bir bulgu olarak, ortalama test uzunluğunun azalmasının mutlak hatayı artırdığı; bir başka deyişle BBST'de daha az sayıda madde kullanıldığında mutlak hata değerinin yükseldiği görülmektedir.

Üçüncü Alt Probleme İlişkin Bulgular

Araştırmanın üçüncü alt probleminde sınıflama kriterlerinin, madde seçme yöntemlerinin ve yetenek kestirim yöntemlerinin ayrı ayrı olmak üzere ortalama test uzunluğu (OTU), ortalama sınıflama doğruluğu (OSD), korelasyon (r), yanlılık, RMSE ve ortalama mutlak hata (OMH) bakımından nasıl değiştikleri incelenmiştir. Aşağıda Tablo 3'te tüm bağımsız değişkenlere ilişkin bilgiler özetlenmiştir.

Tablo 3. Sınıflama Kriterlerinin, Madde Seçme Yöntemlerinin ve Yetenek Kestirim Yöntemlerinin OTU, OSD, r, Yanlılık, RMSE ve OMH Değerleri

Bağımsız Değişken		*OTU*	*OSD*	*r*	*Yanlılık*	*RMSE*	*OMH*
Sınıflama Kriterleri	AOOT	39,45	0,97	0,93	0,039	0,311	0,257
	GOO	16,36	0,97	0,85	0,079	0,475	0,394
	GA	11,24	0,96	0,82	0,085	0,523	0,436
Madde Seçme Yöntemleri	MFB	22,36	0,97	0,87	0,067	0,436	0,362
	KLB	22,35	0,97	0,87	0,068	0,438	0,363
	KY	22,00	0,97	0,91	0,008	0,384	0,304
	KN	22,71	0,97	0,83	0,127	0,490	0,421
	MFB-KY	22,02	0,97	0,91	0,007	0,382	0,303
	MFB-KN	22,69	0,97	0,83	0,127	0,489	0,420
	KLB-KY	21,97	0,97	0,91	0,009	0,385	0,305
	KLB-KN	22,72	0,97	0,83	0,127	0,490	0,421
Yetenek Kestirim Yöntemleri	BSD	22,04	0,97	0,87	0,001	0,424	0,352
	AOK	22,65	0,97	0,87	0,135	0,449	0,373

Tablo 3'e göre sınıflama kriterlerinden, yöntemlerin farklı hata düzeyleri veya farksızlık bölgesi değerleri dikkate alınmaksızın, en az sayıda maddeyle sınıflama yapabilen yöntemin yaklaşık 11 maddeyle GA yöntemi olduğu; bunu 16 maddeyle GOO yönteminin takip ettiği ve en çok sayıda maddeyle sınıflama yapabilen yöntemin de 40 maddeyle AOOT olduğu görülmektedir. Sınıflama kriterlerinin ortalama sınıflama doğruluğu bakımından ise benzer sonuçlar verdikleri görülmektedir. Bu bulgu, test etkililiği bakımından GA ve GOO yöntemlerinin BBST uygulamalarında daha kullanışlı olacağına işaret etmektedir.

Tablo 3'te bireylerin türetilen gerçek yetenek düzeyleri ile BBST uygulaması sonucu kestirilen son yetenek düzeyleri arasındaki korelasyon (r) bakımından sınıflama kriterlerinin üçünün de iyi performans gösterdikleri görülmektedir. Korelasyonlar bakımından en iyi sonucu AOOT sınıflama kriteri verirken; GOO ve GA yöntemlerinin benzer şekilde çalıştıkları görülmüştür. Bu sonuç test etkililiği ile birlikte düşünüldüğünde, GOO ve GA sınıflama kriterlerinin uygulamada avantaj sağlayacağı yorumu yapılabilir. Yanlılık, RMSE ve OMH bakımından GA sınıflama kriterinin diğer iki yönteme kıyasla daha kötü performans gösterdiği; en iyi performansı ise AOOT sınıflama kriterinin sergilediği görülmektedir. Bu sonuçlar test etkililiği ile birlikte düşünüldüğünde, AOOT sınıflama kriteri sonuçlarının hatadan daha arınık olmasına karşın; bu yöntemin test etkililiği bakımından kullanışlı olmadığı dikkat çekmektedir.

BBST uygulamalarından beklenen, bireyleri az sayıda maddeyle yüksek doğrulukta sınıflamaktır. Bu açıdan bakıldığında, alt problem için elde edilen tüm sonuçları düşünerek, GOO sınıflama kriterinin diğer iki yönteme kıyasla daha kullanışlı bir seçenek olduğu söylenebilir. Bu alt problem için elde edilen bulgular, Thompson (2011) ile Nydick, Nozawa ve Zhu'nun (2012) çalışmalarının sonuçları ile örtüşmektedir. Bahsedilen çalışmalarda GOO'nun test etkililiği bakımından en kullanışlı sınıflama kriteri olduğu ortaya çıkmıştır.

Tablo 3'e göre, madde seçme yönteminin hangi ölçütü temele aldığı fark etmeksizin, MFB ve KLB madde seçme yöntemlerinin bağımlı değişkenler bakımından birbirine oldukça benzer performans gösterdiği görülmektedir. Bu bulgu Eggen (1999), Lau ve Wang (1999), Cheng ve Liou (2000) ile Lin ve Spray'in (2000) çalışma sonuçlarına benzerlik göstermekte iken; Spray ve Reckase (1994) ile Lau ve Wang'ın (1998) araştırma sonuçlarıyla örtüşmemektedir. Alanyazın incelendiğinde bu iki madde seçme yönteminin performansları hakkında araştırmacılar tarafından fikir birliğinin sağlanamadığı görülmektedir.

Madde seçme yönteminin hangi temele dayandığı incelendiğinde ise kestirilen yeteneğe dayanan (KY) madde seçiminin kesme noktasına dayanan (KN) madde seçimine kıyasla bağımlı değişkenler bakımından daha iyi performans sergilediği görülmektedir. KY ve KN temelli yöntemler ortalama test uzunluğu ve ortalama sınıflama doğruluğu bakımından benzer sonuçlar vermiş olsa da gerçek yetenek düzeyleriyle kestirilen son yetenek düzeyleri arasındaki korelasyon, yanlılık, RMSE ve OMH değerleri incelendiğinde KY temelli madde seçme yönteminin daha başarılı olduğu görülmektedir. Alanyazında bu karşılaştırmaya az sayıda çalışmada yer verilmiş ve bu araştırmalarda da yöntemlerin etkililiği hakkında ortak bir sonuca ulaşılamamıştır. Örneğin Spray ve Reckase (1994) çalışmasında kesme noktasında (KN) en yüksek bilgiyi veren madde seçme yöntemiyle daha kısa test oluştuğunu gösterirken; Thompson (2007b, 2009) araştırmalarında tam aksini, geçici yetenek düzeyinde (KY) en yüksek bilgi veren maddenin seçilmesi durumunda testin kısaldığını ortaya koymuştur. Bu açıdan çalışmanın bu bulgusu Thompson'ın (2007b, 2009) araştırma sonuçlarıyla örtüşmektedir.

Tablo 3'teki dört madde seçme yöntemi incelendiğinde, madde seçme yöntemlerinden en iyi performansı MFB-KY'nin sergilediği; bunu KLB-KY'nin takip ettiği ve MFB-KN ile KLB-KN'nin ise benzer performans gösterdiği görülmektedir. Bu bulgu Eggen ve Straetmans'ın (2000) çalışma sonuçlarıyla örtüşmektedir.

Tablo 3'e göre BSD ve AOK yetenek kestirim yöntemlerinin OTU, OSD ve r bakımından oldukça benzer çalıştıkları ancak yanlılık, RMSE ve OMH değerleri bakımından BSD'nin AOK'a kıyasla göreli olarak daha iyi performans sergilediği görülmektedir. Bu bulgu Wang, Hanson ve Lau'nun (1999) çalışma sonuçlarıyla AOK'un değişken uzunluklu testlerde yüksek yanlılık değerine sahip olması bakımından örtüşmektedir. Ayrıca Yi, Wang ve Ban (2000) araştırmalarında değişken uzunluklu testlerde AOK'un BSD'den daha fazla sayıda madde gerektirdiği sonucuna ulaşmışlardır. Tablo 3'te BSD ile test ortalama 22 maddede sonlanırken; yetenek kestirim yöntemi AOK olduğunda bu ortalama 23'e çıkmaktadır. Çalışmanın bu bulgusu da alanyazın ile örtüşmektedir.

SONUÇLAR ve TARTIŞMA

Bu araştırmada, BBST uygulamalarındaki sınıflama kriterleri, yetenek kestirim yöntemleri ve madde seçme yöntemleri Monte Carlo simülasyonu altında incelenmiştir. Araştırma sonucunda, tüm koşullarda, madde seçme yöntemi ve yetenek kestirim yöntemi fark etmeksizin, bireyleri sınıflamada en az sayıda madde gerektiren sınıflama kriterlerinin sırasıyla Güven Aralığı (GA) yöntemi, Genelleştirilmiş Olabilirlik Oranı (GOO) ve Ardışık Olasılık Oran Testi (AOOT) olduğu; AOOT ve GOO sınıflama kriterleri için farksızlık bölgesi genişledikçe ve GA sınıflama kriteri için ise hata düzeyi değeri küçüldükçe ortalama test uzunluğunun azaldığı; sınıflama kriterlerinden elde edilen ortalama sınıflama doğruluklarının birbirine yakın değerler aldığı ve bu değerlerin oldukça yüksek düzeyde sınıflama doğruluğuna işaret ettiği görülmüştür. Bu sonuçlar bir arada düşünüldüğünde test etkililiği (test uzunluğu ve sınıflama doğruluğu) bakımından gerçek uygulamalarda, daha az sayıda maddeyle yüksek doğrulukta sınıflama yapabilmeleri sebebiyle, GA ve GOO sınıflama kriterlerinin tercih edilmesi önerilmektedir.

Çalışma sonuçlarına göre daha az sayıda maddeyle testin sonlanabilmesi, bireyin ait olduğu kestirilen kategoriye daha kısa zamanda atanabilmesi için farksızlık bölgesinin geniş tutulması (örneğin 0,05 yerine 0,10 değerinin alınması) veya güven aralığı değerinin daha küçük (örneğin %90 yerine %70 olarak) belirlenmesi gerekmektedir. Farksızlık bölgesi daraldıkça veya güven aralığı değeri yükseldikçe bireyin bir kategoriye atanması zorlaşmakta, daha fazla sayıda maddeye ihtiyaç duyulmakta, bu da test etkililiğini düşürmektedir.

Çalışmanın bir diğer sonucu, sınıflama kriterlerinin kestirilen yetenekler ile gerçek yetenekler arasındaki korelasyonlar (r) bakımından yüksek düzeyde ilişki verdiği; buna dayanarak sınıflama kriterlerinin BSD veya AOK ile birlikte yetenek kestiriminde başarılı olduğu; sınıflama kriterlerinden, madde seçme yöntemi ve yetenek kestirim yöntemi fark etmeksizin, yanlılık, RMSE ve ortalama mutlak hata bakımından göreli olarak en iyi performansı AOOT yönteminin gösterdiği; bunu GOO yönteminin takip ettiği ve en kötü performansı ise GA yönteminin gösterdiği belirlenmiştir. Bu noktada dikkati çeken bir durum, daha az sayıda maddeyle testin sonlanmasının mutlak hatayı artırmasıdır. Bu sonuçlara dayanarak ölçme kesinliği bakımından en iyi performansı gösteren sınıflama kriterlerinin AOOT ve GOO yöntemleri olduğu görülmüştür. Bir önceki paragraftaki sonuçlara dayanarak test etkililiği, ölçme kesinliği ile bir arada düşünüldüğünde GOO yönteminin diğer iki sınıflama kriterine kıyasla daha başarılı performans sergilediği görülmüştür ve bu sebeple de uygulayıcılara gerçek uygulamalarda GOO sınıflama kriterinin kullanılmasının daha uygun olacağı önerilmektedir.

Çalışmada incelenen madde seçme yöntemleri olan MFB ve KLB'nin, sınıflama kriteri ve yetenek kestirimi yöntemi fark etmeksizin, ortalama test uzunluğu, ortalama sınıflama doğruluğu, kestirilen yetenek ile gerçek yetenek düzeyi arasındaki korelasyon, yanlılık, RMSE ve ortalama mutlak hata bakımından birbirine oldukça benzer çalıştıkları; madde seçme yöntemlerinin dayandığı temel bakımından kestirilen yetenek (KY) ve kesme noktası (KN) temelli yöntemlerin ortalama test uzunluğu ve ortalama sınıflama doğruluğu bakımından birbirine benzer performans gösterdikleri ancak kestirilen yetenek ile gerçek yetenek düzeyi arasındaki korelasyon, yanlılık, RMSE ve ortalama mutlak hata bakımından KY'nin KN'ye kıyasla daha iyi performans gösterdiği; madde seçme yöntemlerinden MFB-KY yönteminin ortalama test uzunluğu, ortalama sınıflama doğruluğu, kestirilen yetenek ile gerçek yetenek düzeyi arasındaki korelasyon, yanlılık, RMSE ve ortalama mutlak hata bakımından diğer yöntemlere kıyasla daha iyi performans gösterdiği belirlenmiştir. Bu sonuçlar bir arada düşünüldüğünde uygulayıcılara, test etkililiği ve ölçme kesinliği bakımından daha iyi performans göstermiş olması sebebiyle, MFB-KY'nin tercih edilmesinin uygun olacağı düşünülmektedir.

Yetenek kestirim yöntemlerinin ortalama test uzunluğu, ortalama sınıflama doğruluğu ve kestirilen yetenek ile gerçek yetenek düzeyi arasındaki korelasyon açısından benzer çalıştıkları ancak BSD'nin çok düşük yanlılık değerine; göreli olarak daha küçük RMSE ve ortalama mutlak hata değerine sahip olduğu görülmüştür. Bu sonuçla birlikte değişken uzunluklu BBST uygulamalarında BSD'nin AOK'a kıyasla daha iyi bir kestirici olduğu görülmüştür. Bu sonuca dayanarak da gerçek uygulamalarda test etkililiği ve ölçme kesinliği bakımından yetenek kestirim yöntemlerinden BSD'nin tercih edilmesi önerilmektedir.

KAYNAKÇA

Boyd, A. M. (2003). Strategies for controlling testlet exposure rates in computerized adaptive testing systems. (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database. (UMI No. 3110732)

Cheng, P. E. & Liou, M. (2000). Estimation of trait level in computerized adaptive testing. *Applied Psychological Measurement, 24*(3), 257–265

Dooley, K. (2002). Simulation research methods. In J. Baum (Ed.). Companion to organizations. London: Blackwell.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*(3), 249-261

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

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologist*. London: Lawrence Erlbaum Associates Publishers

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

Huebner, A. (2012). Item overexposure in computerized classification tests using sequential item selection. *Practical Assessment, Research & Evaluation, 17(12), 1-9.*

Jiao, H. & Lau, A. C. (2003). The Effects of Model Misfit in Computerized Classification Test. The annual meeting of the National Council of Educational Measurement. Chicago, IL, April 2003. [Online: http://iacat.org/sites/default/files/biblio/ji03-01.pdf , Accessed date: 17.5.2018.]

Kingsbury, G. G. & Weiss, D. J. (1980). A Comparison of Adaptive, Sequential and Conventional Testing Strategies for Mastery Decisions. Research Report 80-4. [Online: http://iacat.org/sites/default/files/biblio/ki80-04.pdf , Accessed date: 17.5.2018.]

Lau, C. A. & Wang, T. (1998, April). Comparing and combining dichotomous and polytomous items with SPRT procedure in computerized classification testing. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Lau, C. A. & Wang, T. (1999, April). Computerized classification testing under practical constraints with a polytomous model. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Lin, C. J. & Spray, J. (2000). Effects of item-selection criteria on classification testing with the sequential probability ratio test. ACT Research Report Series 2000-8. [Online: https://eric.ed.gov/?id=ED445066, Accessed date: 17.5.2018.]

McBride, J. R. (1985). Computerized adaptive testing. *Educational Leadership, 43*(2), 25 -28

Miller, I. & Miller, M. (2004). *John E. Freund's mathematical statistics with applications*. New Jersey: Prentice Hall

Nydick, S. W., Nozawa, Y. & Zhu, R. (2012, April). Accuracy and efficiency in classifying examinees using computerized adaptive tests: an application to a large scale test. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Nydick, S. W. (2013). Multidimensional mastery testing with CAT. (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database. (UMI No. 3607925)

Nydick, S. W. (2014). *catirt: An R Package for Simulating IRT-Based Computerized Adaptive Tests*. [Online: https://cran.r-project.org/web/packages/catIrt/catIrt.pdf, Accessed date: 17.5.2018.]

R Core Team. (2013). *R: A language and environment for statistical computing*, (Version 3.0.1), Vienna, Austria: R Foundation for Statistical Computing. Online: http://www.R-project.org/

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.). New horizons in testing: latent trait theory and computerized adaptive testing. New York: Academic Press.

Spray, J. A. & Reckase, M. D. (1994, April). The selection of test items for decision making with a computer adaptive test. Paper presented at the annual meeting of the National Council on Measurement in Education, NewOrleans, LA.

Spray, J. A. & Reckase, M. D. (1996). Comparison of SPRT and sequential bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21(4),* 405-414.

Şencan, H. (2005). *Sosyal ve davranışsal ölçümlerde güvenirlilik ve geçerlilik*. Ankara: Seçkin Yayıncılık.

Thompson, N. A. & Ro, S. (2007). Computerized classification testing with composite hypotheses. In D. J. Weiss (Ed.). Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing. Accessed date: [17.5.2018] from http://iacat.org/sites/default/files/biblio/cat07nthompson.pdf

Thompson, N. A. (2007a). *A comparison of two methods of polytomous Computerized classification testing for multiple cutscores*. (Unpublished Doctoral Dissertation). University of Minnesota.

Thompson, N. A. (2007b). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment Research & Evaluation, 12(1)*, 1-13

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69(5)*, 778-793

Thompson, N. A. (2011). Termination criteria for computerized classification testing. *Practical assessment, Research & Evaluation, 16(4)*, 1-7

van der Linden, W. J. (1990). Applications of decision theory to test-based decision making. In R. K. Hambleton & J. N. Zaal (Eds.). Advances in educational and psychological measurement. Massachusetts: Kluwer-Nijhof.

Wainer, H. (2000). *Computerized adaptive testing: a primer*. New Jersey: Lawrence Erlbaum Associates

Wald, A. (1947). *Sequential analysis*. New York: John Wiley

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

Wang, S. & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement, 25(4)*, 317–331

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54(3)*, 427-450

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6(4)*, 473-492

Weiss, D. J. & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21(4)*, 361-375

Wouda, J. T. & Eggen, T. J. H. M. (2009). Computerized classification testing in more than two categories by using stochastic curtailment. In D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. Accessed date: [17.5.2018] from http://iacat.org/sites/default/files/biblio/cat09wouda.pdf

Yang, X, Poggio, J. C. & Glasnapp, D. R. (2006). Effects of estimation bias on multiple category classification with an IRT-based adaptive classification procedure. *Educational and Psychological Measurement, 66(4)*, 545-564

Yi, Q., Wang, T. & Ban, J. (2000). Effects of scale transformation and test termination rule on the precision of ability estimates in CAT. ACT Research Report Series, 2000-2. [Online: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1745-3984.2001.tb01127.x, Accessed date: 17.5.2018.]

EXTENDED ABSTRACT

Introduction

Because of the advantages of Item Response Theory (IRT) such as invariance of item parameters and person parameters Computerized Adaptive Testing (CAT) is getting more attention in last years. When the CAT applications are used to classify individuals into two or several groups according to one or more cut-point, Computerized Adaptive Classification Testing (CACT), which is a sub-field of CAT becomes a current issue. CACT aims to classify the persons with the highest classification accuracy using the least number of items according to one or more predefined cut-points and has six components: *(i)* Response model; *(ii)* Item pool; *(iii)* Starting rule; *(iv)* Item selection method; *(v)* Ability estimation method and *(vi)* Classification criteria. The efficiency of the classifications varies by item pools, classification criteria, item selection methods and ability estimation methods. According to this, in the CACT, forming different patterns and identification of these patterns under Monte Carlo (MC) simulations are important for real applications.

In this study, different classification criteria, various methods for item selection and ability estimation in the CACT, are compared using classification accuracy, test length and precision of measurement under the MC simulations. In our research, as classification criteria, Sequential Probability Ratio Test (SPRT), Generalized Likelihood Ratio (GLR) and Confidence Interval (CI) methods; as ability estimation methods, Expected a Posteriori (EAP) and Weighted Likelihood Estimation (WLE) methods; and as item selection methods, Maximum Fisher Information (MFI) and

Kullback-Leibler Information (KLI) methods on the basis of cut-point (CP) and estimated ability (EA) have been examined. The importance of this study comes from the 48 conditions that have not been investigated before Therefore, it is expected to provide some information about classification criteria, item selection methods and ability estimation methods to the researchers or practitioners.

Method

The purpose of this study was to identify the effects of the classification criteria, item selection methods and ability estimation methods in CACT on the classification accuracy, test length and precision of measurement. To achieve this aim, a pool of 500 items, which is based on 3 PLM and informs at the arbitrary cut-point and around, was generated. In this pool, items were simulated from uniform distribution as U[0,5; 2,0] for a parameters; normal distribution as N(1, 1,5) for b parameters and as N(0,15, 0,05) for c parameters. Individual abilities were derived from normal distribution as N(0,1) between (-3,+3) ability levels for 3000 individuals. The item response patterns were generated randomly in R software.

After the data production process, the CACT simulation study was performed for the 48 conditions (*6 classification criteria* ***x*** *4 item selection methods* ***x*** *2 ability estimation methods*) in R with the codes (with 25 *for cycles* for each condition) written by the researcher. At the end of the CACT simulations, the mean values of Average Test Length (ATL), Average Classification Accuracy (ACA), correlation between the true thetas and estimated thetas (r), bias, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for 25 replications have been calculated.

Results and Discussion

According to results of the study, it has been observed that the GLR and the CI classification criteria perform better compared to the SPRT in terms of test efficiency; however the SPRT works better compared to the other two methods in terms of bias, RMSE and MAE. It has also been deduced that the ATL decreases and test efficiency increases as the indifference region of classification criteria expands or the error value decreases.

In addition, it has been concluded that all classification criteria have considerably high level of the classification accuracy in all conditions; and both ability estimation methods, the EAP and the WLE, have successful estimation results in terms of the correlation between true and estimated thetas (r); whereas the EAP relatively performs better than the WLE in terms of the bias, RMSE and MAE. It has also been observed that, all of the item selection methods work similarly to each other however, the MFI-EA performs better for all conditions in terms of all dependent variables.

In conclusion, it can be said that the GLR method is the most preferable classification criteria in terms of test efficiency and precision of measurement and it is necessary to expand the indifference region of the SPRT or the GLR; or to decrease the error value of the CI in order to increase the test efficiency of CACT. In addition, because the EAP performs better than the WLE in terms of the precision of measurement, EAP can be used in real CACT applications. Lastly, the MFI item selection method on the basis of estimated ability (MFI-EA) can be the most appropriate item selection method for the real CACT.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(2);178-201

Toplam Test ve Alt Test Puanlarının Kestiriminin Hiyerarşik Madde Tepki Kuramı Modelleri ile Karşılaştırılması*

Comparison of Estimation of Total Score and Subscores with Hierarchical Item Response Theory Models

Sümeyra SOYSAL ** Hülya KELECİOĞLU ***

Öz

Bu araştırmada güvenilir alt test ve toplam test puanı kestirimleri konusuna katkı sağlamak amacıyla alt test ve toplam test arasındaki ilişki hiyerarşik madde tepki kuramı modelleri ile araştırılmak istenmiştir. Çalışmada Üst Düzey Sıralı (Higher Order), İki Faktör (Bi-factor) ve hiyerarşik çok boyutlu madde tepki kuramı (ÇBMTK) modelleri ile kestirilen toplam test puanının ve alt test puanlarının RMSE ve güvenirlik değerleri alt test sayısı, alt test uzunluğu ve alt testler arasındaki korelasyonların büyüklüğü koşulları altında karşılaştırılmıştır. Ayrıca TEOG 2015 verileri üzerinde araştırmada kullanılan üç kestirim modelinin performansı incelenmiştir. Araştırmanın sonucunda iki ve üç boyutlu verilerde hemen hemen tüm koşullarda alt test uzunluğu ve alt testler arasındaki korelasyonun arttıkça üç kestirim modelinden elde edilen toplam test puanı için yetenek parametreleri kestirim hatasının azaldığı, kestirim güvenirliğinin ise arttığı bulunmuştur. Toplam test puanları için Hiyerarşik ÇBMTK model ile tüm koşullarda en düşük RMSE değeri ve en yüksek güvenirlik değeri elde edilmiştir. Ayrıca korelasyonun 0.8 düzeyinde toplam test puanı için tüm modeller birbirine yakın RMSE ve güvenirlik değerleri ile kestirim yapmıştır. İki ve üç boyutlu verilerde alt test puanı için kestirilen yetenek parametrelerinin RMSE değerleri, Hiyerarşik ÇBMTK modelde alt test uzunluğu arttıkça azalırken alt testler arasındaki korelasyon düzeyinden etkilenmediği; Üst Düzey Sıralı modelde alt test uzunluğu ve alt testler arasındaki korelasyon arttıkça azaldığı; İki Faktör modelde ise alt test uzunluğu arttıkça azalırken alt testler arasındaki korelasyon arttıkça önemli düzeyde arttığı bulunmuştur.

Anahtar Kelimeler: Alt test puan kestirimi, toplam test puan kestirimi, hiyerarşik madde tepki kuramı modelleri, üst düzey sıralı model, iki faktör model

Abstract

In this study, the relationship between subtest and total test was investigated by using hierarchical item response theory models in order to contribute to reliable subtest and total test score estimates. The RMSE and reliability of the total test score and subtest scores estimated by the Higher Order, Bi-factor and hierarchical MIRT models in the study were compared under the conditions of the size of the correlations between the subtests, subtest length and number of subtests. In addition, the performance of three models used in the research was examined on TEOG 2015 data. As a result of the study, in almost all conditions, when the correlation between the subtest and the subtest length increased, the RMSE of the ability parameters decreased and the reliability increased for the total test score obtained from the three estimation models. Under all conditions, the lowest RMSE values and the highest reliability values were yielded from Hierarchical MIRT model for subtest score recovery and from Hierarchical MIRT model for total test score recovery. In addition, all models estimated RMSE and reliability values close to each other at 0.8 level of correlation for total test score recovery. The RMSE values of the ability parameters for the subtest scores in two and three dimensional data were found to be not affected by the correlation level between the subtests while the subtest length decreased in the Hierarchical MIRT model; were found to decrease as the correlation between subtest and subtest length in the Higher Order model and were found to decrease as the subtest length increased, but significantly increased as the correlation between the subtests increased in the Bi-factor model.

* Bu çalışma, ilk yazarın, ikinci yazar danışmanlığında tamamladığı "Toplam Test Puanı ve Alt Test Puanlarının Kestiriminin Hiyerarşik Madde Tepki Kuramı Modelleri ile Karşılaştırılması" isimli doktora tezinden üretilmiştir.

**Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, sumeyrasoysal@hotmail.com, ORCID ID: orcid.org/0000-0002-7304-1722

*** Prof. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, hulyakelecioğlu@gmail.com, ORCID ID: orcid.org/0000-0002-0741-9934

Geliş Tarihi: 11.03.2018
DOI: 10.21031/epod.404089 *Kabul Tarihi: 13.06.2018*

Keywords: subtest scoring, overall test scoring, hierarchical item response theory models, higher order model, bi-factor model

GİRİŞ

Pek çok gelişmiş ülkede, geniş ölçekli standart testler, eğitimde ve psikolojide kullanılan en yaygın ölçme araçlarıdır. Ülkemizde bu araçlar, bir eğitim programına giriş, sertifika alımı ya da personel seçimi gibi önemli kararların verildiği durumlarda sıklıkla kullanılmaktadır. Birçok ülke geniş ölçekli testleri kendi eğitim sistemlerine yönelik bilgiler toplama, eğitime yönelik karar verme ve planlama aşamalarında sıkça kullanmaktadır. SAT ve ACT gibi Amerika'da yapılan geniş ölçekli sınavların burs verme ve eyaletlerin eğitim politikalarının değerlendirilmesi gibi ikincil amaçları vardır. Öğrenci gelişimlerinin izlenmesinde, okulların başarı durumlarının/performanslarının yıllara göre incelenmesinde, öğretim programlarının araştırılması, değerlendirilmesi ve geliştirilmesinde geniş ölçekli testlerden yararlanılmaktadır. Bu testlerin sonuçlarından öğrenciler, öğretmenler, veliler, yöneticiler ve diğer paydaşlar farklı şekillerde faydalanmaktadır.

Geniş ölçekli sınavlar, genellikle hem farklı yapıları hem de bir yapının farklı alt alanlarını ölçen alt bölümlerden oluşur. Bu alt bölümlere genellikle alt test, alt bölümlerden elde edilen puanlara da alt test puanı denir. Örneğin, KPSS sınavındaki eğitim bilimleri alt testi, kendi içinde sekiz konu alanına ayrılır (öğrenme psikolojisi, gelişim psikoloji, ölçme ve değerlendirme, rehberlik ve özel eğitim, öğretim ilke ve yöntemleri, program geliştirme, sınıf yönetimi, öğretim teknolojileri ve materyal tasarımı). Bir alt testteki maddeler bir yeteneği, bir konu alanını ya da bir örtük yapıyı ölçmek için düzenlenirler. Ülkemizde bu testlerin sonuçları alt testlerin ağırlıklı ortalamalarından elde edilen toplam puanlar ile ifade edilir. Böyle birleşik puanlar genel olarak birey başarısını değerlendirmek için yeterli bilgiyi sağlayabilirler. Çünkü geniş ölçekli testlerin en başta oluşturulma amacı bireyleri sıralamaktır.

ABD'de 2001 yılında kabul edilen Hiçbir Çocuk Geride Kalmasın (No Child Left Behind) Yasası gereğince her eyalet okul ilerleme durumunu ölçmek amacıyla toplam test puanı yanında öğrencilerin matematik, okuma, yazma, fen bilimleri gibi temel konu alanlarındaki puanlarını da raporlaması gerekmektedir. Brennan (2012) test puanlarını kullananların çoğunlukla toplam test puanıyla birlikte alt testlerin tanısal amaçlar için raporlanmasını talep ettiğini belirtmiştir. Ayrıca Haladyna ve Kramer (2005), testin başlangıçtaki temel amacı ne olursa olsun, öğrenciler, öğretmenler, veliler, yöneticiler ve diğer paydaşlar tarafından farklı alt alanlar ya da alt bölümlere ait puanların büyük talep gördüğünü raporlamışlardır (akt. Ling, 2012).

Alt testler, formatif (biçimlendirici) ve summatif (özetleyici) değerlendirmelere, eğitim programlarının değerlendirilmesine ve öğretmen değerlendirmelerine bilgi sağlayabilecek bir potansiyele sahiptir. Benzer şekilde, alt testler, toplam puanla karşılaştırıldığında, bireylerin yeteneklerinin farklı alanlarda nasıl değiştiğini/ çeşitlendiğini belirlemek için daha bilgilendirici olabilmektedir. Bilişsel bir yapıyı, bir yeteneği ya da psikolojik bir yapıyı temsil eden ve bunlara yönelik tanılayıcı bilgiler sağlayan alt testler, sınıf içi ve dışı etkinliklerin düzenlenmesinde de yararlı olabilmektedir. Testlerde başarısız olan bireyler, testin kapsamında yer alan konu alanları, yeterlik alanları ya da bilişsel yapılar içinde başarılı ve başarısız oldukları noktaları bilmek istemektedir. Böylece bireyler, çalışma planlarını eksik ya da zayıf oldukları konuları tamamlayabilmek için daha etkili şekilde düzenleyebilme imkanı bulmaktadır (Haladyna ve Kramer, 2004). Ayrıca alt testler öğrencilerin güçlü ve zayıf oldukları noktalar hakkında bilgi sağlayarak öğretmenlerin ders programlarını düzenlemesine katkı sağlayabilmektedir. Yine alt testlerin sağladığı bilgilerle veliler çocuklarının durumları ile ilgili bilgilenirken, onların eksik veya başarısız oldukları konular için destekleyici tedbirler alma ya da onların potansiyellerine göre yönlendirici imkanlar sağlama konusunda daha etkili çözüm üretebilmektedir.

Alt testlerden elde edilebilecek bilgilerden yararlanabilmek için öncelikle alt test puanlarının test geliştirme süreçleri açısından bazı önemli özellikleri sağlaması gerekmektedir. (ETS, 2014; Ferrara ve DeMauro, 2007). İlk olarak, alt test puanları güvenirlik, geçerlik, ayırt edicilik açısından yeterli

psikometrik niteliklere sahip olmalıdır. Psikolojide ve Eğitimde Test Geliştirme Standartları 5.12'ye (AERA, APA, NCME,1999) göre test puanlarının geçerliği, güvenirliği ve karşılaştırılabilirliği sağlanmadıkça raporlanmaması gerekir. Yine aynı standartların 1.12'ye göre, bir test birden fazla puan sağlıyorsa farklı puanların ayırıcılıklarının gösterilmesi gerekir. Benzer şekilde, Ferrara ve DeMauro (2007) orta düzeyde ilişkili ve yüksek güvenirliğe sahip alt testlerin raporlanmasını, düşük güvenirlikli alt testlerin ise raporlanmaması gerektiğini belirtmişlerdir. Alt testlerin güvenirliğine ek olarak testin yapısının da incelenmesi gerekir. Messick'e (1989, s.43) göre maddeler arası ilişkiler yapının alt testlerini ya da alt alanlarını yansıtmalı ve bu da test puanları ve onların yorumlanması düzeyinde ele alınmalıdır. İkinci olarak, Haberman (2008) ve Haberman, Sinharay ve Puhan (2009) alt test puanlarının toplam puan üzerinde bir değeri olup olmadığının belirlenmesi gerektiğini belirtmişlerdir.

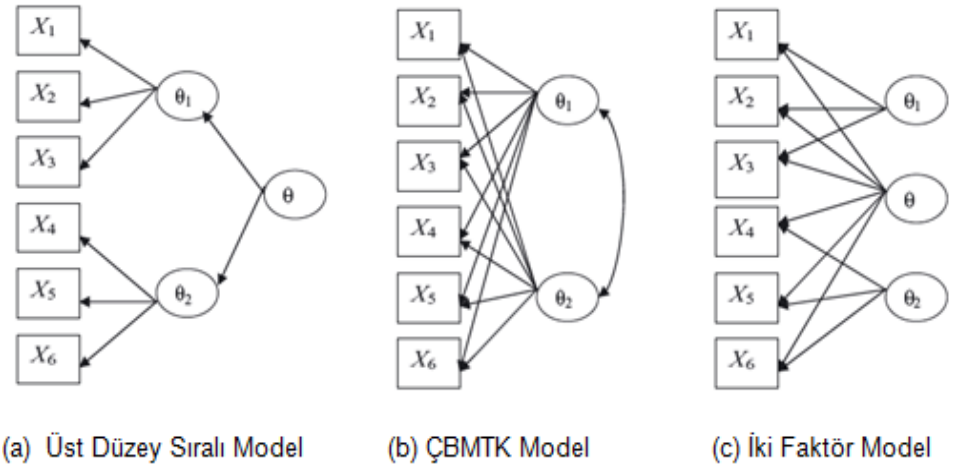
Bir testin alt birimlerinden (alt yeterlik alanları, alt testler, alt ölçekler vb) elde edilen puanların genel olarak zayıf psikometrik özelliklere sahip olduğu belirtilmiştir (Monaghan, 2006; Skorupski ve Carvajal, 2010). Alt birimler testin toplamına göre daha az sayıda madde içermesi nedeniyle daha düşük güvenirliğe sahip olabilmektedir. Sinharay (2010) yetersiz test uzunluğuna sahip olması nedeniyle alt testlerden alınan puanların güvenilir olmasa dahi potansiyel tanı değerleri nedeniyle raporlanmasının yararlı olabileceğini belirtmesine rağmen güvenilir alt test puanlarının raporlanması gerekli ve önemlidir (Haberman, 2008). Bu durumda, "Az sayıda madde içermesi nedeniyle bir testin daha küçük alt birimlerinden güvenilir puanlar elde edilebilir mi veya alt testler arası korelasyon, alt test uzunluğu gibi test özelliklerinin alt test puanı kestirimleri üzerindeki etkisi nedir?" gibi sorular ortaya çıkmaktadır.

Alt test puan ya da birleşik toplam puan kestirimlerinde geleneksel Klasik Test Kuramı'na (KTK) dayalı toplam puanı ya da düzeltilmiş toplam puanı kullanan yöntemler bulunmaktadır (Kelley, 1927, 1947; Wainer ve diğerleri, 2001). KTK'nın madde ve birey örneklemine bağlı olması nedeniyle daha güvenilir alt test puan kestirimlerinde tek boyutlu Madde Tepki Kuramı'na (MTK) dayalı yöntemler geliştirilmiştir (Wainer ve diğerleri 2001;Yen, 1987). Her ne kadar alt test puanı kestirimlerinde tek boyutlu MTK'ya dayalı yöntemlerin KTK'ya dayalı yöntemlere göre daha güvenilir sonuçlar verdiği gösterilse de tek boyutlu yaklaşımlar alt testler arası ilişkileri göz ardı etmektedir. Geniş ölçekli testlerde toplam puanların alt testlerin tek boyutlu olduğu varsayımı altında ve birbirleri arasındaki ilişkilerden bağımsız olarak analiz edilmesi durumunda eğer test maddeleri yerel bağımlı ise güvenirliğin ve test parametrelerinin yanlı kestirilmesi beklenen bir sonuçtur (Brandt ve Duckor, 2013; Wang ve Wilson, 2005;Yen, 1980). Bu sorunun çözümünde alt testlerden oluşan testler için toplam test yetenek kestirimlerinde MTK 'ya dayalı çeşitli modellerin kullanıldığı görülmektedir. Alt testlerden kaynaklanan yerel bağımlılığın, test yapısına (madde demetleri kullanımı gibi) veya ölçülen psikolojik yapıya (alt yeterlik alanları gibi) bağlı olup olmadığına göre bu modeller sırasıyla testlet modeller (Bradlow, Wainer ve Wang, 1999; Wang ve Wilson, 2005) ya da hiyerarşik modeller (de la Torre ve Song, 2009; Gibbons ve Hedeker, 1992; Sheng ve Wikle, 2008) olarak gösterilmektedir.

Alt testleri çok boyutlu MTK çerçevesinde analiz eden birkaç çalışma bulunmaktadır (de la Torre, Song ve Hong, 2011; Sheng ve Wikle, 2007; Wang, Chen ve Cheng, 2004; Yao, 2010; Yao ve Boughton, 2007). Ayrıca, ülkemizde yapılan bilimsel araştırmalar incelendiğinde çok boyutlu yetenek parametresi kestirimleri veya alt testlere yönelik çalışmaların da oldukça az olduğu dikkat çekmektedir (Çakıcı Eser, 2015; Köse, 2010; Özkan, 2012). Çok boyutlu MTK'nın alt test puanı kestirimlerinde kullanılması konusunda daha fazla araştırmaya ihtiyaç olduğu düşünülmektedir. Alt testlerden oluşan ölçme araçlarından elde edilen puanların güvenirliği; ölçme aracının ölçtüğü teorik yapıların açıklanması ya da bileşke yapıya ilişkin kurulan modelin test edilmesi gibi yapı geçerliği; sınıflama geçerliği; çapraz geçerliği gibi ölçme araçlarının psikometrik özellikleri üzerinde MTK'ya dayalı yöntemlerin performansının daha fazla incelenmesi gerektiği düşünülmektedir. Belirlenen bu ihtiyaca bağlı olarak, ölçme araçlarının psikometrik özelliklerinden güvenirlik üzerinde çalışılmaya karar verilmiştir. Güvenilir alt test ve toplam test puanı kestirimleri konusuna katkı sağlamak amacıyla alt test ve toplam test arasındaki ilişki hiyerarşik MTK modelleri ile araştırılmak istenmiştir.

Araştırmada Kullanılan Hiyerarşik MTK Modelleri

Araştırmada kullanılan çok boyutlu madde tepki kuramı modelleri alt testler ile toplam test arasındaki ilişkiyi hiyerarşik düzeyde ele alması nedeniyle "hiyerarşik MTK modelleri" olarak adlandırılmıştır. Modellerin yapısal gösterimleri Şekil 1'de sunulmuştur.



Şekil 1. Araştırmada Kullanılan Hiyerarşik Çok Boyutlu MTK Modelleri

Araştırmada Hiyerarşik ÇBMTK Model için alt testlerin madde ve yetenek parametre kestirimlerinde matematiksel formülü Reckase (1997) tarafından ifade edilen çok boyutlu üç parametreli lojistik model kullanılmıştır. Hiyerarşik ÇBMTK Model'de toplam test puanı için yetenek kestirimleri Yao ve Schwarz (2006) tarafından tanımlanan Maksimum Bilgi Yöntemi ile elde edilmiştir. Maksimum test bilgi yöntemi ile mümkün olan tüm açı değerleri için elde edilen varyans değerini en küçük yapacak açı değeri hesaplanmaktadır. Böylece maksimum bilgiye sahip en güvenilir θ_α -birleşik (composite) puan- elde edilmektedir. Sonuç olarak bu yöntem ile Hiyerarşik ÇBMTK modelde toplam/genel test puanı ile alt test puanları arasında lineer bir ilişki kurulmaz. Aksine toplam puan ve alt test puanları arasındaki ilişkilerin farklı yetenek düzeylerinde veya farklı puan düzeylerinde farklılaşabileceği gerçeği dikkate alınarak toplam test puanları elde edilmektedir (Yao, 2010).

Üst Düzey Sıralı Model'de alt test puanları ile toplam test puanı arasında lineer bir ilişki kurulmaktadır. Bu ilişki toplam test puanı ile alt test puanları arasındaki korelasyonlara dayanmaktadır. Bu yaklaşıma göre alt testler kendi içlerinde tek boyutludur ama bütün alt testler dolaylı olarak genel bir boyutla ilişkilidir. Alan yazında böyle yapılara çoklu-tek boyutlu (multi-unidimensional) test yapıları da denilmektedir (Sheng ve Wickle, 2007).

İki Faktör Model'de bir maddenin hem spesifik bir alt boyutla hem de genel bir boyutla ilişkili olduğu varsayılır. Ölçme modeli açısından İki faktör Model hem Üst Düzey Sıralı Model hem de ÇBMTK model ile birbirine benzerlik gösterse de modeller arasında önemli farklılıklar vardır. ÇBMTK modelde alt testler arasındaki korelasyon model kestirimlerinde serbest bırakılırken İki faktör modelde genel boyut ile alt testlerin birbirine dik olduğu varsayılmaktadır. Üst Düzey Sıralı ve ÇBMTK modeldeki maddeler genel boyut ile doğrudan ilişkili olmazken İki Faktör modelde maddeler genel boyut ile doğrudan ilişkilidir. Ayrıca İki Faktör modelin temel amacı genel boyuta ilişkin yeteneği kestirmektir. Alt testler ikincil çıktılar olarak kabul edilmektedir ve genel yeteneğin artıklarından açıklanmaktadır. Schmid ve Leiman (1957) tarafından İki Faktör model ile Üst Düzey Sıralı modelin belirli matematiksel sınırlamalar altında eşit olduğu belirtilmiştir.

Araştırmanın Amacı

Ülkemizde yapılan geniş ölçekli sınavlarda olduğu gibi PISA, TIMSS, PIRLS gibi uluslararası geniş ölçekli sınavlarda da çoğu zaman politikacıların, eğitimcilerin veya velilerin daha çok ülke performansı/sıralaması üzerinde durduğu ve testlerin içeriğindeki alt ölçeklerin, alt yeterlik alanlarının ya da alt konuların üzerinde çok fazla durulmadığı dikkat çekmektedir. Oysaki eğitim araştırmacıları ve uygulayıcıları için böyle uluslararası geniş ölçekli testlerin anlamlı daha küçük alt birimlerinden yola çıkarak yapılacak çok boyutlu analiz sonuçlarının potansiyel olarak öğrencilerin öğrenme şeklinin ve dolaysıyla öğrenme çıktılarının belirlenmesi açısından daha anlamlı ve otantik olacağı söylenebilir. Bu bağlamda araştırmanın temel amacı alt testlerden elde edilebilecek tanısal bilgilere dikkat çekmek ve güvenilir alt test puanı kestirimlerine katkı sağlamaktır.

Alt testlerden oluşan geniş ölçekli testlere ilişkin hem genel hem de alt test puanı kestirimlerine ülkemizde kullanılan klasik yöntemlerden farklı bir bakış açısı getirmesi bu çalışmanın en önemli özelliğidir. Bu çalışmada, Türkiye'de uygulanan geniş ölçekli sınavların test yapısını çok boyutlu ve hiyerarşik modeller çerçevesinde ele almanın ve analiz etmenin alt test ve toplam test puan kestirimlerinin doğruluğuna ve güvenirliğine etkisinin nasıl olacağı da araştırılmak istenmiştir. TEOG veri setinin özelliklerine göre güvenirliğe etki edebilecek hiyerarşik çok boyutlu MTK Model, İki faktör Model ve Üst Düzey Sıralı Model'in alt test ve toplam test puanı kestirimleri üzerindeki performansı incelenmiştir. Bu amaçla "alt test uzunluğu (20,30,40), alt test sayısı (2,3) ve alt testler arasındaki korelasyonların büyüklüğü (0.0, 0.3, 0.5, 0.8) koşulları altında üretilen veriler ile gerçek verilerin toplam test puanı ve alt test puanları kestirimlerinin doğruluğu ve güvenirlikleri Üst Düzey Sıralı Model, İki Faktör Model ve çok boyutlu hiyerarşik MTK modele göre nasıl değişmektedir?" sorusuna yanıt aranmıştır.

YÖNTEM

Deneysel desenler değişkenler arasındaki neden sonuç ilişkilerini test etmeyi amaçlayan araştırma desenleridir. Bu amacı gerçekleştirebilmek için Fraenkel, Wallen ve Hyun'a (2011, s.265-266) göre deneysel desenler, bağımsız değişken/lerin bağımlı değişken/ler üzerindeki etkisini incelemek için en az iki koşulun karşılaştırılmasını ve bağımsız değişkenin araştırmacı tarafından doğrudan değişimlenmesini (manipüle edilmesini) gerektirir. Ayrıca, araştırmacılar iç geçerliği korumak için dışsal değişkeni (ilgilenilmeyen ya da istenmeyen değişken) kontrol altına alarak bağımlı değişken üzerinde ölçme yapmalıdır (Kerlinger, 1973, s.300-313; Gall, Gall ve Borg, 2003, s.367-368). Simülasyon çalışmaları doğası gereği araştırmacılara bağımsız değişkenleri değişimleme ve dışsal değişkenleri kontrol altına alma imkânı sağlar. Bir bölümünde simülasyon verisi kullanılan bu araştırmada farklı test koşullarında üretilmiş verilerin toplam test puanı ve alt test puanları kestirimlerinin doğruluğu ve güvenirliği farklı modeller ve farklı test koşulları açısından karşılaştırıldığından çalışma simülatif verilerle yürütülen deneysel araştırma özelliği taşımaktadır. Ayrıca araştırmanın gerçek veri uygulaması içeren diğer bölümü, TEOG sınavı ile ilgili mevcut duruma ait bilgiler vereceğinden bu çalışmanın betimsel araştırma özelliği de bulunmaktadır.

Simülasyon Koşulları

Çalışma Grubu

Alan yazın çalışmalarından yetenek parametresi kestirimlerinde 1000 ve üzeri örneklemler arasında fark gözlenmediği (de a Torre ve Song, 2009; Yao ve Boughton, 2009) bulgusuna dayalı olarak bu araştırmada örneklem büyüklüğü bağımsız değişken olarak seçilmemiş ve araştırmanın verileri N=3000 olacak şekilde üretilmiştir.

Alt Test Sayısı

Alan yazındaki çok boyutlu ya da alt testlerden oluşan ölçme yapıları için parametre doğrulama çalışmaları incelendiğinde hem simülasyon hem de gerçek veri setleri üzerinde yapılan

araştırmalarda kullanılan ölçme araçlarının iki ila altı boyut arasında değişen alt boyutlarda/alt testlerde olduğu görülmüştür (Edwards ve Vevea, 2006; Lee, 2012; Yao, 2017).Ülkemizde yapılan çeşitli geniş ölçekli sınavlar incelendiğinde TEOG sınavının aynı oturumda yapılan sınavlarının üç toplamda altı, KPSS'de aynı oturumda yapılan çeşitli sınavların 2 (Genel Kültür ve Genel yetenek gibi) ya da beş (Çalışma ekonomisi, Ekonometri, İstatistik, Kamu Yönetimi ve Uluslararası ilişkiler gibi), ALES sınavının ise 2 (sayısal ve sözel) alt testten oluştuğu görülmüştür. Alan yazında yapılan araştırmalar ve ülkemizde yapılan geniş ölçekli sınavlar ve test uzunlukları göz önüne alındığında gerçek durumları temsil etmesi açısından bu araştırmada alt test sayısı iki ve üç olarak belirlenmiştir.

Alt Test Uzunluğu

Alan yazındaki çok boyutlu ya da alt testlerden oluşan ölçme yapıları için parametre doğrulama çalışmaları incelendiğinde hem simülasyon hem de gerçek veri setleri üzerinde yapılan araştırmalarda kullanılan ölçme araçlarının 5-60 madde arası alt test uzunluğuna sahip olduğu görülmektedir. Ülkemizde yapılan geniş ölçekli sınavlar incelendiğinde TEOG sınavlarının toplam altı alt testten ve 20'şer maddeden oluştuğu, ALES sınavının sayısal bölümünün 40'ar sorudan oluşan sayısal1 ve sayısal2 olarak iki at testten, sözel bölümünün ise 40'ar sorudan oluşan sözel1 ve sözel2 olarak iki alt testten oluştuğu belirlenmiştir. KPSS sınavları incelendiğinde aynı oturumda yapılan sınavlardan genel kültür ve genel yetenek alt testlerinin 60'ar sorudan, A grubu sınavlarından aynı oturumda yapılan sınavlardan Hukuk, İktisat, İşletme, Maliye ve Muhasebe alt testlerinin 30'ar sorudan ve Din Hizmetleri Alan Bilgisi Testi'nin (DHAB) ise DHAB1 ve DHAB2 alt testlerinin 20'şer sorudan oluştuğu görülmüştür. Alan yazın ve ülkemizde yapılan sınavlara dayalı olarak gerçek durumları temsil etmesi açısından bu araştırmada alt test uzunluğu 20, 30 ve 40 madde olarak belirlenmiştir.

Alt Testler Arasındaki Korelasyon

Alan yazındaki çok boyutlu ya da alt testlerden oluşan ölçme yapıları için parametre doğrulama çalışmaları incelendiğinde alt testler arasındaki korelasyonların parametre kestirimi üzerinde etkisinin olduğu belirtilmiştir (de la Torre ve Patz, 2005; Shin, 2007; Shin, Ansley, Tsai, ve Mao, 2005; Yao, 2010). Yine alan yazındaki araştırmalarda alt testler ya da alt testler arası korelasyon koşulu için 0.0-1.0 arasında değişen düzeylerde büyüklükler seçildiği görülmüştür. Ülkemizde 29 Nisan 2015'te yapılan TEOG sınavının altı alt testi arasındaki korelasyonların 0.0 civarında olması da göz önünde bulundurularak bu araştırma için alt testler arası korelasyon düzeyleri 0.0, 0.3, 0.5 ve 0.8 olarak belirlenmiştir.

Tablo 1: Simülasyon Koşulları ve Düzeyleri

Alt Test Sayısı	Al Test Uzunluğu	Alt Testler Arası Korelasyon
2	20	0.0
3	30	0.3
	40	0.5
		0.8

Tekrar (replikasyon) sayısı:50

Verilerin Üretilmesi

Araştırmada kullanılan veri setleri 29 Nisan 2015'te yapılan TEOG sınavının psikometrik özelliklerine dayalı olarak üretilmiştir. Bu sınav verisinin faktör yapısını belirlemek için Factor 10.3 (Lorenzo-Seva ve Ferrando, 2006) programında Ağırlıklandırılmamış En Küçük Kareler (ULS) yöntemi ve varimax döndürme tekniğine göre veriler analiz edilmiştir. Analiz sonucunda elde edilen

faktör yükü ve alt testler arası korelasyon matrislerinin incelenmesi sonucunda TEOG verisinin basit yapılı örtük yetenek konfigürasyonuna sahip olduğu görüldüğünden bu çalışmada veriler basit yapılı olacak şekilde üretilmiştir. TEOG 2015 sınavının 3PL modele göre elde edilen madde parametreleri ve alan yazında yer alan simülasyon çalışmaları göz önünde bulundurularak bu çalışmada kullanılan verilerin madde ayırt edicilik parametresi ranjı [0.8-3] arasında olacak şekilde ortalaması 1.5 ve varyansı 0.5 olan bir normal dağılımdan; güçlük parametresi ranjı [-2-2] arasında olacak şekilde ortalaması 0.0 ve varyansı 1.0 olan bir normal dağılımdan ve en düşük asimtot (şans) parametresi ise (6,16) olan bir beta dağılımdan üretilmiştir. Yetenek parametreleri çok değişkenli normal dağılıma θi ~ MVN (0, Σ) dayalı olarak ortalaması sıfır (0), varyansı ise araştırma koşullarında belirlenmiş olan varyans-kovaryans matrisine göre üretilmiştir. Üretilen madde ve yetenek parametreleri kullanılarak Tablo 1'de özetlenen koşullar altında iki kategorili 3000 kişilik veri setleri 50 tekrara dayalı olarak SimuMIRT (Yao, 2003) programı kullanılarak üretilmiştir. Harwell, Stone, Hsu ve Kirisci (1996) monte carlo simülasyon çalışmaları için optimal koşulları belirleme, mevcut programları inceleme ve simülasyon çalışmalarının kavramsallaştırılmasının önemini açıklama konusundaki çalışmalarında, simülasyon çalışmalarında en az 25 replikasyon kullanılması gerektiğini belirtmişlerdir. Bu konuda alan yazın incelendiğinde, Yao'nun (2010) ve Huang, Wang ve Chen (2013) 20 tekrar, Çakıcı Eser'in (2015) 25 tekrar ve de la Torre'nin (2009) 100 tekrar ile çalışmalarını yürüttükleri görülmüştür. Bu araştırmada ise tekrar sayısı 50 olarak belirlenmiştir.

Verilerin Analizi

Gerçek parametreler ile kestirilen parametrelerin karşılaştırılabilmesi için kestirilen parametreler ile gerçek değerlerin aynı ölçekte olması gerekir. Bunu sağlayabilmek için parametre kestirimlerinde popülasyon parametrelerinin onların gerçek değerine sabitlenmesi gerekir. Normalde gerçek değerleri bilinemez ama üretilen verilerin ortalama ve varyans-kovaryans matrisi, madde parametrelerinin dağılımları gerçek değer yerine kullanılabilir. Yao (2010) yetenek parametreleri kestirimlerinde madde parametrelerinin sabitlenmesi ile sabitlenmemesi yaklaşımı arasında bir fark olmadığını belirtse de bu araştırmada üretilen verilerin özellikleri önsel bilgi (prior) olarak kullanılmıştır. Ayrıca kestirim modelleri ve simülasyon koşullarına göre elde edilen RMSE değerleri ortalaması arasında anlamlı fark olup olmadığı varyans analizi ile test edilmiştir. Varyans analizi sonucunda en az iki grup arasında anlamlı farklılığın gözlendiği durumlar için farklılığın hangi gruplar arasında olduğunu belirlemek amacıyla çoklu karşılaştırma testi yapılmıştır. Karşılaştırma testi olarak Fisher'in LSD Testi kullanılmıştır. SPSS programı ara yüzü seçeneklerine göre yalnızca ana etkiler için karşılaştırma testi yapabilmesi nedeniyle etkileşimler için karşılaştırma testleri syntax yazılarak yapılmıştır.

Değerlendirme Kriteri

Araştırmada alt test puanlarının ve toplam test puan kestirimlerinin doğruluklarını değerlendirmek için RMSE (Ortalama hata kareler kökü) ve güvenirlik istatistikleri kullanılmıştır. RMSE değeri, gerçek parametre ile kestirilen parametreler arasındaki farkların ortalamasının karekökünü ifade etmektedir. Güvenirlik değeri ise gerçek parametre ile kestirilen parametreler arasındaki korelasyonun kareler ortalamasını ifade etmektedir. Bu istatistiklere ait matematiksel ifadeler aşağıdaki gibidir:

$$\text{RMSE}\left(\tau_j\right) = \sqrt{\frac{1}{n*N}\sum_{d=1}^{n}(\tau_j^* - \tau_j)^2} \qquad Güvenirlik = \frac{1}{n}\sum_{d=1}^{n} cor\,(\tau_j^*, \tau_j)^2$$

τ_j: j parametresinin gerçek değeri

τ_j^*: j parametresinin kestirilen değeri

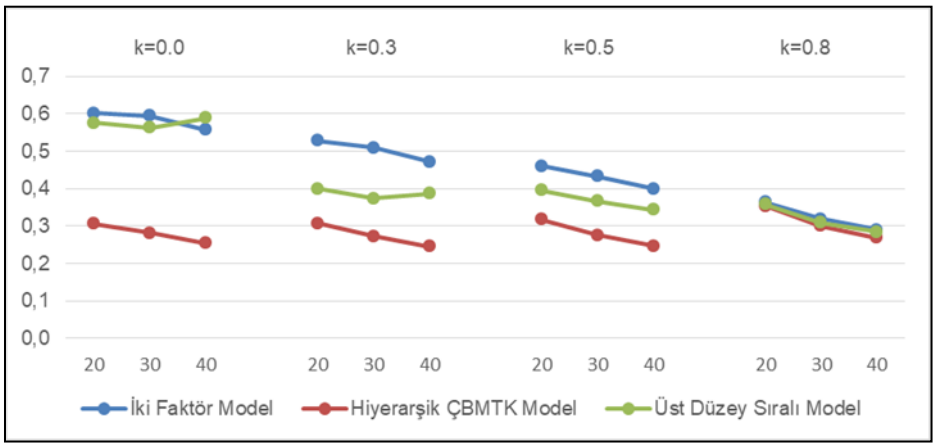
n: tekrar (replikasyon) sayısı

N: örneklem büyüklüğü

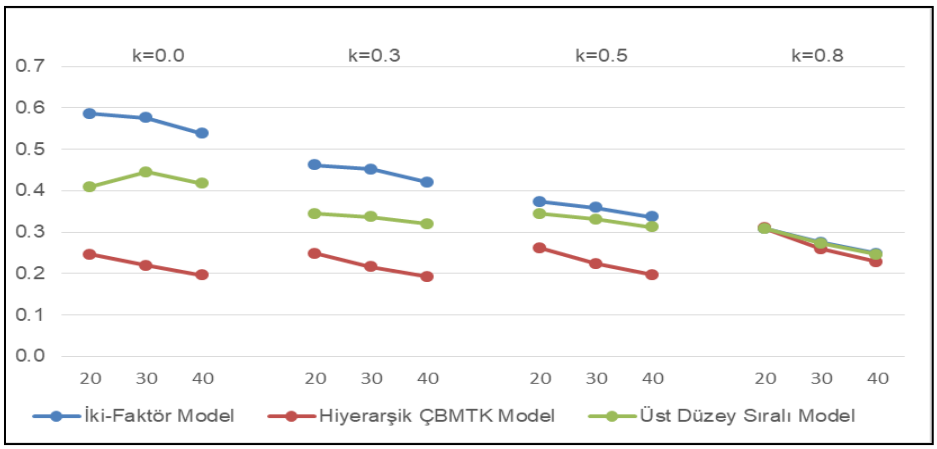
BULGULAR

Toplam Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait RMSE Değerlerine Yönelik Bulgular

Şekil 2 ve Şekil 3'te araştırmada ele alınan koşullara dayalı olarak sırasıyla iki ve üç boyutlu veri setlerinden toplam test puanı için üç hiyerarşik madde tepki kuramı modeli kullanılarak kestirilen yetenek parametrelerine ilişkin RMSE değerleri verilmiştir.



Şekil 2. İki Boyutlu Veri Setlerinden Toplam Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait RMSE Değerleri



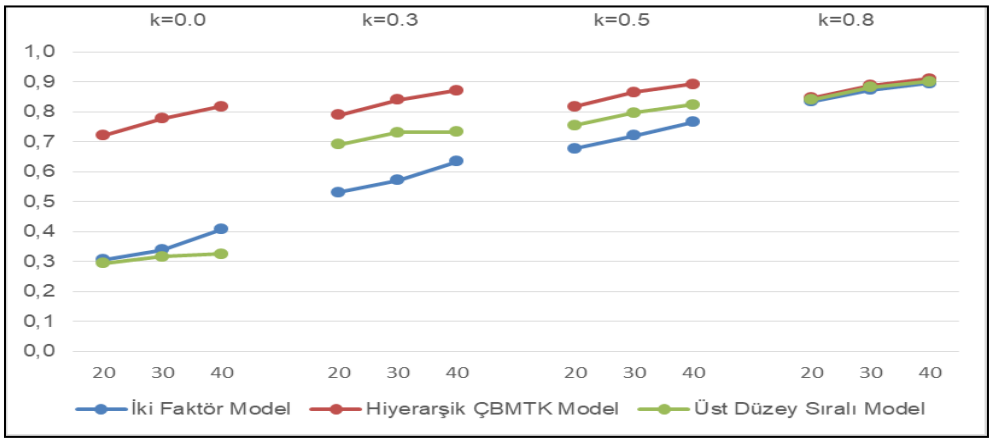
Şekil 3. Üç Boyutlu Veri Setlerinden Toplam Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait RMSE Değerleri

Şekil 2 ve Şekil 3'e göre üç kestirim modelinin tüm koşullar altındaki performansları karşılaştırıldığında, en düşük hata düzeyine sahip kestirimlerin İki Faktör modelinden elde edildiği görülürken en düşük hata düzeyine sahip kestirimlerin Hiyerarşik ÇBMTK modelden edildiği görülmektedir. Alt testler arası korelasyon düzeyi attıkça Hiyerarşik ÇBMTK dışında modellerin kestirim hatalarının genel olarak azaldığı ve üç modele ait sonuçların birbirine yaklaştığı gözlenmektedir. Özellikle alt testler arası korelasyon düzeyinin 0.8 olduğu durumda üç yöntemin benzer hatalar ile parametre kestirimi yaptığı söylenebilir. Alt test uzunluğundaki artışın üç yöntem için de genel olarak parametre kestirim hatasını azalttığı görülmektedir. İki-faktör Model'e ait kestirim hataları ile alt test uzunluğu ve alt testler arası korelasyon arasında azalan doğrusal yönde bir ilişki olduğu gözlenirken Üst Düzey Sıralı Model için bazı koşullarda değişken düzeyde ama çoğu koşul için benzer bir ilişki gözlenmektedir. Hiyerarşik ÇBMTK Model'de ise kestirim hataları ile alt test uzunluğu arasında azalan fakat alt testler arası korelasyon arasında artan doğrusal yönde

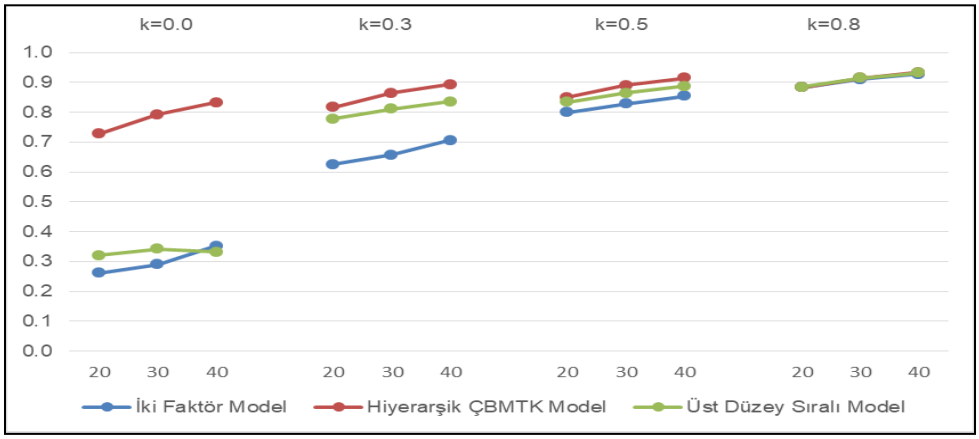
bir ilişki olduğu görülmektedir. İki ve üç boyutlu veri setlerine ait değerler birlikte değerlendirildiğinde alt test sayısındaki artışın üç modelin toplam test puanı için yetenek parametresi kestirim hatalarını azalttığı gözlenmektedir. İki faktör modeli için alt test sayısı ikiden üçe çıkarıldığında yetenek parametresi kestirim hatası alt testler arası korelasyon koşulunun her bir düzeyi için sırasıyla ortalama %3, %12, %17 ve %15 oranında azalmaya neden olurken bu durum Hiyerarşik ÇBMTK model için sırasıyla ortalama %22, %21, %19 ve %14 oranında, Üst Düzey Sıralı Model için sırasıyla %26, %14, %11 ve %13 oranında azalmaya neden olduğu görülmektedir.

Toplam Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait Güvenirlik Değerlerine Yönelik Bulgular

Şekil 4 ve Şekil 5'te araştırmada ele alınan koşullara dayalı olarak sırasıyla iki ve üç boyutlu veri setlerinden toplam test puanı için üç hiyerarşik madde tepki kuramı modeli kullanılarak kestirilen yetenek parametrelerine ilişkin güvenirlik değerleri verilmiştir.



Şekil 4. İki Boyutlu Veri Setlerinden Toplam Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait Güvenirlik Değerleri



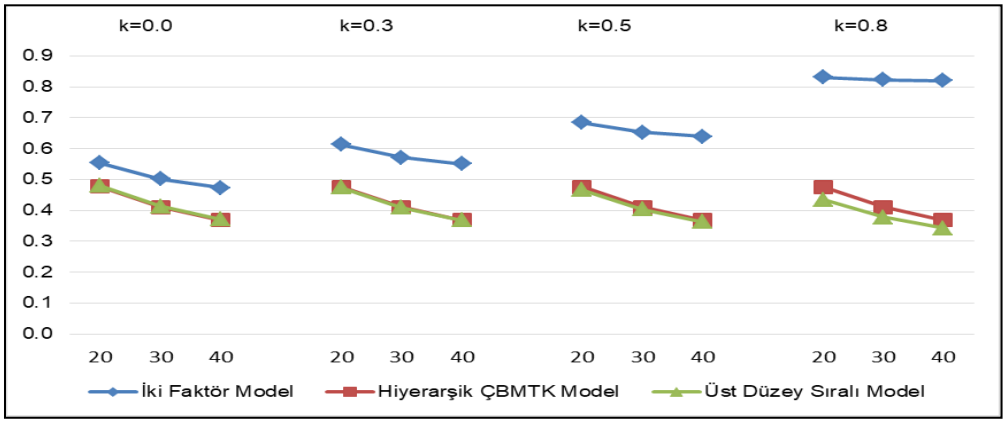
Şekil 5. Üç Boyutlu Veri Setlerinden Toplam Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait Güvenirlik Değerleri

Şekil 4 ve Şekil 5'e göre hem iki hem de üç boyutlu veri setleri için üç kestirim modelinin tüm koşullar altındaki performansları karşılaştırıldığında kabul edilebilir en güvenilir kestirimlerin Hiyerarşik ÇBMTK modelden edildiği görülmektedir. Alt testler arası korelasyon düzeyi attıkça

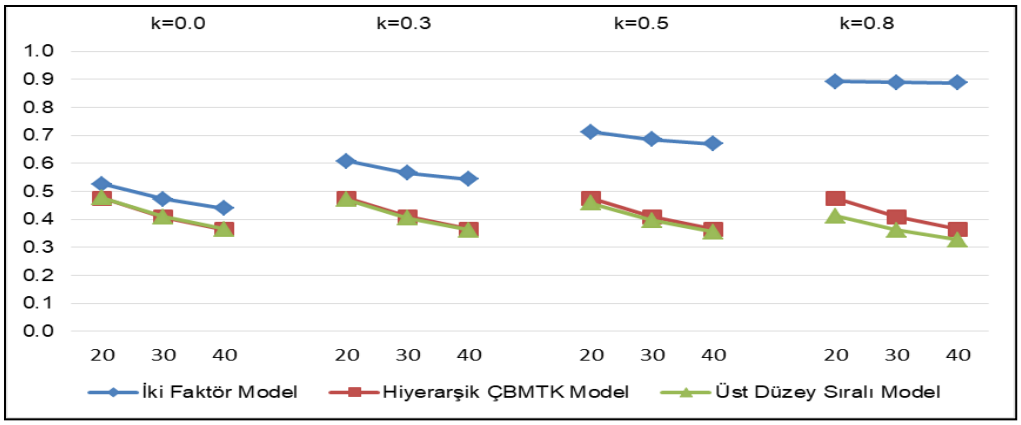
modellerin yetenek parametresi kestirim güvenirliğinin arttığı, Üst Düzey Sıralı Model kestirim güvenirliğinin korelasyonun 0.3 ve üzeri düzeylerde kabul edilebilir düzeylerde olduğu ve üç modelin kestirim güvenirliğinin korelasyonun 0.5 ve üzeri düzeylerde birbirine yaklaştığı gözlenmektedir. Özellikle alt testler arası korelasyon düzeyinin 0.8 olduğu durumda üç yöntemin benzer güvenirlik ile parametre kestirimi yaptığı söylenebilir. Alt test uzunluğu ve alt testler arası korelasyon düzeyindeki artışın her üç yöntem için de genel olarak parametre kestirim güvenirliğini arttırdığı görülmektedir. İki-faktör Model ve Üst Düzey Sıralı Model'in alt testler arası korelasyonun 0.0 düzeyinde yetenek parametre kestirimlerinin kabul edilemez düzeyde güvenirlik ile kestirildiği dikkat çekmektedir. Her üç modele ait kestirim güvenirliği ile alt test uzunluğu ve alt testler arası korelasyon arasında artan doğrusal yönde bir ilişki olduğu gözlenmektedir. İki ve üç boyutlu veri setlerine ait değerler birlikte değerlendirildiğinde boyut sayısındaki artışın üç modelin toplam test puanı için yetenek parametresi kestirim güvenirliğini koşulların çoğunda arttırdığı gözlenmektedir. İki faktör modelde alt test sayısındaki artışın kabul edilebilir düzeyde parametre kestirimi koşullarında iyileşme sağladığı görülmektedir. İki Faktör Model için alt test sayısını ikiden üçe çıkarmanın yetenek parametresi kestirim güvenirliğinin alt testler arası korelasyon düzeyinin 0.0 olduğu durum dışındaki diğer düzeyleri için sırasıyla ortalama %15, %15 ve %5 oranında artışa neden olduğu görülmektedir. Bu durum Hiyerarşik ÇBMTK model için alt testler arası korelasyon koşulunun tüm düzeylerinde ortalama %3 oranında artışa neden olurken Üst Düzey Sıralı Model için boyut sayısı artışı ile alt testler arası korelasyon koşulunun tüm düzeyleri için yetenek parametresi kestirim güvenirliğinde sırasıyla ortalama %7, %13, %9 ve %4 oranında artış olduğu gözlenmektedir.

Alt Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait RMSE Değerlerine Yönelik Bulgular

Şekil 6 ve Şekil 7'de araştırmada ele alınan koşullara dayalı olarak sırasıyla iki ve üç boyutlu veri setlerinden alt test puanı için üç hiyerarşik madde tepki kuramı modeli kullanılarak kestirilen yetenek parametrelerine ilişkin RMSE değerleri verilmiştir.



Şekil 6. İki Boyutlu Veri Setlerinden Alt Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait RMSE Değerleri



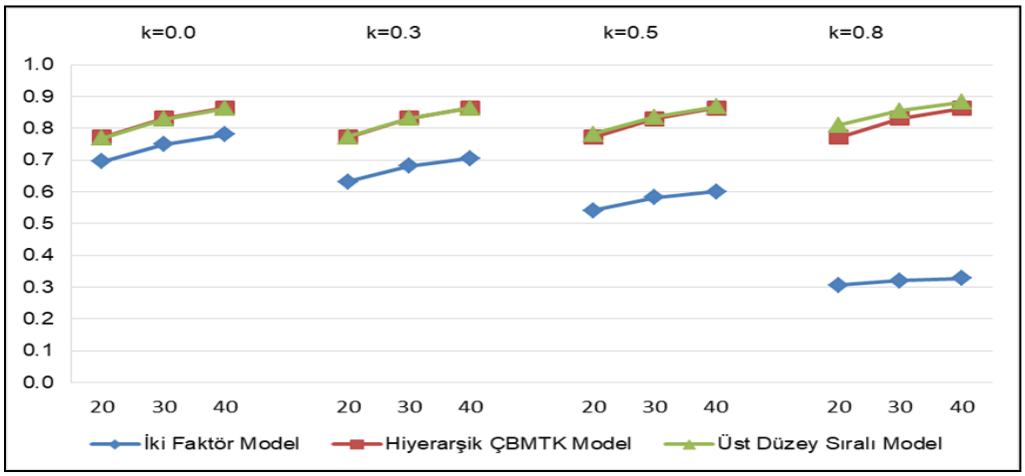
Şekil 7. Üç Boyutlu Veri Setlerinden Alt Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait RMSE Değerleri

Şekil 6 ve Şekil 7'ye göre üç kestirim modelinin tüm koşullar altındaki performansları karşılaştırıldığında, en düşük hata düzeyine sahip kestirimlerin İki Faktör modelinden elde edildiği görülürken en düşük hata düzeyine sahip kestirimlerin Üst Düzey Sıralı modelden edildiği görülmektedir. Alt testler arası korelasyon düzeyi attıkça Üst Düzey Sıralı Model ve Hiyerarşik ÇBMTK Model için alt test yetenek kestirim hatalarının azaldığı gözlenirken İki-faktör Model için hataların arttığı gözlenmektedir. Alt testler arası korelasyon koşulunun ilk iki düzeyinde Üst Düzey Sıralı Model ve Hiyerarşik ÇBMTK Model'in alt test yetenek kestirim hatalarının benzer olduğu görülürken korelasyon düzeyi arttıkça Üst Düzey Sıralı Model'in daha iyi performans gösterdiği görülmektedir. İki-faktör Model'e ait kestirim hataları ile alt test uzunluğu arasında azalan fakat alt testler arası korelasyon arasında artan doğrusal yönde bir ilişki olduğu gözlenirken Üst Düzey Sıralı Model'e ait kestirim hataları ile alt test uzunluğu ve alt testler arası korelasyon koşulları arasında azalan doğrusal yönde bir ilişki olduğu gözlenmektedir. Hiyerarşik ÇBMTK Model'e ait kestirim hataları ile korelasyon koşulu arasında bir ilişki olmadığı ama alt test uzunluğu ile kestirim hataları arasında azalan doğrusal yönde bir ilişki olduğu görülmektedir.

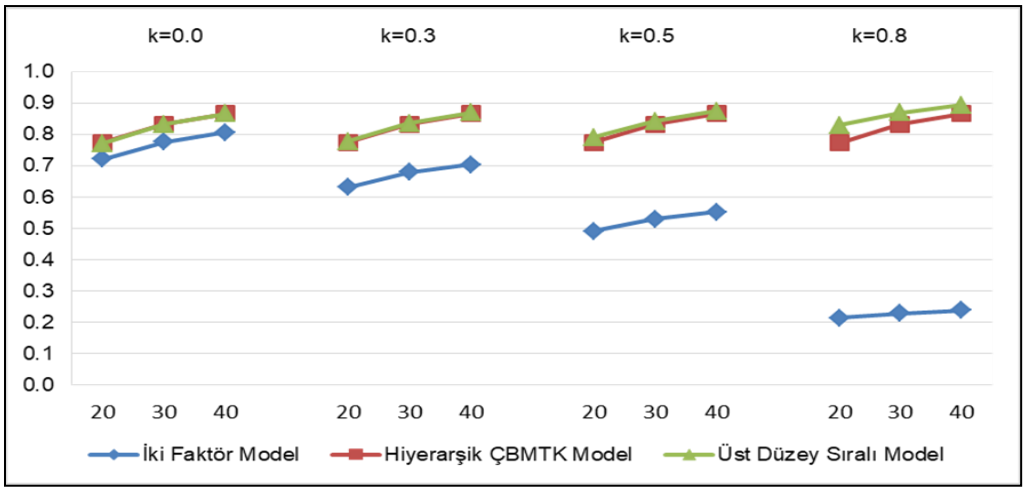
İki ve üç boyutlu veri setlerine ait değerler birlikte değerlendirildiğinde boyut sayısındaki artışın üç modelin alt test puanı için yetenek parametresi kestirim hataları üzerindeki etkisinin değişkenlik gösterdiği görülmektedir. İki-faktör modeli için boyut sayısı ikiden üçe çıkarıldığında yetenek parametresi kestirim hatası alt testler arası korelasyon koşulunun düşük düzeylerinde azalmaya neden olurken korelasyonun yüksek düzeylerinde artmaya neden olduğu görülmektedir. Bu model için boyut sayısı arttırıldığında alt testler arası korelasyonun 0.0 düzeyinde 20, 30 ve 40 maddelik alt testlerden elde edilen yetenek parametre kestirim hatalarının sırasıyla yaklaşık olarak %5, %6 ve %7 oranlarında azalmasına neden olurken korelasyonun 0.3 düzeyinde yaklaşık olarak %1 oranda azalmasına neden olmaktadır. Yine İki faktör model için boyut sayısı arttırıldığında alt testler arası korelasyonun 0.5 düzeyinde 20, 30 ve 40 maddelik alt testlerden elde edilen yetenek parametre kestirim hatalarının sırasıyla yaklaşık olarak %4, %5 ve %5 oranlarında artmasına neden olurken korelasyonun 0.8 düzeyinde sırasıyla yaklaşık olarak %1 oranda artmasına neden olmaktadır. Hiyerarşik ÇBMTK model için boyut sayısının ikiden üçe çıkarılmasının alt test yetenek kestirim hatalarında bir etkisi olmadığı görülmektedir. Üst Düzey Sıralı Model için boyut sayısı ikiden üçe çıkarıldığında alt testler arası korelasyon koşulunun düşük düzeylerinde minimal düzeyde olmakla birlikte alt test yetenek parametresi kestirim hatasının korelasyonun artışı ile azalmaya neden olduğu görülmektedir. Bu model için boyut sayısı arttırıldığında alt testler arası korelasyonun 0.0 ve 0.3 düzeylerinde 20, 30 ve 40 maddelik alt testlerden elde edilen yetenek parametre kestirim hatalarının yaklaşık olarak %0-%2 arası oranlarında azalmasına neden olurken korelasyonun 0.3 düzeyinde yaklaşık olarak %1-%2 arası oranlarda ve 0.8 düzeyinde yaklaşık olarak %4-%5 arası oranda azalmasına neden olmaktadır.

Alt Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait Güvenirlik Değerlerine Yönelik Bulgular

Şekil 8 ve Şekil 9'da araştırmada ele alınan koşullara dayalı olarak sırasıyla iki ve üç boyutlu veri setlerinden alt test puanı için üç hiyerarşik madde tepki kuramı modeli kullanılarak kestirilen yetenek parametrelerine ilişkin güvenirlik değerleri verilmiştir.



Şekil 8. İki Boyutlu Veri Setlerinden Toplam Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait Güvenirlik Değerleri



Şekil 9. Üç Boyutlu Veri Setlerinden Toplam Test Puanı İçin Kestirilen Yetenek Parametrelerine Ait Güvenirlik Değerleri

Şekil 8 ve Şekil 9'a göre üç kestirim modelinin tüm koşullar altındaki performansları karşılaştırıldığında kabul edilebilir en güvenilir kestirimlerin Üst Düzey Sıralı Modelden elde edildiği görülmektedir. Alt testler arası korelasyon düzeyi attıkça Üst Düzey Sıralı Model ve Hiyerarşik ÇBMTK Model'in yetenek parametresi kestirim güvenirliğinin arttığı ve kestirim güvenirliğinin korelasyonun tüm düzeylerinde kabul edilebilir düzeylerde olduğu gözlenmektedir. Ayrıca bu iki modelin alt test yetenek kestirim güvenirliğinin korelasyonun ilk üç düzeyinde aynı/benzer olduğu ve korelasyonun 0.8 düzeyinde Üst Düzey Sıralı Modelin daha güvenilir sonuçlar verdiği söylenebilir. İki faktör Modelin alt test yetenek kestirim güvenirliğinin korelasyon

düzeyi arttıkça önemli düzeyde azaldığı ve bu yöntemin kabul edilebilir düzeyde güvenilir kestirimlerinin düşük korelasyon düzeyinde elde edildiği gözlenmektedir. Fakat düşük korelasyon düzeyinde dahi İki faktör Model parametre kestirim güvenirliğinin Üst Düzey Sıralı Model ve Hiyerarşik ÇBMTK Model kestirimlerine göre daha düşük olduğu dikkat çekmektedir. Alt test uzunluğu ve alt testler arası korelasyon düzeyindeki artışın her üç yöntem için de parametre kestirim güvenirliğini arttırdığı görülmektedir. Alt test yetenek parametresi kestirim güvenirliği ile alt test uzunluğu ve alt testler korelasyon arasında Üst Düzey Sıralı Model ve Hiyerarşik ÇBMTK Model için artan doğrusal yönde bir ilişki olduğu gözlenirken İki faktör Model için güvenirlik ile alt test uzunluğu arasında artan fakat korelasyon ile azalan doğrusal yönde bir ilişki olduğu gözlenmektedir.

İki ve üç boyutlu veri setlerine ait değerler birlikte değerlendirildiğinde boyut sayısındaki artışın üç modelin alt test puanı için yetenek parametresi kestirim güvenirliğine etkisinin değişken olduğu gözlenmektedir. İki faktör modeli için boyut sayısını ikiden üçe çıkarmanın alt test kestirim güvenirliğini korelasyon düzeyinin 0.0 olduğu durumda yaklaşık olarak %4 oranında arttırdığı, korelasyonun 0.3 düzeyinde etkisinin olmadığı, korelasyonun 0.5 düzeyinde yaklaşık olarak %9 oranında azalttığı ve korelasyonun 0.8 düzeyinde yaklaşık olarak %28 oranında azalttığı görülmektedir. Hiyerarşik ÇBMTK model için boyut sayısının ikiden üçe çıkarılmasının alt test yetenek kestirim hatalarında bir etkisi olmadığı görülmektedir. Üst Düzey Sıralı Model için boyut sayısını ikiden üçe çıkarmanın alt test kestirim güvenirliğine korelasyon düzeyinin 0.0 düzeyinde etkisinin olmadığı, korelasyonun 0.3 ve 0.5 düzeyinde güvenirliği yaklaşık olarak %1 oranında arttırdığı, korelasyonun 0.8 düzeyinde güvenirliği yaklaşık olarak %2 oranında artırdığı görülmektedir.

Toplam Test Puanlarına ait RMSE ve Güvenirlik Değerleri İçin Varyans Analizi Sonuçları

Simülasyon çalışmasında tüm koşullar altında elde edilen toplam test puanlarına ait RMSE ve güvenirlik değerleri üzerinde model, alt test sayısı, alt test uzunluğu ve alt testler arasındaki korelasyonun etkisini incelemek için yapılan varyans analizi sonuçları sırasıyla Tablo 2 ve Tablo 3'te verilmiştir.

Tablo 2. Toplam Test Puanlarına Ait RMSE Değerleri İçin Varyans Analizi Sonuçları

Varyans kaynağı	Kareler toplamı	df	Kareler ortalaması	F	p	Kısmi η^2
Alt test sayısı	2,964	1	2,964	2239,519	0,000	0,388
Alt test uzunluğu	1,482	2	0,741	559,900	0,000	0,241
Korelasyon	10,010	3	3,337	2521,279	0,000	0,682
Model	19,153	2	9,576	7236,319	0,000	0,804
Alt test sayısı * alt test uzunluğu	0,016	2	0,008	6,104	0,002	0,003
Alt test sayısı * korelasyon	0,127	3	0,042	31,884	0,000	0,026
Alt test sayısı * model	0,066	2	0,033	24,815	0,000	0,014
Alt test uzunluğu * korelasyon	0,167	6	0,028	21,091	0,000	0,035
Alt test uzunluğu * model	0,136	4	0,034	25,756	0,000	0,028
Korelasyon * model	8,567	6	1,428	1078,886	0,000	0,647
Alt test sayısı * Alt test uzunluğu * korelasyon	0,009	6	0,002	1,196	0,305	0,002
Alt test sayısı * Alt test uzunluğu * model	0,018	4	0,004	3,358	0,009	0,004
Alt test sayısı * korelasyon * model	0,687	6	0,114	86,494	0,000	0,128
Alt test uzunluğu * korelasyon * model	0,056	12	0,005	3,501	0,000	0,012
Alt test sayısı * Alt test uzunluğu * korelasyon * model	0,016	12	0,001	0,999	0,447	0,003

Tablo 2 incelendiğinde toplam test puanlarına ait RMSE değerleri üzerinde tüm ana ve tüm ikili ortak etkilerin anlamlı düzeyde etkisi olduğu gözlenmektedir. Ana etkiler açısından etki büyüklükleri incelediğinde en fazla etkiye sahip değişkenlerin sırasıyla model (kısmi η^2 =0.804) ve alt testler arası korelasyon (kısmi η^2 =0.682) olduğu görülmektedir. İkili ortak etkiler incelendiğinde toplam testlere ait RMSE değerlerinin varyansını en fazla açıklayan etkileşimin korelasyon*model (kısmi η^2 =0.647) olduğu ve diğer ikili etkileşimlerin etki büyüklerinin (kısmi η^2 ≤0.035) çok düşük olduğu gözlenmektedir. Üçlü ortak etkiler içerisinde en fazla etkiye sahip etkileşimin alt test uzunluğu*korelasyon*model (kısmi η^2 =0.128) olduğu görülürken diğer üçlü etkileşimlerin etkisinin ya olmadığı (alt test sayısı*alt test uzunluğu*korelasyon etkileşimi, p=0.305) ya da çok düşük (kısmi η^2 ≤0.012) olduğu görülmektedir. Dörtlü ortak etkinin RMSE değerlerinin varyansına anlamlı bir katkısı olmadığı görülmektedir (p=0.447).

Çoklu karşılaştırma testi sonucunda ana etkiler için koşulların tüm düzeyleri arasında anlamlı farklılık olduğu bulunmuştur. Korelasyon, alt test uzunluğu ve alt test sayısı koşullarının düzeyleri arttıkça hatanın azaldığı gözlenmiştir. Modeller açısından en az hatalı kestirim yapan modeller sırasıyla Hiyerarşik ÇBMTK, Üst Düzey Sıralı ve İki Faktör modeldir. Korelasyon*model ikili etkileşimin modellere göre ikili karşılaştırma testi sonucunda yalnızca 0.8 korelasyon düzeyi için İki Faktör Model ile Üst Düzey Sıralı Model arasında anlamlı farklılık gözlenmezken diğer tüm ikili karşılaştırmalar anlamlı bulunmuştur. Yine, korelasyon*model etkileşimin korelasyona göre ikili karşılaştırma testi sonucunda ise İki Faktör ve Üst Düzey Sıralı modeller için korelasyonun tüm ikili etkileşimleri arasında anlamlı farklılık gözlenirken Hiyerarşik ÇBMTK model için yalnızca 0.8 korelasyon ile diğer korelasyon düzeyleri ve 0.5 korelasyon ile 0.3 korelasyon düzeyi arasında anlamlı farklılık gözlenmiştir.

Tablo 3. Toplam Test Puanlarına Ait Güvenirlik Değerleri İçin Varyans Analizi Sonuçları

Varyans kaynağı	Kareler toplamı	df	Kareler ortalaması	F	p	Kısmi η^2
Alt test sayısı	1,509	1	1,509	497,963	0,000	0,124
Alt test uzunluğu	2,721	2	1,361	448,974	0,000	0,203
Korelasyon	90,152	3	30,051	9915,606	0,000	0,894
Model	24,928	2	12,464	4112,629	0,000	0,700
Alt test sayısı * alt test uzunluğu	0,020	2	0,010	3,371	0,034	0,002
Alt test sayısı * korelasyon	0,806	3	0,269	88,663	0,000	0,070
Alt test sayısı * model	0,154	2	0,077	25,379	0,000	0,014
Alt test uzunluğu * korelasyon	0,038	6	0,006	2,117	0,048	0,004
Alt test uzunluğu * model	0,162	4	0,040	13,349	0,000	0,015
Korelasyon * model	25,631	6	4,272	1409,532	0,000	0,706
Alt test sayısı * Alt test uzunluğu * korelasyon	0,009	6	0,002	0,499	0,810	0,001
Alt test sayısı * Alt test uzunluğu * model	0,008	4	0,002	0,637	0,636	0,001
Alt test sayısı * korelasyon * model	0,497	6	0,083	27,322	0,000	0,044
Alt test uzunluğu * korelasyon * model	0,140	12	0,012	3,842	0,000	0,013
Alt test sayısı * Alt test uzunluğu * korelasyon * model	0,015	12	0,001	0,400	0,964	0,001

Tablo 3 incelendiğinde toplam test puanlarına ait güvenirlik değerleri üzerinde tüm ana ve tüm ikili ortak etkilerin anlamlı düzeyde etkisi olduğu gözlenmektedir. Ana etkiler açısından etki büyüklükleri incelediğinde en fazla etkiye sahip değişkenlerin sırasıyla alt testler arası korelasyon (kısmi η^2

=0.894) ve model (kısmi η^2=0.700) olduğu görülmektedir. İkili ortak etkiler incelendiğinde toplam testlere ait güvenirlik değerlerinin varyansını en fazla açıklayan etkileşimin korelasyon*model (kısmi η^2=0.706) olduğu ve diğer ikili etkileşimlerin etki büyüklerinin (kısmi η^2≤0.070) çok düşük olduğu gözlenmektedir. Üçlü ortak etkiler içerisinde yalnızca alt test sayısı*korelasyon*model ve alt test uzunluğu*korelasyon*model ortak etkileşimlerinin anlamlı fakat çok düşük düzeyde (kısmi η^2 ≤0.044) etkiye sahip olduğu görülmektedir. Dörtlü ortak etkinin güvenirlik değerlerinin varyansına anlamlı bir katkısı olmadığı görülmektedir (p=0.964). Çoklu karşılaştırma testi sonucunda ana etkiler için koşulların tüm düzeyleri arasında anlamlı farklılık olduğu bulunmuştur. Korelasyon, alt test uzunluğu ve alt test sayısı koşullarının düzeyleri arttıkça güvenirliğin arttığı gözlenmiştir. Modeller açısından en güvenilir kestirim yapan modeller sırasıyla Hiyerarşik ÇBMTK, Üst Düzey Sıralı ve İki Faktör modeldir.

Alt Test Puanlarına ait RMSE ve Güvenirlik Değerleri İçin Varyans Analizi Sonuçları

Simülasyon çalışmasında tüm koşullar altında elde edilen alt test puanlarına ait RMSE ve güvenirlik değerleri üzerinde model, alt test sayısı, alt test uzunluğu ve alt testler arasındaki korelasyonun etkisini incelemek için yapılan varyans analizi sonuçları sırasıyla Tablo 4 ve Tablo 5'te verilmiştir.

Tablo 4. Alt Test Puanlarına Ait RMSE Değerleri İçin Varyans Analizi Sonuçları

Varyans kaynağı	Kareler toplamı	df	Kareler ortalaması	F	p	Kısmi η^2
Alt test sayısı	0,002	1	,002	104,263	0,000	,029
Alt test uzunluğu	4,617	2	2,309	114331,070	0,000	,985
Korelasyon	5,680	3	1,893	93761,951	0,000	,988
Model	45,951	2	22,976	1137853,566	0,000	,998
Alt test sayısı * alt test uzunluğu	,000	2	,000	11,190	0,000	,006
Alt test sayısı * korelasyon	,101	3	,034	1662,740	0,000	,586
Alt test sayısı * model	,091	2	,045	2245,775	0,000	,560
Alt test uzunluğu * korelasyon	,092	6	,015	760,861	0,000	,564
Alt test uzunluğu * model	,425	4	,106	5257,361	0,000	,856
Korelasyon * model	16,664	6	2,777	137549,079	0,000	,996
Alt test sayısı * Alt test uzunluğu * korelasyon	0,001	6	,000	9,986	0,000	,017
Alt test sayısı * Alt test uzunluğu * model	0,000	4	,000	5,141	0,000	,006
Alt test sayısı * korelasyon * model	0,308	6	,051	2538,759	0,000	,812
Alt test uzunluğu * korelasyon * model	0,088	12	,007	364,595	0,000	,554
Alt test sayısı * Alt test uzunluğu * korelasyon * model	0,001	12	,000	4,206	0,000	,014

Tablo 4 incelendiğinde alt test puanlarına ait RMSE değerleri üzerinde tüm ana ve tüm ikili ortak etkilerin anlamlı düzeyde etkisi olduğu gözlenmektedir. Ana etkiler açısından etki büyüklükleri incelediğinde alt test puanlarına ait RMSE değerleri üzerinde alt test sayısı değişkeni (kısmi η^2 =0.029) dışındaki diğer değişkenlerin yüksek düzeyde (kısmi η^2≥0.985) etkiye sahip olduğu görülmektedir. İkili ortak etkiler incelendiğinde alt testlere ait RMSE değerlerinin varyansını en fazla açıklayan etkileşimin sırasıyla alt test uzunluğu* model (kısmi η^2=0.996) ve korelasyon*model (kısmi η^2=0.856) olduğu; en düşük etkiye sahip etkileşimin ise alt test sayısı*alt test uzunluğu (kısmi η^2=0.006) olduğu gözlenmektedir. Üçlü ortak etkiler içerisinde en fazla etkiye sahip etkileşimlerin sırasıyla alt test sayısı*korelasyon*model (kısmi η^2=0.812) ve alt test uzunluğu*korelasyon*model (kısmi η^2=0.554) olduğu görülürken diğer üçlü etkileşimlerin etkisinin çok düşük (kısmi η^2≤0.017) olduğu görülmektedir. Dörtlü etkilerin RMSE değerlerinin varyansına

katkısının çok düşük (kısmi η^2=0.014) olduğu görülmektedir. Çoklu karşılaştırma testi sonucunda ana etkiler için koşulların tüm düzeyleri arasında anlamlı farklılık olduğu bulunmuştur. Korelasyon arttıkça hatanın arttığı fakat alt test uzunluğu arttıkça hatanın azaldığı gözlenmiştir. İki ve üç boyutlu alt testlerden elde edilen alt test puanı kestirim hataları arasında bir fark olmadığı görülmüştür. Her bir boyut için elde edilen RMSE ortalaması sırasıyla 0,490 ve 0,491'dir. Varyans analinizde gözlenen anlamlı etkinin örneklem büyüklüğünden kaynaklandığı düşünülmektedir. Modeller açısından en az hatalı kestirim yapan modeller sırasıyla Üst Düzey Sıralı, Hiyerarşik ÇBMTK ve İki Faktör modeldir.

Tablo 5. Alt Test Puanlarına Ait Güvenirlik Değerleri İçin Varyans Analizi Sonuçları

Varyans kaynağı	Kareler toplamı	df	Kareler ortalaması	F	p	Kısmi η^2
Alt test sayısı	0,037	1	0,037	1487,220	0,000	0,297
Alt test uzunluğu	3,835	2	1,918	76934,827	0,000	0,978
Korelasyon	11,430	3	3,810	152859,783	0,000	0,992
Model	56,748	2	28,374	1138361,878	0,000	0,998
Alt test sayısı * alt test uzunluğu	0,000	2	0,000	2,827	0,059	0,002
Alt test sayısı * korelasyon	0,163	3	0,054	2175,031	0,000	0,649
Alt test sayısı * model	0,235	2	0,118	4724,061	0,000	0,728
Alt test uzunluğu * korelasyon	0,080	6	0,013	532,873	0,000	0,475
Alt test uzunluğu * model	0,115	4	0,029	1152,489	0,000	0,566
Korelasyon * model	28,721	6	4,787	192049,700	0,000	0,997
Alt test sayısı * Alt test uzunluğu * korelasyon	0,000	6	0,000	1,993	0,063	0,003
Alt test sayısı * Alt test uzunluğu * model	0,001	4	0,000	5,515	0,000	0,006
Alt test sayısı * korelasyon * model	0,450	6	0,075	3008,391	0,000	0,837
Alt test uzunluğu * korelasyon * model	0,059	12	0,005	197,838	0,000	0,402
Alt test sayısı * Alt test uzunluğu * korelasyon * model	0,001	12	0,000	2,238	0,008	0,008

Tablo 5 incelendiğinde alt test puanlarına ait güvenirlik değerleri üzerinde tüm ana etkilerin anlamlı düzeyde etkisi olduğu gözlenmektedir. Ana etkiler açısından etki büyüklükleri incelediğinde alt test puanlarına ait güvenirlik değerleri üzerinde alt test sayısı değişkeni (kısmi η^2=0.297) dışındaki diğer değişkenlerin yüksek düzeyde (kısmi η^2≥0.978) etkiye sahip olduğu görülmektedir. İkili ortak etkiler incelendiğinde alt testlere ait güvenirlik değerlerinin varyansına alt test sayısı*alt test uzunluğu etkileşiminin katkı sağlamadığı görülürken (p=0.059) en fazla katkı sağlayan etkileşimin korelasyon*model (kısmi η^2=0.997) olduğu ve diğer ikili etkileşimlerin etki büyüklerinin (kısmi η^2 ≥0.475) en az orta düzeyde olduğu görülmektedir. Üçlü ortak etkiler içerisinde yalnızca alt test sayısı*alt test uzunluğu*korelasyon etkileşimin anlamlı etkisinin olmadığı gözlenirken en fazla etkiye sahip üçlü ortak etkileşimin alt test sayısı*korelasyon*model (kısmi η^2=0.837) olduğu, en az etkiye sahip etkileşimin ise alt test sayısı*alt test uzunluğu*model (kısmi η^2=0.006) olduğu görülmektedir. Dörtlü etkinin alt testlere ait güvenirlik değerlerinin varyansına katkısının çok düşük düzeyde (kısmi η^2=0.997) olduğu görülmektedir. Çoklu karşılaştırma testi sonucunda ana etkiler için koşulların tüm düzeyleri arasında anlamlı farklılık olduğu bulunmuştur. Korelasyon arttıkça güvenirliğin azaldığı fakat alt test uzunluğu ve alt test sayısı arttıkça güvenirliğin arttığı gözlenmiştir. Modeller açısından en güvenilir kestirim yapan modeller sırasıyla Üst Düzey Sıralı, Hiyerarşik ÇBMTK ve İki Faktör modeldir.

Gerçek Veri Uygulamasına İlişkin Sonuçlar

Araştırmanın kapsamında "TEOG (2015) verilerinin Üst Düzey Sıralı (Higher Order), İki Faktör (Bi-factor) ve hiyerarşik çok boyutlu madde tepki kuramı modellerine göre alt test puan kestirimlerinin nasıl değiştiği" modellerin kestirdiği sonsal dağılımın ortalama ve standart sapması ve alt testler arası korelasyon değerleri ve standart sapması ile incelenmiştir. TEOG 2015 verisinin üç kestirim modeline göre her bir alt test için kestirilen sonsal dağılımın ortalama ve standart sapma değerleri Tablo 6'da verilirken kestirilen alt testler arası korelasyon matrisi ise Tablo 7'de verilmiştir.

Tablo 6. Sonsal Dağılımın Ortalama ve Standart Sapması

Alt testler	İki Faktör Model		Hiyerarşik ÇBMTK Model		Üst Düzey Sıralı Model	
	Ortalama	Std. Sapma	Ortalama	Std. Sapma	Ortalama	Std. Sapma
Din Kültürü	0.004	0.009	0.009	0.012	0.009	0.013
Fen Bilgisi	-0.008	0.014	-0.004	0.010	-0.001	0.009
İngilizce	-0.017	0.021	-0.026	0.028	-0.024	0.026
Matematik	-0.015	0.017	-0.017	0.019	-0.014	0.017
Tarih	-0.005	0.009	-0.005	0.009	-0.002	0.008
Türkçe	0.002	0.008	0.005	0.013	0.000	0.008

Tablo 7. Modellerin Kestirdiği ve TEOG Verisinin Alt Testler Arası Korelasyon Matrisi

Model	Alt testler	Din Kültürü	Fen Bilgisi	İngilizce	Matematik	Tarih	Türkçe
İki Faktör	Fen Bilgisi	-0.032	1	0.001	-0.009	-0.012	-0.007
	İngilizce	-0.037	0.001	1	-0.005	-0.018	-0.009
	Matematik	-0.036	-0.009	-0.005	1	0.000	-0.009
	Tarih	-0.037	-0.012	-0.018	0.000	1	0.006
	Türkçe	-0.041	-0.007	-0.009	-0.009	0.006	1
ÇBMTK	Fen Bilgisi	0.009	1	0.006	-0.007	-0.006	-0.005
	İngilizce	-0.008	0.006	1	-0.004	-0.012	-0.006
	Matematik	-0.005	-0.007	-0.004	1	0.005	-0.005
	Tarih	0.012	-0.006	-0.012	0.005	1	0.010
	Türkçe	-0.007	-0.005	-0.006	-0.005	0.010	1
Üst Düzey	Fen Bilgisi	0.008	1	0.006	-0.005	-0.005	-0.003
	İngilizce	-0.008	0.006	1	-0.004	-0.015	-0.010
	Matematik	-0.004	-0.005	-0.004	1	0.007	-0.007
	Tarih	0.015	-0.005	-0.015	0.007	1	0.013
	Türkçe	-0.006	-0.003	-0.010	-0.007	0.013	1
TEOG	Fen Bilgisi	-0.002	1	0.006	-0.005	-0.005	-0.003
	İngilizce	-0.018	0.013	1	-0.004	-0.015	-0.010
	Matematik	0.013	-0.011	-0.007	1	0.007	-0.007
	Tarih	0.017	0.011	-0.009	0.000	1	0.013
	Türkçe	0.040	-0.001	0.006	-0.081	0.010	1

Tablo 6 incelendiğinde İki Faktör modelin Hiyerarşik ÇBMTK modele göre Din Kültürü, İngilizce, Matematik ve Türkçe alt testlerinde Üst Düzey Sıralı modele göre ise Din Kültürü ve İngilizce alt testlerinde daha düşük sonsal standart sapmaya sahip olduğu görülmektedir. Üst Düzey Sıralı

modelin Hiyerarşik ÇBMTK modele göre Fen Bilgisi, İngilizce, Matematik, Tarih ve Türkçe alt testlerinde İki Faktör modele göre ise Fen Bilgisi ve Tarih alt testlerinde daha düşük sonsal standart sapmaya sahip olduğu görülmektedir. Hiyerarşik ÇBMTK modelin İki Faktör modele göre yalnızca Fen Bilgisi alt testinde ise Üst Düzey Sıralı modele göre ise yalnızca Din Kültürü alt testinde daha düşük sonsal standart sapmaya sahip olduğu görülmektedir. Tüm yöntemlerin bütün alt testler açısından sonsal dağılım ortalama ve standart sapmaları bir arada değerlendirildiğinde üç yönteminde genel olarak düşük sonsal standart sapmaya sahip olduğu fakat İki Faktör modelin diğer yöntemlere göre az farkla daha iyi sonuç verdiği söylenebilir.

Tablo 7'de üç model tarafından kestirilen alt testler arası korelasyon matrisi incelendiğinde üç modelin de gerçek veri matrisinin korelasyon matrisine çok benzer kestirimler yaptığı görülmektedir. Tablo 7'deki korelasyon matrisinin standart sapma değerleri ise EK-1'de verilmiştir. Standart sapma değerleriyle birlikte korelasyon değerleri incelendiğinde Hiyerarşik ÇBMTK model ile Üst Düzey Sıralı modelin birbirine çok benzer sonuçlar verdiği ve çok az farkla İki Faktör modelin diğer modellere göre standart sapma değerlerinin yüksek olduğu söylenebilir. Gerçek veri setinden elde edilen bu sonucun simülasyon çalışmasının 0.0 korelasyon ve 20 maddelik alt testlerden elde edilen sonuçları ile uyumludur.

SONUÇLAR ve TARTIŞMA

Araştırmada öncelikle iki ve üç boyutlu Hiyerarşik ÇBMTK'ya göre üretilen verilerin aynı modele dayalı elde edilen toplam ve alt test yetenek parametresi kestirimlerinin belirli bir hata düzeyinde elde edildiği gözlenmiştir. Simülasyon çalışmalarında verilerin belirli bir hata düzeyinde üretilmesi beklenen bir durumdur. RMSE istatistiği için belirli bir sınır olmadığından yalnızca daha düşük değerlerin daha iyi olduğu belirtilmektedir. Parametre doğrulama çalışmalarında yetenek parametresine ait hataların madde parametresine ait hatalardan daha yüksek elde edildiği görülmektedir (Çakıcı Eser, 2014; Jiang, Wang ve Weiss,2016; Lee, 2012). Ayrıca de la Torre ve Patz (2005) ile Yao'nun (2010) çok boyutlu MTK'ya dayalı ürettikleri verileri ve de la Torre, Song ve Hong'un (2011) Üst Düzey Sıralı Modele dayalı ürettikleri verileri aynı model ile analiz etmeleri sonucu yetenek parametre kestirimlerinde bu araştırma ile benzer düzeyde hataların elde edildiği görülmüştür. Bu durumun veri üretilirken yetenek parametresinin geniş bir normal dağılımdan gelmesi ve az sayıda madde örneklemi ile birey yeteneğinin kestirilmesinden kaynaklandığı düşünülmektedir.

Hem iki hem de üç boyutlu veri setlerinde alt testler arasındaki korelasyon düzeyinin artmasıyla Hiyerarşik ÇBMTK Model'den toplam test puanı için elde edilen hataların artmasının nedeni olarak testin boyutluluk derecesindeki azalma gösterilebilir. Bir başka ifadeyle alt testler arasındaki yüksek düzeyde ilişkiler testin tek boyutluluğa yakınlaşmasına neden olmaktadır. Aynı nedenle, İki Faktör Model'de toplam test puanı kestirimlerinde alt testler arasındaki korelasyon düzeyi arttıkça hataların azaldığı görülmektedir. Çünkü İki Faktör Model'de toplam test puanı testteki tüm maddelerin kalibrasyonundan elde edilir. Korelasyon düzeyindeki artışı ile Üst Düzey Sıralı Model'de toplam test puanı kestirimlerinde hataların azalmasının nedeni ise modelde kullanılan regresyon katsayılarının alt testler arası ilişkilerden türetilmesidir. Yukarıdaki sayılan benzer nedenler ile hem iki hem üç boyutlu verilerde İki Faktör ve Üst Düzey Sıralı Model için korelasyon arttıkça toplam test güvenirliği artmaktadır. Ayrıca alt testler arası korelasyonlar yüksek olsa da yapı modelinin maddeleri tek bir boyutla ilişkilendirmesi ve verilerin yine aynı modelle üretilmesi nedeniyle tüm koşullarda Hiyerarşik ÇBMTK Model en düşük hata ve en yüksek güvenilirlik düzeyinde sonuçlar vermektedir. Daha uzun alt testlerde üç yöntem için de kestirim hatalarının azalması ve güvenirliğin artması beklenen bir durumdur.

Hiyerarşik ÇBMTK Model'in toplam test puanı kestirimlerinde maksimum bilgi yöntemini kullanması nedeniyle alt test sayısındaki artışın yetenek parametre kestirimleri üzerinde en fazla katkı sağladığı model bu modeldir. Alt test sayısındaki artış toplam madde sayısını arttırdığı için İki Faktör Model'de toplam test yetenek kestirimlerine katkısı alt testler arası korelasyon arttıkça daha fazla artmaktadır. Üst Düzey sıralı modelde ise genel yetenek kestirimde kullanılan birinci düzey

değişken sayısının artması nedeniyle alt test sayısındaki artış kestirim hatalarını azaltmakta ve güvenirliği arttırmaktadır.

Hem iki hem de üç boyutlu veri setlerinde alt testler arasındaki korelasyon düzeyinin Hiyerarşik ÇBMTK Model'den elde edilen alt test puan kestirim hataları ve güvenirliği üzerinde bir etkisi olmadığı sonucuna ulaşılmıştır. Fakat Bulut (2013) ve Yao (2010) çalışmalarında alt testler arası korelasyon düzeyindeki artışın ÇBMTK Model kestirimlerinin güvenirliği ve doğruluğunu arttırdığını bulmuşlardır. Alt test sayısındaki artışın bu çalışmada kestirim hataları ve güvenirliği üzerinde bir etkisi gözlenmezken Bulut (2013) alt test sayısının güvenirlik üzerinde minimal düzeyde etkisi olduğunu belirtmiştir. Bu çalışma ile diğer çalışmalar arasındaki farkın veri üretme koşulları arasındaki farklılıktan kaynaklanabileceği düşünülmektedir.

Hem iki hem de üç boyutlu veri setlerinde alt testler arasındaki korelasyon düzeyindeki artış ile İki Faktör Model'den elde edilen alt test puan kestirim hatalarının arttığı ve güvenirliğin azaldığı sonucu Yao (2010) ve Chang'nın (2015) araştırma sonuçlarıyla uyumludur. Bu modelde alt testler arasındaki ilişkilerin dik olduğu varsayımı nedeniyle model yüksek ilişki gösteren alt testlerde yüksek hatalı ve düşük güvenirlikli kestirimler yapmaktadır. Dolayısıyla kabul edilebilir düzeyde güvenilir kestirimlerin ancak 0.0 korelasyon düzeyinde ve 0.3 korelasyon düzeyinin de uzun alt test düzeylerinde elde edilmektedir. Ayrıca alt test sayısındaki artışın bu model için optimal koşul olan düşük korelasyon düzeylerinde hataların azalması ve güvenirliğin artması benzer nedenlerden kaynaklanmakladır. Optimal koşullardan uzaklaştıkça alt test sayısını arttırmak daha düşük güvenirlikli ve daha yüksek hatalı kestirimlere sebep olmaktadır.

Hem iki hem de üç boyutlu veri setlerinde alt testler arasındaki korelasyon düzeyindeki artış ile Üst Düzey Sıralı Model'den elde edilen alt test puanı kestirim hataların azalması modelin doğası gereği genel ve alt boyutlar arasındaki ilişkileri kullanmasının doğal bir sonucudur. De la Torre, Song ve Hong (2011) Üst Düzey Sıralı Model'e dayalı ürettikleri veriler üzerinde de bu araştırmayla benzer sonuçlara ulaşmıştır. Aynı zamanda 0.8 korelasyon düzeyinde Üst Düzey Sıralı Model'in biraz daha doğru ve güvenilir kestirimler yapmasıyla birlikte bu model ile Hiyerarşik ÇBMTK model benzer performans göstermiştir. Bu durumun iki modelin de alt testler arasındaki korelasyonları kullanarak yetenek parametresi kestirmesinden kaynaklandığı düşünülmektedir. Bu araştırma sonuçlarından farklı olarak Yao (2010) bu iki modeli benzer ama minimal farkla ÇBMTK modelin daha iyi performans gösterdiğini bulurken de la Torre, Song ve Hong (2011) bu iki modelin performansını birbirine eşit bulmuştur. Bu araştırmanın alt test sayısındaki artış ile Üst Düzey Sıralı Model'in alt test puan kestirimlerinin hatası ve güvenirliği üzerinde minimal düzeyde iyileşme sağladığı sonucu ile de la Torre, Song ve Hong'un (2011) araştırma sonuçları benzedir. Daha uzun alt testlerde üç yöntem için de kestirim hatalarının azalması ve güvenirliğin artması beklenen bir durumdur.

Gerçek veri uygulamasında TEOG 2015 gerçek verisinin faktör analizi sonucunda elde edilen alt testler arası korelasyon matrisi ile Hiyerarşik ÇBMTK Model, Üst Düzey Sıralı Model ve İki Faktör Model tarafından kestirilen matris karşılaştırıldığında modellerin gerçek duruma çok benzer kestirimler yaptığı görülmüştür. Din Kültürü, Fen Bilgisi, İngilizce, Matematik, Tarih ve Türkçe alt testlerinin her bir yöntemden elde edilen sonsal dağılımın ortalama standart sapma değerleri karşılaştırıldığında üç yönteminde genel olarak düşük sonsal standart sapmaya sahip olduğu bulunmuştur. Fakat daha fazla alt düşük sonsal dağılım ortalama ve standart sapma değerleri verdiği için İki Faktör modelin diğer yöntemlere göre az farkla daha iyi sonuç verdiği sonucuna ulaşılmıştır.

Araştırmadan elde edilen bulgulara dayanarak; hiyerarşik modellerin varsayımının hem toplam test hem de alt test puanlarının kestiriminde farklı performans göstermesi nedeniyle daha doğru ve güvenilir alt test ve toplam test puanı kestirimleri için öncelikle mevcut testin yapısal modeli tespit edilmelidir. Toplam test puan kestirimlerinde araştırmada ele alınan tüm koşullar altında ve alt test puanların kestirimde ise hemen hemen tüm koşullarda en düşük hatalı ve en güvenilir kestirimlerin hiyerarşik ÇBMTK modelden elde edilmesi nedeniyle geniş ölçekli testlerin raporlanmasında bu modelin kullanımı önerilebilir. Alt testler arasında orta ve düşük düzeyde ilişkilerin olduğu bilinen sınavların raporlanmasında hiyerarşik ÇBMTK Model'e alternatif olarak bu modelle çok yakın analizler yapabilen Üst Düzey Sıralı Model'in kullanımı da tercih edilebilir. Alan yazında alt testler arasında yüksek düzeyde ilişki olduğu bilinen sınavlarda toplam test puanı kestirimleri için İki

Faktör Model'in kullanımı önerilirken bu araştırmada ele alınan koşullara sahip sınavlar için alt test puan kestirimlerinde bu yöntemin kullanımı önerilmez. Toplam test süresi göz ününde bulundurmak koşuluyla hem toplam hem de alt test puanları kestirimin doğruluğunu ve güvenirliğini arttıracağı için alt test uzunluklarının arttırılması önerilebilir.

KAYNAKÇA

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (1999*). Standards for educational and psychological testing*. American Educational Research Association, Washington, DC.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*(2), 153–168, doi: 10.1002/j.2333-8504.1998.tb01752.x

Brandt, S., & Duckor, B. (2013). Increasing unidimensional measurement precision using a multidimensional item response model approach. *Psychological Test and Assessment Modeling, 55*(2), 148-161.

Brennan, R. L. (2012). *Utility indexes for decisions about subscores* (No. 33). Center for Advanced Studies in Measurement and Assessment (CASMA). Retrieved from https://education.uiowa.edu/sites/education.uiowa.edu/files/documents/centers/casma/publications/casma-research-report-33.pdf

Bulut, O. (2013). *Between-person and within-person subscore reliability: Comparison of unidimensional and multidimensional IRT models*. (Doctoral Dissertation). Retrieved from https://conservancy.umn.edu/bitstream/handle/11299/155592/Bulut_umn_0130E_13879.pdf?sequence=1&isAllowed=y

Chang, Y. F. (2015). *A Restricted Bi-factor Model of Subdomain Relative Strengths and Weaknesses.* (Doctoral Dissertation) Retrieved from https://conservancy.umn.edu/bitstream/handle/11299/175551/CHANG_umn_0130E_16452.pdf?sequence=1&isAllowed=y

Çakıcı Eser, D. (2015). *Çok boyutlu madde tepki kuramının farklı modellerinden çeşitli koşullar altında kestirilen parametrelerin incelenmesi.* (Doktora tezi). Erişim adresi: http://tez2.yok.gov.tr/

de la Torre, J., & Patz, R.J. (2005). Making the most of what we have: A practical application of multidimensional IRT in test scoring. *Journal of Educational and Behavioral Statistics, 30*(3), 295–311, doi: 10.3102/10769986030003295

de la Torre, J. (2009). Improving the quality of ability estimates through multidimensional scoring and incorporation of ancillary variables. *Applied Psychological Measurement, 33*(6), 465–485, doi: 10.1177/0146621608329890

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement, 33*(8), 620-639, doi: 10.1177/0146621608326423

de la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT subscoring. *Applied Psychological Measurement, 35(*4), 296-316, doi: 10.1177/0146621610378653

Edwards, M. C., & Vevea, J. L. (2006). An empirical Bayes approach to subscore augmentation: How much strength can we borrow?. *Journal of Educational and Behavioral Statistics, 31*(3), 241-259, doi: 10.3102/10769986031003241

ETS. (2014). *ETS standards for quality and fairness*. Educational Testing Service. .Retreived from https://www.ets.org/s/about/pdf/standards.pdf

Ferrara, S., & DeMauro, G. E. (2007). Standardized assessment of individual achievement in K–12. In R. L. Brennan (Eds.). *Educational measurement*, 579–622. Westport, CT: Praeger.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2011). *How to design and evaluate research in education*. (8th edition). Boston: McGraw – Hill.

Gall M. D., Gall, J. P., & Borg, W., R. (2003). *Educational research: An introduction*. (7th. Edition). Pearson Education, Inc.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item Bi-factor analysis. *Psychometrika, 57*, 423–436.

Haberman, S. J. (2008). When can subscores have value*? Journal of Educational and Behavioral Statistics, 33*(2), 204–229, doi:10.3102/1076998607302636

Haberman, S., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology, 62*(1), 79–95, doi:10.1348/000711007X248875

Haladyna, T. M., & Kramer, G. A. (2004). The validity of subscores for a credentialing Test. *Evaluation & The Health Professions, 27*(4), 349–368, doi: 10.1177/0163278704270010

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, doi: 10.1177/014662169602000201

Huang, H. Y., Wang, W. C., Chen, P. H., & Su, C. M. (2013). Higher-order item response models for hierarchical latent traits. *Applied Psychological Measurement, 37*(8), 619-637, doi: 10.1177/0146621613488819

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in psychology, 7*(109), 1-10, doi: 10.3389/fpsyg.2016.00109

Kelley, T. L. (1927). *The interpretation of educational measurements*. New York: World Book.

Kelley, T. L. (1947). *Fundamentals of statistics*. Cambridge: Harvard University Press Kerlinger.

Kerlinger, F.N. (1973). *Foundation of behavioural research*. New York. Holt. Rinehand and Hinston.

Köse, İ.A. (2010). *Madde tepki kuramına dayalı tek boyutlu ve çok boyutlu modellerin test uzunluğu ve örneklem büyüklüğü açısından karşılaştırılması*. (Doktora Tezi). Erişim adresi: http://tez2.yok.gov.tr/

Lee, J. (2012). *Multidimensional item response theory: an investigation of interaction effects between factors on item parameter recovery using Markov Chain Monte Carlo*. (Doctoral Dissertation). Retrieved from https://d.lib.msu.edu/islandora/object/etd:1577/datastream/OBJ/download/Multidimensional_item_response_theory__an_investigation_of_interaction_effects_between_factors_on_item_parameter_recovery_using_Markov_Chain_Monte_Carlo.pdf

Ling, G. (2012). *Why the major field test in business does not report subscores: Reliability and construct validity evidence* (No. RR-12-11). ETS Research Report. Retrieved from https://www.ets.org/Media/Research/pdf/RR-12-11.pdf

Lorenzo-Seva, U., & Ferrando, P.J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavioral Research Methods, Instruments and Computers, 38*(1), 88-91.

Messick, S. (1989). Validity. In R. L. Linn (Eds.). *Educational measurement*, 13-103, New York, NY: Macmillan.

Monaghan, W. (2006). The fact about subscores (No. RDC-04). ETS Research Report. Retrieved from https://www.ets.org/research/policy_research_reports/rdc-04

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*. (Doktora Tezi). Erişim adresi: http://tez2.yok.gov.tr/

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21, 25–36, doi: 10.1177/0146621697211002

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*(1), 53-61.

Sheng, Y., & Wikle, C. K. (2007). Comparing Multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement, 67*(6) 899–919, doi: 10.1177/0013164406296977

Sheng, Y., & Wikle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement, 68*(3), 413–430, doi: 10.1177/0013164407308512

Shin, D. (2007). *A comparison of methods of estimating subscale scores for mixed-format tests*. Report for Pearson Educational Measurement. Retreived from https://images.pearsonassessments.com/images/tmrs/tmrs_rg/EstimatingSubscaleScoresforMixedFormatItemsforPEMreportfinal.pdf?WT.mc_id=TMRS_A_Comparison_of_Methods_of_Estimating

Shin, C. D., Ansley, T., Tsai, T., & Mao X. (2005, April). *A comparison of methods of estimating objective scores*. Annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47*(2),150-174.

Skorupski, W. P., & Carvajal, J. (2010). A comparison of approaches for improving the reliability of objective level scores. *Educational and Psychological Measurement, 70*(3), 357-375, doi: 10.1177/0013164409355694

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L.,Swygert, K. A., & Thissen, D. (2001). Augmented score—''borrowing strength'' to compute scores based on small numbers of items. In D. Thissen and H. Wainer (Eds.). *Test scoring*, (343-387). Mahwah, Lawrence Erlbaum Associates, Inc

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126–149, doi: 10.1177/0146621604271053

Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods, 9*(1), 116, doi: 10.1037/1082-989X.9.1.116

Yao, L. (2003). SimuMIRT [Software]. Monterey, CA: Defense Manpower Data Center. Retreived from http://www.bmirt.com

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement, 47*(3), 339-360, doi: 10.1111/j.1745-3984.2010.00117.x

Yao, L. (2017). Comparing methods for estimating the abilities for the multidimensional models of mixed item types. *Communications in Statistics-Simulation and Computation*, 1-18, doi: 10.1080/03610918.2016.1277749

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, doi: 10.1177/0146621606291559

Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement, 30*(6), 469–492, doi: 10.1177/0146621605284537

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for 2 latent trait models. *Journal of Educational Measurement, 17*(4), 297–311, doi: 10.1111/j.1745-3984.1980.tb00833.x

Yen, W. M. (1987,June). *A Bayesian/IRT index of objective performance*. Annual meeting of the Psychometric Society, Montreal, Quebec, Canada.

EXTENDED ABSTRACT

Introduction

In many developed countries, large-scale standardized tests are the most common measurement tools used in education and psychology. These tests have multiple components which consists of subsets of items that measure specific content or structure. Although the total test score estimate is useful for important decisions, the subscores complement the total test score estimate by providing finer grained diagnosis of weakness and strengths of examinees. In this context, there is an increasing interest in subscores in educational testing. Reliable subscores should be obtained to make valid inferences about attributes of a student in the subtests. In practice, conventional analysis of tests with multiple components ignores its multidimensionality and only responses specific to each subtest are used in estimating the subscores of examinees.

In this study, the relationship between subtest and total test was investigated by using hierarchical item response theory models in order to contribute to reliable subtest and total test score estimates. The RMSE and reliability of the total test score and subtest scores estimated by the Higher Order, Bi-factor and hierarchical MIRT models in the study were compared under the conditions of the size of the correlations between the subtest number, subtest length and number of subtests. In addition, the performance of three models used in the research was examined on TEOG 2015 data.

Method

To generate data sets based on the item parameters of the TEOG 2015 data, item discrimination parameters were drawn from normal distribution with a mean of 1.5 and a variance of 0.5; item difficulty parameters were drawn from normal distribution with a mean of 0.0 and a variance of 1.0, and guessing (lower asymptote) parameters were drawn from beta distribution with (6,16). The true subtest abilities were drawn from a multivariate normal distribution with variance-covariance matrix based on the correlations between the dimensions explained under simulation conditions. Finally, given subtest abilities and item parameters, binary responses were simulated for number of subtest (2,3), subtest length (20,30,40) and correlation between subtest (0.0, 0.3, 0.5, 0.8) by SimuMIRT software. The simulated data and TEOG 2015 data was analyzed by BMIRT software. For the parameter estimates, 3PL model and MCMC estimation method were used.

Results and Discussion

As a result of the study, when the correlation between the subtests and the subtest length increased, the RMSE of the ability parameters decreased and the reliability increased for the total test score obtained from Higher Order and Bi factor Models. But with higher levels of correlation between the subtests in both two- and three-dimensional datasets, more errors were obtained for total test score from the Hierarchical MIRT model. This might be caused by the decrease in the test dimensionality. For the same reason, it was observed that as the correlation level between subtests increased in the total test score estimates from the Bi Factor Model, the errors decreased since the total test score in the Bi Factor Model is obtained from the calibration of all the items in the test. As a result, it was found that Hierarchical MIRT Model outperformed Higher Order and Bi Factor Models with regards to total test score recovery for all conditions. But when the correlation between subtests was .8 all three methods performed similarly. Furthermore, the reliability values obtained from Hierarchical MIRT model under all conditions were at least .7. The increase in the number of subtests contributed to the more accurate total test score estimates for three models, with Hierarchical MIRT maximum information model performing slightly better than the two other methods.

Under all conditions, the lowest RMSE value and the highest reliability value were yielded from Hierarchical MIRT model for subscores recovery and Bi factor model performed the worst. When the correlation between the subtests increased in both two- and three-dimensional datasets, the RMSE of the ability parameters decreased and the reliability increased for the subscores obtained from Higher Order whereas RMSE of the ability parameters increased and the reliability decreased for the subscores obtained from Bi Factor model. These results for Bi Factor models are similar with the ones obtained from the paper of Yao (2010) and Chang (2015). Also, it was unexpectedly deduced that there was no effect of the level of correlation between the subtests on the subscores estimation errors and reliability obtained from the Hierarchical MIRT Model. However, it was found in the studies of Bulut (2013) and Yao (2010) that when the correlation between the subtests increased, the reliability and accuracy of the Hierarchical MIRT model subscores estimates increased. It was found from real data analysis that all three methods gave similar estimates for subscores. This was consistent with results obtained for the condition in which the correlation between subtests was .0.

According to the results of the research, to report total test scores with reliability greater than .8: The correlation between subtests has to be higher than .3 to use either Hierarchical MIRT or Higher Order model and has to be than .5 to use Bi factor model for a test of 20 items for each subtest. Also, the subtest length has to be at least 30 when the correlation between subtests is at .0 for Hierarchical MIRT model. Both Hierarchical MIRT and Higher Order model can give subscore estimates with greater than .8 at all level of correlation between subtests for at least a test of 20 items for each subtest.

Based on findings from the study; the use of the Hierarchical MIRT model is recommended for the reporting of large scale tests. In reporting exams known to have moderate and low correlations among the sub-tests, it may also be preferable to use the Higher Order model, which is able to perform close analyzes with the Hierarchical MIRT Model, as an alternative to the Hierarchical MIRT Model.

Ekler

Tablo EK-1. Modellerin Kestirdiği Alt Testler Arası Korelasyonların Standart Sapma Matrisi

Model	Alt testler	Din Kültürü	Fen Bilgisi	İngilizce	Matematik	Tarih	Türkçe
İki Faktör	Fen Bilgisi	0.033	1.000	0.007	0.011	0.014	0.009
	İngilizce	0.039	0.007	1	0.009	0.020	0.011
	Matematik	0.038	0.011	0.009	1	0.006	0.010
	Tarih	0.038	0.014	0.020	0.006	1	0.008
	Türkçe	0.041	0.009	0.011	0.010	0.008	1
ÇBMTK	Fen Bilgisi	0.010	1	0.009	0.009	0.008	0.007
	İngilizce	0.010	0.009	1	0.008	0.013	0.008
	Matematik	0.008	0.009	0.008	1	0.007	0.008
	Tarih	0.013	0.008	0.013	0.007	1	0.011
	Türkçe	0.009	0.007	0.008	0.008	0.011	1
Üst Düzey	Fen Bilgisi	0.010	1	0.010	0.009	0.008	0.007
	İngilizce	0.011	0.010	1	0.008	0.017	0.012
	Matematik	0.007	0.009	0.008	1	0.009	0.009
	Tarih	0.016	0.008	0.017	0.009	1	0.014
	Türkçe	0.008	0.007	0.012	0.009	0.014	1

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(2);202-215

Otizm Sosyal Beceriler Profili Ölçeğinde Puanlayıcılar Arası Güvenirliğin Farklı Kuramlara Göre Karşılaştırılması*

Comparison of Interrater Reliability Based on Different Theories for Autism Social Skills Profile

Zeynep PEKİN ** Sevda ÇETİN *** Neşe GÜLER ****

Öz

Bu araştırmada, "Otizm Sosyal Beceriler Profili" (OSBP) ölçeğinin beş puanlayıcı tarafından puanlanması ile elde edilen puanların klasik test kuramı ve genellenebilirlik (G) kuramı ile puanlayıcılar arası güvenirliğinin karşılaştırılması amaçlanmıştır. G kuramında puanlayıcıların birlikte ve dönüşümlü puanlama yapmasıyla oluşturulan farklı desenlerden ve klasik test kuramından elde edilen güvenirlik katsayılarının düzeyleri saptanmış ve hangi kuramın daha fazla bilgi sunduğu belirlenmeye çalışılmıştır. Araştırmada elde edilen veriler klasik test kuramında her bir puanlayıcı için puanların iç tutarlılık güvenirliği Cronbach-alfa (α) katsayısı; puanlayıcılar arası güvenirlik, Kendall'ın uyuşum katsayısı, puanlayıcılar arası korelasyon katsayısı ve puanlayıcıların verdikleri puanlar arasında fark olup olmadığı ise ilişkili örneklemlerde varyans analizi ile hesaplanmıştır. Genellenebilirlik teorisinde, değerlendiricilerin ortaklaşa ve alternatif derecelendirmelerine göre iki farklı tasarım oluşturulmuştur. G kuramı kapsamında bireylerin (*b*) aynı maddeler (*m*) doğrultusunda puanlayıcıların (*p*) her biri tarafından puanlandığı *bxmxp* çapraz deseni ve bireylerin tüm maddeler doğrultusunda farklı puanlayıcılar tarafından puanlandığı (*p:b*)*xm* yuvalanmış deseni için ayrı ayrı G ve K çalışmaları yapılmış ve sonuçlar birbirleriyle karşılaştırılmıştır.

Anahtar Kelimeler: Klasik test kuramı, genellenebilirlik kuramı, puanlayıcılar arası güvenirlik, Kendall'ın uyuşum katsayısı, sosyal becerilerin değerlendirilmesi

Abstract

In this study, interrater reliability was compared based on both classical test theory and generalizability theory according to the scores which were obtained from five raters' ratings with Autism Social Skills Profile. Levels of reliability coefficients obtained from classical test theory and different designs in generalizability theory formed by five raters' jointly and alternatively ratings were determined and which theory presented more information was tried to be specified. In the classical test theory, Cronbach-Alpha (α) coefficient for internal consistency, Kendall's coefficient of concordance for inter-rater reliability and correlation coefficients of five raters' scores were calculated and it was investigated whether there was a difference among the means of raters' scores with F test. In the generalizability theory, two different designs were formed according to raters' jointly and alternatively ratings. Several G and D studies were made for crossed design pxixr (p: person, i: item and r: rater) which people were scored by all raters through all items and nested design (r:p)xi which people were scored by different raters through all items and the results were compared to each other.

Keywords: Classical test theory, generalizability theory, interrater reliability, Kendall's coefficient of concordance, evaluation of social skills

* Bu çalışma, Zeynep Pekin' in Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsünde, Dr. Öğr. Üyesi Sevda Çetin'in danışmanlığında hazırlanan yüksek lisans tezinden üretilmiştir.

** Arş. Gör., Yeditepe Üniversitesi, Eğitim Fakültesi, İstanbul-Türkiye, zynppknn@gmail.com:, http://orcid.org/0000-0002-9976-1218
*** Dr. Öğr. Üyesi, Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, tsevda@hacettepe.edu.tr http://orcid.org/0000-0001-5483-595X
**** Doç. Dr., İzmir Demokrasi Üniversitesi, Eğitim Fakültesi, İzmir-Türkiye, gnguler@gmail.com: https://orcid.org/0000-0002-2836-3132

Geliş Tarihi: 01.02.2018
DOI: *10.21031/epod.388590* *Kabul Tarihi: 18.05.2018*

GİRİŞ

Sosyal beceriler, sosyal bir ortamda kabul edilebilir bir şekilde başkalarıyla iletişime geçebilme yeteneğidir (Combs & Salaby, 1977). Bu becerilerde görülen yetersizlikler bireylerin sosyal ve eğitim hayatlarında iletişim kurmalarında sıkıntı yaşamalarına sebep olmaktadır. Bu durumdan yüksek düzeyde etkilenen özel gereksinim gruplarından birini de otizm spektrum bozukluğu olan bireyler oluşturmaktadır. Nitekim alanyazında otizm spektrum bozukluğu (OSB), sosyal etkileşimlerde yüksek seviyede yetersizlikler ve stereotip davranışlar ile karakterize nöro-gelişimsel bir bozukluk olarak tanımlanmaktadır (Özdemir, Diken, Diken & Şekercioğlu, 2013). Otizmli bireylerin sosyal becerilerindeki sınırlılıkları, başkalarına yaklaşmada sıra dışı özellikler gösterme (nerede duracağını bilememek), arkadaşlık kurmada sıkıntı yaşama, grup etkinliklerinde zorlanma, yalnızlığı yeğleme, başkalarının dikkatini çekme çabası göstermeme, sözel övgüler karşısında tepkisiz kalma, başkalarına karşı ilgisiz olma, başkalarının duygularını anlamada yetersiz olma şeklinde sıralanabilmektedir (Kırcaali-İftar, 2012).

Otizmli bireyler iletişim kurmada sıkıntı yaşamaları sebebiyle aile ve akranlarından kopuk bir yaşam sürmektedirler. Bu sebeple sosyal becerileri doğal ortamda gözlemleyerek arkadaş ve ailelerinden öğrenememektedirler. Bu becerilerin öğrenilebilmesi için bireyin özelliklerine göre yapılandırılmış sosyal beceri öğretim programlarına ihtiyaç duyulmaktadır (Hall & Schlesinger, 1997). Bu programların temelini ise otizmli bireylerin sosyal becerilerinin değerlendirilmesi oluşturmaktadır (Merrel, 2001). Ancak, özel eğitim alanında yapılan çalışmalarda çoğunlukla otizmli bireylerin kendilerini değerlendirecek yeterlilikte olmamaları sebebiyle değerlendirmeler aile, öğretmen gibi bireye yakın kişiler tarafından dereceleme ölçekleriyle yapılmaktadır. Bu tarz değerlendirmelerde, dikkatsizlik, kişisel yanlılık, halo etkisi, merkeze kayma etkisi, genelleme hatası, gözlem yetersizliği vb. gibi puanlayıcı kaynaklı hatalar karışabilmektedir (Turgut & Baykul, 2014). Bu sebeple, güFnilir sonuçların elde edilebilmesi için özel eğitim alanında değerlendirme yapılırken puanlayıcılar arası güvenirliğin test edilmesi oldukça önem kazanmaktadır. Puanlayıcılar arası güvenirliğin belirlenmesinde kullanılan çeşitli kuram ve uygulamaları bulunmaktadır. Ancak bu araştırmada, klasik test kuramı ve genellenebilirlik kuramı kapsamındaki yöntemler kullanıldığı için sadece bu kuramlara dair bilgiler yer almaktadır.

Klasik Test Kuramı

Klasik test kuramında (KTK), gerçek puan varyansı ve hata varyansının toplamından gözlenen puan varyansı oluşmaktadır (Lord & Novick, 1968). Gerçek varyans hariç kalan tüm varyansın farklı hata kaynaklarından gelebileceği düşünülmektedir. KTK'da, ele alınan hata kaynağına göre güvenirlik hesaplama yöntemleri farklılık göstermektedir (Baykul, 2000). Güvenirlik hesaplanırken olası hata kaynağı, test-tekrar test yönteminde ele alınan zaman, paralel (eşdeğer) formlar yönteminde formlar, iç tutarlılık anlamında hesaplama yapılırken ise görevler ya da maddeler olmaktadır (Shavelson & Webb, 1991; Brennan, 2001; Güler & Gelbal, 2010). Birden fazla puanlayıcının yer aldığı ölçmelerde ise puanlayıcılar hata kaynağı olarak düşünülmekte ve puanlayıcıların verdikleri puanlar arasındaki tutarlılığa 'puanlayıcılar arası güvenirlik' denilmektedir (Güler, 2008). Birden fazla puanlayıcının değerlendirme yaptığı çalışmalarda, puanlayıcılar önemli bir hata kaynağı olmakla birlikte çalışmadaki tek hata kaynağı değildir. Sonuçların kolay yorumlanması sebebiyle sıkça tercih edilmesine rağmen, KTK kapsamında yapılan güvenirlik yöntemlerinde hata kaynaklarının ayrı ayrı ele alınması önemli bir sınırlılık olarak karşımıza çıkmaktadır (Güler & Gelbal, 2010). Genellenebilirlik (G) kuramı, birden fazla hata kaynağının olduğu durumlarda güvenilirlik kestirimi için klasik test kuramının bir uzantısı olarak geliştirilen bir yaklaşımdır. G kuramı, tek bir analizle hata kaynaklarının aynı anda kestirimini sağlamakla birlikte, farklı uygulamalara katkı sağlayacak hata varyanslarına ilişkin bilgi edinilmesine de imkân tanımaktadır. (Cardinet, Johnson & Pini, 2009).

Genellenebilirlik Kuramı

Genellenebilirlik (G) kuramı, davranış değerlendirilmesinde güvenirliğin hesaplanmasını, güvenilir gözlemlerin genellenebilirlik (G) ve karar (K) çalışmalarıyla tasarlanıp, araştırılmasını ve çeşitli hata kaynaklarını göz önünde bulundurarak tek bir güvenirlik katsayısının bulunmasını sağlayan istatiksel bir kuramdır (Eason, 1989).

G kuramında; zaman, puanlayıcı, madde, formlar gibi benzerlik gösteren ölçme durumlarına, değişkenlik kaynaklarına yüzey (facet) adı verilir ve faktör analizinde yer alan faktörler gibi düşünülebilir (Güler, Uyanık & Teker, 2012). Her bir yüzeyin düzeyleri ise koşul olarak adlandırılmaktadır (Crocker & Algina, 1986; Shavelson & Webb, 1991; Brennan, 2001). Örneğin, "madde" bir yüzey olarak ele alınırsa; maddelerin her biri bir koşul olmaktadır. Araştırmacının, genellemek istediği yüzeyin koşullarına genelleme evreni, alınabilecek tüm koşulların oluşturacağı evrene ise kabul edilebilir gözlemler evreni denir (Crocker & Algina, 1986). Ölçme sonucunda, istenilen kararların alınacağı ölçmenin hedefi durumundaki değişkenlik kaynağı ise ölçme objesi olarak adlandırılmaktadır. G kuramında, ölçme objesine yüzey adı verilmemektedir. Ölçme objesi genellikle sistematik varyans içeren bireylerdir. Çünkü bireyler, doğası gereği farklılık gösterir. Ancak bireyler her zaman ölçmenin objesi olmak zorunda değildir. Ölçme durumuna göre, sistematik varyans içeren diğer madde, durum vb. de ölçme objesi olabilir (Eason, 1989; Shavelson & Webb, 1991; Mushquash & O'connor, 2006).

G kuramında, G çalışması ve K çalışması adı verilen iki tür çalışma yer almaktadır. G çalışmasının amacı, kabul edilebilir gözlemler evrenine ilişkin varyans bileşenlerini belirlemek ve bu varyans kaynakları hakkında bilgi edinmektir (Shavelson & Webb, 1991; Keiffer, 1998; Brennan, 2001). Varyans bileşenleri belirlendikten sonra 'Eğer ki?' sorularına cevap aramak için K çalışmasına geçilir. K çalışması sürecinde araştırmacı, madde, form ya da puanlayıcı sayıları gibi yüzeylerin koşulları üzerinde değişiklikler yaparak daha yüksek güvenirlik ve daha düşük hata içeren sonuçlar elde edebilecek senaryolar üretebilir (Kieffer, 1998).

Değişkenlik kaynaklarının ele alınış şekillerine göre G kuramında iki tür desen vardır. Bir ölçmedeki değişkenlik kaynaklarının (ölçme objesi ve yüzeyler dâhil) tüm koşulları, diğer değişkenlik kaynaklarının tüm koşullarıyla etkileşim gösteriyorsa bu desene çaprazlanmış desen denir ve bu desende değişkenlik kaynakları arasına *'x'* işareti konulur. Bir ölçmedeki bir değişkenlik kaynağının bazı koşulları, diğer bir değişkenlik kaynağının bazı koşullarıyla etkileşim gösteriyorsa; bu desene de yuvalanmış desen denir ve bu desende de değişkenlik kaynakları arasında *':'* işareti yer alır (Shavelson & Webb, 1991; Brennan, 2001; Mushquash & O'Connor, 2006). Ayrıca, G kuramı, mutlak ve bağıl olmak üzere iki tür değerlendirmeye olanak sağlar ve bu iki değerlendirme için iki farklı güvenirlik katsayısının hesaplanmasına imkân tanır. Bağıl değerlendirmeler için G-katsayısı, mutlak değerlendirmeler için ise Phi-katsayısı (Φ) hesaplanarak güvenirlik kestirilmektedir (Crocker & Algina, 1986; Shavelson & Webb, 1991; Brennan, 2001; Goodwin, 2001).

Alanyazında puanlayıcılar arası güvenirliğin KTK ve G kuramına göre karşılaştırıldığı çalışmalara rastlanılmaktadır (Rae & Hyland, 2001; Yelboğa & Tavşancıl, 2010; Öztürk, 2011; Deliceoğlu & Çıkrıkçı-Demirtaşlı 2012; Yıldıztekin, 2014, Polat-Demir, 2016). Örneğin, Rae ve Hyland (2001)'ın çalışmasında KTK ve G kuramı tutarlı sonuçlar vermiş ve puanlayıcılar arası güvenirlik yüksek bulunmuştur. Öte yandan, Öztürk (2011)'ün çalışmasında puanlayıcılar arası güvenirlik KTK ve G kuramı kapsamında düşük elde edilmiştir. Yelboğa ve Tavşancıl (2010)'ın çalışmalarında KTK'da puanlayıcılar arası güvenirlik için Kendall'ın uyuşum katsayısı ve G kuramında da G ve Phi katsayıları hesaplanmış ve her iki kuramdan elde edilen katsayıların birbirleriyle tutarlı oldukları sonucuna varılmıştır. Yıldıztekin (2014)'in çalışmasında KTK'da Pearson momentler çarpım korelasyon katsayısı, Spearman sıra farkları korelasyon katsayısı, Kappa ve Krippendorf Alfa katsayıları ile G kuramı karşılaştırılmış ve elde edilen sonuçlara göre puanlayıcılar arası güvenirlik yüksek bulunmuştur.

Genel olarak alanyazında puanlayıcılar arası güvenirliğin KTK ve G kuramına göre karşılaştırıldığı çalışmalar yer alsa da, özel eğitim alanında KTK ve G kuramının karşılaştırıldığı çalışmalara rastlanılmamıştır. Özel eğitim alanında otizmli bireylerin değerlendirilmesinin öğretmen gibi bireye

yakın kişiler tarafından yapıldığı göz önünde bulundurulduğunda, puanlayıcılar arası güvenirlik çalışmasının önemli olduğu düşünülmektedir. Bu alanda çalışan araştırmacıların, çalışmalarında puanlayıcılar arası güvenirliği genellikle KTK'ya dayalı yöntemlerle belirledikleri görülmektedir (Irmak, Sütçü, Aydın & Sorias, 2007; Girli & Atasoy, 2010; Sucuoğlu & Demir, 2017). Alanyazında birden fazla puanlayıcının değerlendirme yaptığı çalışmalarda G kuramı, KTK'ya göre daha kapsamlı ve güvenilir sonuçlar üretmesi nedeniyle daha çok önerilmektedir (Shavelson & Webb, 1991; Yelboğa & Tavşancıl, 2010; Güler & Gelbal, 2010; Güler, 2011). Bu doğrultuda gerçekleştirilen araştırmanın amacı, Otizm Sosyal Beceri Profili (OSBP) ile otizmli çocuk ve gençlerin birden fazla puanlayıcı tarafından puanlanması sonucu elde edilen sonuçlarda puanlayıcılar arası güvenirliğin KTK ve G kuramına göre karşılaştırılmasıdır. Bu araştırma ile kuramlardan elde edilen bilgilerin karşılaştırılmasına, gelecek araştırmalara ve özel eğitim alanında ölçme ve değerlendirme sürecine yönelik bir katkı sağlanabileceği düşünülmektedir. Ayrıca, bu çalışmada G kuramı kapsamında puanlayıcıların birlikte ve dönüşümlü puanlama yapmasıyla oluşturulan farklı desenlerin sonuçlarının karşılaştırılması amaçlanmaktadır. Her puanlayıcının her bireyi bütün maddeler doğrultusunda puanladığı çapraz desenler uygulama açısından zaman zaman çok pratik olamamaktadır. Bununla birlikte puanlayıcıların yorulması sebebiyle değerlendirmelerin güvenirliği riske girebilmektedir. Bu nedenle çalışmada her iki desenle elde edilen sonuçların karşılaştırılarak, yeterli düzeyde güvenilir sonuçlar için yuvalanmış desenin uygunluğunun belirlenmesinin gelecek çalışmalara katkı sağlayacağı düşünülmektedir.

Problem Cümlesi

Otizm Sosyal Beceriler Profili (OSBP) ölçeğinin sosyal karşılıklılık alt boyutunun birden fazla puanlayıcı tarafından puanlanması sonucu elde edilen puanlayıcılar arası güvenirliğin, KTK ve G kuramındaki farklı desen çalışmaları sonuçları nasıldır?

Alt Problemler

1. KTK'na göre; OSBP sosyal karşılıklılık alt boyutundan elde edilen puanların puanlayıcılara ilişkin iç tutarlılık düzeyi ve beş farklı puanlayıcı arasındaki tutarlılık derecesi nedir?
2. G kuramına göre; birey (*b*), madde (*m*) ve puanlayıcı (*p*) değişkenlerinin çaprazlandığı *bxmxp* deseninin sonuçları nasıldır?
3. G kuramına göre; birey (*b*) ve puanlayıcı (*p*) değişkenlerinin yuvalandığı, madde (*m*) değişkeninin ise çaprazlandığı *(p:b)xm* deseninin sonuçları nasıldır?
4. *bxmxp* ve *(p:b)xm* desenlerinden elde edilen G çalışması sonuçlarının karşılaştırılması nasıldır?

YÖNTEM

Araştırmanın Türü

Araştırma, puanlayıcılar arası güvenirliğin hesaplanmasında KTK ve G kuramından hangisinin daha çok bilgi sağladığını belirlemesi ve iki kuramdan elde edilen sonuçları karşılaştırması açısından karşılaştırmalı bir araştırmadır. Aynı zamanda araştırma, G kuramı ve KTK ile OSBP ölçeğine ait özelliklerin belirlenmesi yönüyle durum belirleme çalışması olduğundan betimsel bir araştırma niteliği taşımaktadır.

Araştırma Grubu

Araştırmanın çalışma grubunu, Ankara'da bir özel eğitim ve rehabilitasyon merkezinde eğitim almakta olan, "Otizm", "Asperger Sendromu", "Başka Şekilde Tanımlanamayan" tanılarını almış ya

da bu gruplardan herhangi birine dahil edilemeyen ancak otizm spektrum bozukluğu belirtileri gösteren "Yaygın Gelişimsel Bozukluk" tanısı almış 6-17 yaş arasında olan 50 çocuk ve genç oluşturmuştur. Örneklem grubu seçilirken, puanlayıcıların en az bir yıldır birlikte çalıştıkları öğrenciler seçilmiştir. 50 öğrenci, OSBP ölçeğinin "sosyal karşılıklılık" alt boyutunda bulunan 15 madde ile beş puanlayıcı tarafından puanlanmıştır. Puanlayıcı grubunu ise aynı kurumda 3-5 yıldır görev yapmakta olan iki özel eğitim öğretmeni, bir psikolog (özel eğitim alanıyla ilgilenen), bir fizyoterapist ve bir sosyal psikolog (aynı zamanda davranış terapisti) oluşturmaktadır.

Veri Toplama Aracı

OSBP, Bellini ve Hopf (2007) tarafından otizmli çocukların sosyal beceri yetersizliklerinin belirlenmesi ve bulgular doğrultusunda uygun müdahale programlarının oluşturulması gayesiyle geliştirilmiştir. Araştırmanın örneklemini, otizm spektrum bozukluğuna sahip 6-17 yaş arası 340 çocuk ve genç oluşturmuştur. 45 maddelik OSBP, sosyal karşılıklılık, sosyal katılım/kaçınma ve zarar verici sosyal davranışlar şeklinde adlandırılan üç alt boyuttan meydana gelmektedir. Bu çalışmada, OSBP ölçeğinin "sosyal karşılıklılık" alt ölçeğinde yer alan 15 madde kullanılmıştır. Ölçeğin Cronbach-alfa iç tutarlık katsayısı 0.92 iken, sosyal karşılıklılık alt ölçeği için yine bu değer 0.92 olarak bulunmuştur.

OSBP'nin Türkçe uyarlaması Demir (2009) tarafından 208 otizmli çocuk ve gencin ebeveynleri tarafından değerlendirilmesi sonucu elde edilen verilerle gerçekleştirilmiştir. Cronbach-alfa katsayıları alt ölçeklerden sosyal karşılıklılık için 0.91 ve ölçeğin tamamı için 0.84 bulunmuştur. Bu sonuçlara göre OSBP ölçeğinin geçerli ve güvenilir bir araç olduğuna karar verilmiştir (Demir, 2009).

Verilerin Analizi

Dörtlü likert tipi bir ölçek olan OSBP kullanılarak 50 öğrenci, beş uzman tarafından değerlendirilmiştir. Puanlayıcılar birbirlerinden bağımsız puanlama yapmışlardır. Elde edilen veriler, KTK ve G kuramındaki çaprazlanmış *bxmxp* (birey, madde, puanlayıcı) deseni ve birey ile puanlayıcıların yuvalandığı, maddelerin ise çaprazlandığı yuvalanmış desen olan *(p:b)xm* deseniyle analiz edilmiştir. Yapılan analizlerde SPSS 20.0 ve EduG 6.1 bilgisayar programlarından yararlanılmıştır.

BULGULAR

Çalışmada yer alan beş puanlayıcının ölçekte yer alan 15 madde doğrultusunda 50 bireye verdikleri puanların betimsel istatistikleri Tablo 1'de verilmiştir.

Tablo 1. Elde Edilen Puanların Beş Puanlayıcıya İlişkin Betimsel İstatistikleri

İstatistikler	***Puanlayıcılar***				
	1	***2***	***3***	***4***	***5***
Minimum	15	15	15	17	15
Maksimum	59	56	60	53	55
Ortalama	33.98	30.68	33.38	29.88	23.94
Std. Sapma	12.28	13.00	12.73	10.06	9.41
Çarpıklık	-0.023	0.312	-0.181	0.654	1.33
Basıklık	1.33	-1.25	-1.26	-0.437	1.44

Tablo 1. incelendiğinde, en yüksek ortalamanın birinci puanlayıcıya (33.98), en düşük ortalamanın ise beşinci puanlayıcıya (23.94) ait olduğu görülmektedir. Basıklık çarpıklık katsayılarının ise -1.5 ile +1.5 arasında değerler aldığı gözlemlenmektedir. Tabachnick ve Fidell (2013)'e göre bu katsayıların -1.5 ile +1.5 arasında olması verilerin normal dağılımına işaret etmektedir.

Birinci Alt Probleme İlişkin Bulgular

"KTK'ya göre; OSBP'nin sosyal karşılıklılık alt boyutundan elde edilen puanların puanlayıcılara ilişkin iç tutarlılık düzeyi ve beş farklı puanlayıcı arasındaki tutarlılık derecesi nedir?"

OSBP ölçeğindeki maddelerin kendi içinde tutarlı ölçme yapıp yapmadığının belirlenmesi amacıyla her bir puanlayıcı için iç tutarlılık güvenirliğini ifade eden Cronbach-alfa (α) katsayıları hesaplanmıştır. Ölçeğin "sosyal karşılıklılık" alt ölçeğinden elde edilen puanların her bir puanlayıcı için iç tutarlılığı Tablo 2'de verilmiştir.

Tablo 2. OSBP Ölçeği ile Yapılan Puanlamalara ait Cronbach-alfa (α) Değerleri

	Puanlayıcılar				
	1	***2***	***3***	***4***	***5***
Cronbach α	0.981	0.979	0.988	0.961	0.953

Tablo 2 incelendiğinde, her bir puanlayıcı için hesaplanan iç tutarlılık değerleri 0.95 ile 0.98 arasında değişmektedir. Daha sonra, puanlayıcılar arası tutarlılık derecesi, parametrik olmayan istatistiksel bir teknik olan Kendall'ın uyuşum katsayısı ile analiz edilmiştir. Analiz sonucunda uyuşum katsayısı 15 madde için 0.201 olarak bulunmuştur (X^2= 40.272, *sd*=4, *p*= .00 < . 05). Ayrıca, beş puanlayıcının 15 madde üzerinden verdikleri puanlar arasındaki korelasyon değerleri Tablo 3'te gösterilmiştir.

Tablo 3. Beş Puanlayıcının 15 Maddeye Verdikleri Puanlar Arasındaki Korelasyon Katsayıları

	1.Puanlayıcı	***2.Puanlayıcı***	***3.Puanlayıcı***	***4.Puanlayıcı***	***5.Puanlayıcı***
1.Puanlayıcı	-	0.606*	0.921*	0.722*	0.362*
2.Puanlayıcı		-	0.578*	0.727*	0.585*
3.Puanlayıcı			-	0.678*	0.398*
4.Puanlayıcı				-	0.535*
5.Puanlayıcı					-

**p<0.01*

Tablo 3 incelendiğinde, beş puanlayıcının 15 madde üzerinden verdiği puanlar arasındaki korelasyon katsayıları 0.362 ile 0.901 arasında değiştiği gözlenmektedir. Birinci ve üçüncü puanlayıcı, birinci ve dördüncü puanlayıcı son olarak da ikinci ve dördüncü puanlayıcı arasında yüksek; diğer puanlayıcılar arasında ise orta derecede anlamlı ilişki vardır.

Araştırmada korelasyon değerleri hesaplandıktan sonra, elde edilen puanların ortalamaları arasında farklılık olup olmadığı, ilişkili örneklemlerde tek faktörlü varyans analizi ile test edilmiş ve istatiksel olarak anlamlı bir farklılık bulunmuştur (*F*=8.261, p= .00<.05). Bu sonuç üzerine puanlayıcıların puan ortalamalarının ikili karşılaştırılması için çoklu karşılaştırma çalışması yapılmıştır. Çoklu karşılaştırma çalışmasında, beşinci puanlayıcı ile diğer tüm puanlayıcılar arasında ve birinci ile dördüncü puanlayıcı arasında anlamlı bir farklılık bulunmuştur.

İkinci Alt Probleme İlişkin Bulgular

"G kuramına göre; birey (b), madde (m) ve puanlayıcı (p) değişkenlerinin çapraz tasarlandığı bxmxp deseninin sonuçları nasıldır?"

İkinci alt problem için birey (*b*), madde (*m*) ve puanlayıcı (*p*) değişkenlerinin çapraz tasarlandığı *bxmxp* deseni ele alınmıştır. Bu aşamada ilk olarak, G çalışması sonucunda kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri incelenmiştir. G çalışması sonucunda kestirilen

varyans bileşenleri ve toplam varyansı açıklama yüzdeleri *b, m* ve *p* ana etkileri ile *bm, bp, mp* ve *bmp* ortak etkileri Tablo 4'te verilmiştir.

Tablo 4. *bxmxp* Desenine ait G Çalışması Sonucunda Kestirilen Varyans Bileşenleri ve Toplam Varyansı Açıklama Yüzdeleri

Varyans Kaynağı	*sd*	*Toplam Kareler*	*Kareler Ortalaması*	*Varyans*	*%*
b	49	1529.890	31.222	0.36541	39.9
m	14	76.195	5.442	0.01675	1.8
p	4	210.878	52.719	0.06412	7.0
bm	686	217.937	0.317	0.02522	2.8
bp	196	723.254	3.690	0.23323	25.5
mp	56	63.289	1.130	0.01877	2.1
bmp	2744	525.777	0.191	0.19161	20.9
Toplam	3749	3347.223			100

b:birey, m: madde, p:puanlayıcı

Tabloda verilen G çalışması sonucunda kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri incelendiğinde, en büyük oranda birey (*b*) ana etkisinin (%39.9), daha sonra birey *x* puanlayıcı (*bp*) ortak etkisinin (%25.5) toplam varyansı açıkladığı, en az ise %1.8 değer ile madde (*m*) ana etkisinin toplam varyansa katkı sağladığı görülmektedir.

Ana etkilere ait varyans bileşenleri ve toplam varyansı açıklama yüzdeleri incelendiğinde, birey (*b*) ana etkisi, toplam varyansın % 39.9'unu açıklamaktadır. Bu değer ile toplam varyansa en çok katkı sağlayan birey varyans bileşeni, bireylerin ölçülen özellik bakımından birbirlerinden farklılaştığını göstermektedir. Madde (*m*) ana etkisi ise toplam varyansın % 1.8'ini açıklamaktadır. Bu değer ile toplam varyansı açıklamada en son sırada yer almaktadır. Madde varyans bileşeninin toplam varyansı açıklama yüzdesinin düşük olması, her maddenin benzer güçlük düzeyinde olduğuna işaret etmektedir. Son olarak, puanlayıcı (*p*) ana etkisi, toplam varyansın % 7'sini açıklamakta olup; bu varyans bileşeni, ana etkilerden toplam varyansa en çok katkı sağlayan ikinci bileşen olarak gözlenmektedir. Bu durum, puanlayıcıların birbirleriyle çok benzer puanlama yapmadıklarını belirtmektedir.

Ortak etkilerden, birey *x* madde (*bm*) ortak etkisi, toplam varyansın % 2.8'ini açıklamaktadır. Birey *x* madde etkileşiminin toplam varyansa katkısının düşük olması, maddelerin güçlük düzeylerinin bireyden bireye farklılık göstermediğine işaret eder. Birey *x* puanlayıcı etkileşimi (*bp*) ölçülmek istenmeyen puanlayıcı etkisinin, ölçülmek istenen birey etkisini etkilemesi sonucu ortaya çıkan değişkenlik kaynağıdır. Birey x puanlayıcı ortak etkisi, toplam varyansın % 25.5'ini açıklayarak, toplam varyanstaki payı en yüksek ikinci değişkenlik kaynağıdır. Bu sonuç, bireylerin bağıl durumlarının bir puanlayıcıdan diğerine değiştiği anlamına gelmektedir. Madde x puanlayıcı (*mp*) ortak etkisi varyansı, % 2.1 toplam varyansı açıklama oranıyla, toplam varyans içindeki payı en düşük ikinci değişkenlik kaynağıdır. Toplam varyansı açıklama yüzdesinin düşük olması, maddelere verilen puanların puanlayıcıdan puanlayıcıya çok farklılaşmadığını, puanlayıcıların bireyleri maddelerin güçlük düzeyleri açısından tutarlı puanladıklarını göstermektedir.

Birey *x* madde *x* puanlayıcı (*bmp*) etkisine ait varyans bileşeni artık varyans olarak adlandırılıp, ölçme hatasını da barındırmaktadır. Tablo 4 incelendiğinde, ölçme objesi olan birey (*b*) ve birey x puanlayıcı (*bp*) ortak etkileşiminden sonra toplam varyansa en büyük getirisi olan üçüncü değişkenlik kaynağıdır. Artık varyans, toplam varyansın % 20.9'unu açıklamaktadır. Artık varyans bileşeninin yüksek çıkması; birey, madde ve puanlayıcı ortak etkisi ve/veya tesadüfi hataların büyük olabileceğinin bir göstergesidir.

OSBP ölçeğinin "sosyal karşılıklılık" alt boyutunda yer alan 15 madde için, beş puanlayıcının sayısının artırılıp azaltılarak puanlayıcı sayısının 3, 4, 6 ve 7 olduğu durumlara göre düzenlenen K çalışması senaryoları için kestirilen G ve Phi katsayılarına ilişkin değerler Tablo 5'te verilerek açıklanmıştır.

Tablo 5. *bxmxp* Desenine ait K Çalışmaları ile Puanlayıcı Sayılarının Artırılıp Azaltılmasıyla Yapılan Senaryolara Göre G ve Phi Katsayıları

Puanlayıcı Sayıları									
3		***4***		***5***		***6***		***7***	
G	***Φ***	***G***	***Φ***	***G***	***Φ***	***G***	***Φ***	***G***	***Φ***
0.813	0.774	0.852	0.819	**0.877**	**0.848**	0.895	0.869	0.908	0.885

Araştırmada kullanılan *bxmxp* deseninde, 50 bireyin beş puanlayıcı tarafından 15 madde doğrultusunda puanlanması ile elde edilen G katsayısı 0.877, Phi katsayısı ise 0.848 olarak bulunmuştur. Tablo 5'teki veriler incelendiğinde, puanlayıcı sayısı azaltıldığında G ve Phi katsayılarının azaldığı, puanlayıcı sayısı artırıldığında ise G ve Phi katsayılarının arttığı gözlemlenmektedir.

Üçüncü Alt Probleme İlişkin Bulgular

"G kuramına göre; birey (b) ve puanlayıcı (p) değişkenlerinin yuvalanmış, madde (m) değişkeninin ise çapraz tasarlandığı (p:b)xm deseninin sonuçları nasıldır?"

Üçüncü alt problem için birinci desende kullanılan aynı verilerle birey (*b*) ve puanlayıcı (*p*) değişkenlerinin yuvalanmış, madde (*m*) değişkeninin ise çaprazlandığı *(p:b)xm* deseni çalışılmıştır. Bu aşamada ilk olarak, G çalışması sonucunda kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri incelenmiştir. Daha sonra K çalışmalarında, puanlayıcı sayılarının 3, 4, 6 ve 7 olduğu senaryolara ilişkin G ve Phi katsayılarının değişimine bakılmıştır. *b, m, p:b, bm* ve *mp:b* değişkenleri için G çalışması sonucunda kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri Tablo 6'da verilmiştir.

Tablo 6. *(p:b)xm* Desenine ait G Çalışması Sonucunda Kestirilen Varyans Bileşenleri ve Toplam Varyansı Açıklama Yüzdeleri

Varyans Kaynağı	***Sd***	***Toplam Kareler***	***Kareler Ortalaması***	***Varyans***	***%***
b	4	1529.890	31.222	0.35259	39.1
m	14	76.195	5.442	0.02050	2.3
p:b	200	934.13	4.670	0.29735	33
bm	686	217.97	0.317	0.02146	2.4
mp:b	2800	589.06	0.210	0.21038	23.3
Toplam	3749	3347.223			100

b:birey, m: madde, p:puanlayıcı

Tablo 6'da verilen G çalışması sonucunda kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdelerine bakıldığında, en çok % 39.1 ile birey (*b*) ana etkisinin, daha sonra %33 ile birey ve puanlayıcı (*p:b*) ortak etkisinin toplam varyansı açıkladığı, en az ise %2.3 değer ile madde (*m*) ana etkisinin toplam varyansa katkı sağladığı görülmektedir. Bu sonuçlardan anlaşılacağı üzere, birey (*b*) ana etkisi, toplam varyansa en çok katkı sağlayan varyans bileşenidir. Bu durum, bireylerin ölçülen özellik bakımından farklılaştığına işaret etmektedir. Madde (*m*) ana etkisi, % 2.3 ile toplam varyansa getirisi en düşük olan bileşendir. Bu durum, ölçekte yer alan 15 maddenin zorluk-kolaylık düzeylerinin değişmediğini göstermektedir.

Her bir puanlayıcı, farklı bireyleri puanladığı için çalışmada birey değişkeniyle puanlayıcı değişkeni yuvalanmıştır. Buradaki $\sigma^2(b:p)$ varyans bileşeni, birey varyans bileşenini $\sigma^2(b)$ ve birey puanlayıcı ortak etkileşim varyans bileşenini $\sigma^2(bp)$ temsil etmektedir (Brennan, 2001). (*b:p*) için kestirilen

varyans değeri toplam varyansın % 33'ünü açıklayarak, toplam varyansa getirisi en yüksek ikinci değişkendir. Bu değerin yüksek olması, birey-puanlayıcı etkileşiminin farklılaştığı, bireylerin puanlarının bir puanlayıcıdan diğerine farklılık gösterdiğini belirtmektedir. Ortak etkilerden, birey x madde (*bm*) ortak etkisi ise toplam varyansın % 2.4'ünü açıklamaktadır. Birey x madde etkileşiminin toplam varyansa katkısının düşük olması, bireylerin bağıl durumlarının bir maddeden diğerine çok değişmediğini göstermektedir.

Tablo 6 incelendiğinde, ölçmenin objesi olan birey (*b*) ve birey puanlayıcı (*p:b*) etkileşiminden sonra toplam varyansa en büyük getirisi olan üçüncü değişkenlik kaynağı, artık varyans bileşenidir. Buradaki artık varyans bileşeni $\sigma^2(mp{:}b,e)$, madde puanlayıcı $\sigma^2(mp)$ varyans bileşenini ve birey madde puanlayıcı $\sigma^2(bmp,e)$ varyans bileşenini temsil etmektedir (Brennan, 2001). Artık varyans bileşeninin toplam varyansı açıklama yüzdesi % 23.3'tür. Artık varyans bileşeninin yüksek çıkması birey-madde-puanlayıcı ortak etkileşimi, madde-puanlayıcı ortak etkileşimi ve/veya tesadüfi hataların büyük olabileceğinin bir göstergesidir.

(p:b)xm yuvalanmış deseninde birey ölçme objesi olup; puanlayıcı sayılarının azaltılıp artırılmasıyla düzenlenen senaryolar için yapılan K çalışmaları sonucunda kestirilen G ve Phi katsayılarına ait değerler Tablo 7'de verilerek açıklanmıştır.

Tablo 7. *(p:b)xm* Desenine ait K Çalışmaları ile Puanlayıcı Sayılarının Artırılıp Azaltılmasıyla Yapılan Senaryolara göre G ve Phi Katsayıları

Puanlayıcı Sayıları									
3		***4***		***5***		***6***		***7***	
G	***Φ***	***G***	***Φ***	***G***	***Φ***	***G***	***Φ***	***G***	***Φ***
0.770	0.767	0.816	0.813	**0.846**	**0.844**	0.868	0.865	0.884	0.881

Farklı puanlayıcı senaryolarına göre G katsayısı, Phi katsayısına göre daha yüksek değerlere sahiptir. Düzenlenen senaryolarda puanlayıcı sayısı azaltıldığında G ve Phi katsayılarında azalma, puanlayıcı sayıları artırıldığında ise G ve Phi sayılarında artış görülmektedir.

Dördüncü Alt Probleme İlişkin Bulgular

"bxmxp ve (p:b)xm desenlerinden elde edilen G çalışması sonuçlarının karşılaştırılması nasıldır?

Birey (*b*), madde (*m*) ve puanlayıcı (*p*) değişkenlerinin çaprazlandığı *bxmxp* deseni ile birey (*b*) ve puanlayıcı (*p*) değişkenlerinin yuvalanmış, madde (*m*) değişkeninin ise çaprazlanmış olduğu *(p:b)xm* deseninden elde edilen G çalışması sonuçları karşılaştırıldığında, bireylere ait varyans bileşeninin çapraz desende toplam varyansın % 39.9'unu, yuvalanmış desende ise toplam varyansın % 39.1'ini açıkladığı görülmektedir. Her iki desende de toplam varyansa en çok getirisi olan birey varyans bileşeni, bireylerin sosyal becerileri bakımından farklılaştığını göstermektedir.

Çapraz desende, puanlayıcı (*p*) varyans bileşeninin toplam varyansı açıklama oranı % 7, birey x puanlayıcı (*bp*) ortak etkileşimine ait varyans bileşeninin ise toplam varyansı açıklama oranı % 25.5'tir. Yuvalanmış desende ise puanlayıcı ana etkisine ve birey x puanlayıcı ortak etkisine ait varyans bileşenleri ayrı ayrı değil, (*b:p*) değişkeni altında ortak değerlendirilmektedir. (*b:p*) değişkeninin toplam varyansı açıklama oranı ise % 33'tür.

bxmxp ve *(p:b)xm* desenlerinde puanlayıcı sayılarının artırılıp azaltılmasıyla yapılan K çalışmaları karşılaştırıldığında, her iki desende de puanlayıcı sayıları artırıldığında G ve Phi katsayıları artarken, puanlayıcı sayıları azaltıldığında G ve Phi katsayıları azalmaktadır. Ancak, çapraz desende kestirilen G ve Phi katsayılarının yuvalanmış desende kestirilen G ve Phi katsayılarından daha yüksek olduğu görülmektedir.

SONUÇLAR ve TARTIŞMA

KTK kapsamında ilk olarak iç tutarlılık düzeyinin belirlenmesi için Cronbach-alfa güvenirlik katsayısı oldukça yüksek (α=0.95 - α=0.98) çıkmıştır. Cronbach-alfa değerlerinin 0.90'dan yüksek olması ölçeğin iç tutarlılığın oldukça yüksek olduğunu göstermektedir. Puanlamada kullanılan maddelerin OSBP ölçeğinin sosyal karşılıklılık alt boyutunda yer alan maddelerin tek bir yapıyı ölçmesi, bu değerlerin oldukça yüksek çıkmasını açıklar niteliktedir. Elde edilen Cronbach-alfa katsayılarına göre, puanlayıcıların verdiği puanların kendi içinde tutarlı olduğu yorumu yapılabilir. Beş puanlayıcı arasındaki uyum düzeyinin belirlenmesi için hesaplanan Kendall'ın uyuşum katsayısı, beklenen düzeyden (en az 0.80) düşük bulunmuştur. Bu bulgu, dereceleme ölçekleriyle gerçekleştirilen Deliceoğlu ve Çıkrıkçı-Demirtaşlı (2012)'nın ve Öztürk (2011)'ün çalışmalarında elde ettikleri düşük düzeydeki Kendall'ın uyuşum katsayıları bulgularıyla tutarlılık göstermektedir. Howell (2009)'e göre elde edilen bu sonuç doğrultusunda, puanlayıcılar arası uyum olmadığı yorumu yapılabilir. Ayrıca, her bir puanlayıcının verdiği puanlar ile diğer puanlayıcıların verdikleri puanlar arasındaki ilişkiler, Pearson momentler çarpım korelasyon katsayısı ile hesaplanmıştır. Ancak, Pearson çarpım moment korelasyon katsayısının ortalamadan bağımsız olması sebebiyle bu katsayı, puanlayıcıların verdikleri puanlar arasındaki benzerlik ve farklılıklar hakkında bilgi verememektedir. Bu nedenle, Goodwin (2001) puanlayıcılar arası tutarlılık test edilirken, korelasyonla birlikte ortalamaların da karşılaştırılmasını önermektedir. Bu doğrultuda gerçekleştirilen tek faktörlü varyans analizi sonucu elde edilen anlamlı farklılık, Kendall'ın uyuşum katsayısı ile elde edilen puanlayıcıların puanlamalarının birbirleriyle paralellik göstermediği yorumunu desteklemektedir.

G kuramı kapsamında iki farklı desen ile çalışılmıştır. Bunlardan ilki birey, madde ve puanlayıcı değişkenlerinin çaprazlandığı *bxmxp* desenidir. G çalışması sonucunda kestirilen varyans bileşenleri ve toplam varyansı açıklama yüzdeleri incelendiğinde, toplam varyansı açıklama yüzdesi en yüksek olan değişkenlik kaynağın birey (*b*) ana etkisi olduğu görülmüştür. Birey ana etkisinin, ölçmenin objesi olması nedeniyle bu durum istenilen bir durumdur (Güler vd., 2012). Birey puanlayıcı (*bp*) ortak etkisine ait varyans bileşeni, toplam varyansa katkısı en yüksek olan ikinci değişkenlik kaynağıdır. Bu değerin yüksek olması, bireylerin bağıl durumlarının bir puanlayıcıdan diğerine değiştiğini göstermiştir. Bu nedenle puanlayıcılar arası uyumun düşük olduğu yorumu yapılabilir. Puanlayıcı (*p*) değişkenine ait varyans bileşeninin, ana etkiler arasında ölçme objesinden sonra en yüksek varyansa sahip bileşen olması, bu durumu destekler niteliktedir. G kuramı kapsamında çalışılan bir diğer desen, birey ve puanlayıcıların birbiriyle yuvalandığı, maddelerin ise her ikisiyle çaprazlandığı *(p:b)xm* yuvalanmış desenidir. Bu desende yapılan G çalışması sonucunda toplam varyansa katkı sağlayan beş varyans bileşeninden, yüzdesi en yüksek olan birey (*b*) değişkenine ait varyans bileşenidir. Birey değişkeniyle puanlayıcı değişkeninin yuvalandığı (*p:b*) değişkenine ait varyans değeri, toplam varyansa en çok katkıyı sağlayan ikinci bileşendir. Bu değerin yüksek olması, birey-puanlayıcı etkileşiminin farklılaştığı, bireylerin puanlarının bir puanlayıcıdan diğerine farklılık gösterdiğini belirtmektedir. *bxmxp* çapraz ve *(p:b)xm* yuvalanmış desenlerinde yapılan G çalışmalarıyla elde edilen varyans ve toplam varyansı açıklama yüzdeleri karşılaştırıldığında genel olarak, çapraz ve yuvalanmış desenden elde edilen bulgulara göre kestirilen varyans değerleri arasında paralellik olduğu görülmüştür. Bu bulgu, Alkan (2013)'ün ve Nalbantoğlu (2009)'un çalışma bulguları ile desteklenmektedir. Her iki desende elde edilen sonuçlar birbirleriyle tutarlılık gösterse de, çapraz desenin, birey-puanlayıcı ve madde-puanlayıcı ortak etkileri için de bilgi üretmesi sebebiyle daha detaylı bilgi verdiği yorumu yapılabilir.

KTK ve G kuramı kapsamında yapılan analizler sonucu puanlayıcılar arası uyumun düşük çıkmasının nedenlerinin araştırılması için puanlayıcılarla görüşmeler yapılmıştır. Görüşmelerden elde edilen bilgiler doğrultusunda; puanlayıcıların mesleki alanlarının, sosyal becerileri farklı şekilde puanlamalarına neden olduğu sonucuna varılmıştır. Örneğin, aralarında yüksek ilişki bulunan birinci puanlayıcı (psikolog) ve üçüncü puanlayıcı (sosyal psikolog, davranış terapisti) değerlendirmelere psikolog olarak yaklaştıklarını ve bunun puanlamalara yansımış olabileceğini ifade etmişlerdir. Ayrıca, dereceleme ölçeklerinde güvenirliği tehlikeye sokan dikkatsizlik, kişisel yanlılık, merkeze kayma etkisi, halo etkisi gibi unsurların puanlayıcılar arası uyumu etkilemiş olabileceği

düşünülmektedir. Bu sebeple, farklı puanlayıcıların yer aldığı çalışmalar düzenlenirken, puanlayıcıların nasıl puanlama yapacaklarına dair bilgilendirme eğitimleri verilerek, ön bir uygulama ile puanlayıcıların farklı anladıkları ölçütler/maddeler ve nedenleri araştırılıp, bu farklılıklar giderilerek asıl uygulama gerçekleştirilebilir.

KTK'da sadece bağıl değerlendirmeler için güvenirlik hesaplarken; G kuramında hem bağıl değerlendirme için G-katsayısı hem de mutlak değerlendirmeler için Phi-katsayısı kestirilmektedir (Crocker & Algina, 1986; Shavelson & Webb, 1991; Brennan, 2001; Goodwin, 2001). Düzenlenen farklı puanlayıcı senaryolarında çapraz desende kestirilen G ve Phi katsayıları, yuvalanmış desende kestirilen G ve Phi katsayılarından daha yüksek kestirilmiştir. Ayrıca, gerek çapraz desende gerekse yuvalanmış desende tüm senaryo durumlarına göre G katsayıları Phi katsayılarından daha yüksek çıkıştır. Benzer çalışmalarda (Nalbantoğlu, 2009; Alkan, 2013) da G katsayılarının Phi katsayılarından yüksek kestirildiği görülmektedir. Bu durum, mutlak ve bağıl hata varyansları arasındaki farklılıktan kaynaklandığı için beklenen bir durumdur. Tanımı gereği mutlak hata varyansı, bağıl hata varyansından daha yüksek bir değere sahip olduğu için, Phi katsayısı G katsayısına göre daha düşük bir değere sahip olur. Shavelson ve Webb (1991)'e göre "yüksek" güvenirlik sağlayabilmek için G ve Phi katsayılarının en az 0.80 olması gerektiğini belirtmişlerdir. Yapılan K çalışmalarına göre, bu değerin sağlanabilmesi için her iki desende de puanlayıcı sayısının en az dört olması gerekmektedir. Bu durumda, zaman ve işgücü göz önünde bulundurularak yüksek güvenirlikli en ekonomik uygulamanın dört puanlayıcı ile yapılabileceği söylenebilir.

Elde edilen sonuçlar incelendiğinde, klasik test kuramında puanlayıcılar arası güvenirliğin belirlenebilmesi için birden fazla analize gerek duyulmaktadır. Genellenebilirlik kuramı ise birden fazla hata kaynağını aynı anda göz önünde bulundurarak güvenirlik çalışmalarının tek bir analiz ile yapılmasını sağlamaktadır. Ayrıca, klasik test kuramı kapsamında hesaplanan Kendall'ın uyuşum katsayısı toplam puanlar üzerinden analiz yaparken, G kuramı maddeler bazında analiz yaptığı için daha detaylı bilgi vermektedir. Bu sebeple, puanlayıcılar arası güvenirlik analizlerinde G kuramı tercih edilebilir. KTK'da sadece bağıl değerlendirmeler için güvenirlik hesaplanırken; G kuramında hem bağıl değerlendirme hem de mutlak değerlendirmeler için güvenirlik katsayıları kestirilmektedir. Bununla birlikte, G kuramı karar çalışmaları ile farklı senaryolar için güvenirlik kestirimleri yaparak, daha sonraki çalışmalar için daha yüksek güvenirlik ve daha düşük hata içeren sonuçlar elde edebilecek senaryolar üretebilmektedir. Bu sebeplerden ötürü, G kuramının KTK'na göre daha avantajlı bir kuram olduğu yorumu yapılabilir.

KAYNAKÇA

Alkan, M. (2013). *PISA 2009 okuma becerileri açık uçlu sorularının puanlanmasında genellenebilirlik kuramındaki farklı desenlerin karşılaştırılması* (Doktora Tezi, Hacettepe Üniversitesi, Ankara). https://tez.yok.gov.tr/UlusalTezMerkezi/ adresinden edinilmiştir.

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

Bellini, S., & Hopf, A. (2007). The development of the autism social skills profile: A preliminary analysis of psyhometric properties. *Focus on Autism and Other Developmental Disabilities, 22*(2), 80–87.

Brennan, R. L. (2001*). Generalizability theory*. New York: Springer-Verlog.

Cardinet, J., Johnson, S., & Pini, G. (2009). *Applying generalizability theory using eduG (quantitative methodology series)*. New York, London: Routledge.

Combs, H. L., & Slaby, D. A. (1977). Social skills training with children. In B. B. Lahey & A. E. Kazdin (Eds.), *Advances in clinical childpsy- chology* (Vol. 1, pp. 161-201). New York: Plenum.

Crocker, L. M., & Algina, L. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Deliceoğlu, G. ve Çıkrıkçı-Demirtaşlı, N. (2012). Futbol yetilerine ilişkin dereceleme ölçeğinin güvenirliğinin genellenebilirlik kuramına ve klasik test kuramına dayalı olarak karşılaştırılması. *Spor Bilimleri Dergisi, 23*(1), 1-12.

Demir, Ş. (2009). *Otizmli çocukların sosyal becerilerinin farklı değişkenler açısından değerlendirilmesi* (Yüksek Lisans Tezi, Ankara Üniversitesi, Ankara). https://tez.yok.gov.tr/UlusalTezMerkezi/ adresinden edinilmiştir.

Eason, S. H. (1989, November). *Why generalizability theory yields beter results than classical test theory.* Mid-South Educational Research Association Annual Meeting. Little Rock, AR, USA

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Psychical Education and Exercises Science, 5*(1), 13-34.

Girli, A., & Atasoy, S. (2010). Examining the effectiveness of social skills training program based on cognitive process approach among inclusion students with autism. *Elementary Education Online, 9*(3), 990-1006.

Güler, N. (2008). *Klasik test kuramı genellenebilirlik kuramı ve Rasch modeli üzerine bir araştırma* (Doktora Tezi, Hacettepe Üniversitesi, Ankara). https://tez.yok.gov.tr/UlusalTezMerkezi/ adresinden edinilmiştir.

Güler, N. ve Gelbal S. (2010). Açık uçlu matematik sorularının güvenirliğinin klasik test kuramı ve genellenebilirlik kuramına göre incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri Dergisi, 10*(2), 989–1019.

Güler, N. (2011). Rasgele veriler üzerinde genellenebilirlik kuramı ve klasik test kuramına göre güvenirliğin karşılaştırılması. *Eğitim ve Bilim, 36*(162), 225-234.

Güler, N., Uyanık, G. K. ve Teker, G. T. (2012). *Genellenebilirlik kuramı*. Ankara: Pegem Akademi.

Hall, J. A., Schlesinger, D. J., & Dieen, J. P. (1997). Social skills training in group with developmentally disabled adults. *Research on Social Work Practice, 7*(2), 187-201.

Howell, D. C. (2009). *Statistical methods for psychology.* (7th Edition). USA: Thomson Learning Academic Research Center.

Irmak, T. Y., Sütçü, S. T., Aydın, A. ve Sorias, O. (2007). Otizm davranış kontrol listesinin (abc) geçerlik ve güvenirliğinin incelenmesi. *Çocuk ve Gençlik Ruh Sağlığı Dergisi, 14*(1), 13-23

Kieffer, K. M., (1998,April). *Why generalizability theory is essential and classical test theory is often inadequate?* Paper presented at the annual meeting of the Southwestern Psychological Association, New Orleans, LA, USA.

Kırcaali-İftar, G. (2012). Otizm spektrum bozukluğuna genel bakış. E. Tekin-İftar (Ed). *Otizm spektrum bozukluğu olan çocuklar ve eğitimleri* (s. 17-46). Ankara: Vize.

Lord, F. M., & Novick, R. M. (1968). *Statistical theories of mental test scores.* California: Addison-Wesley Publishing Company.

Mushquash, C., & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods, 38*(3), 542-547.

Nalbantoğlu, F. (2009). *Performans ölçümlerinde genellenebilirlik kuramıyla farklı desenlerin karşılaştırılması* (Yüksek Lisans Tezi, Hacettepe Üniversitesi, Ankara). https://tez.yok.gov.tr/UlusalTezMerkezi/ adresinden edinilmiştir.

Özdemir, O., Diken, İ. H., Diken, Ö. ve Şekercioğlu, G. (2013). Otizm davranış kontrol listesi (autism behavior checklist-ABC) modifiye edilmiş Türkçe versiyonunun geçerlik ve güvenirlik çalışması: Pilot uygulama sonuçları. *International Journal of Early Childhood Special Education (INT-JECSE), 5*(2), 168-186.

Öztürk, M. E. (2011). *Voleybol becerileri gözlem formu ile elde edilen puanların genellenebilirlik ve klasik test kuramına göre karşılaştırılması* (Doktora Tezi, Hacettepe Üniversitesi, Ankara). https://tez.yok.gov.tr/UlusalTezMerkezi/ adresinden edinilmiştir.

Polat-Demir, B. (2016). Vee diyagramından elde edilen puanların güvenirliğinin klasik test kuramı ve genellenebilirlik kuramına göre incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 7*(2), 419-431.

Rae, G., & Hyland, P. (2001). Generalisability and classical test theory analyses of Koppitz's Scoring System for human figure drawing. *British Journal of Educational Psychology, 71*, 369-382.

Shavelson, R. J., & Webb, N. M. (1991*). Generalizability theory: A primer*. New-York: Springer.

Sucuoğlu, B. ve Demir, Ş. (2017). Bağlamsal Değerlendirme Envanteri: Otizmli bireylerin problem davranışlarının bağlamsal değişkenleri. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Özel Eğitim Dergisi, 8*(2), 209-224.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson.

Turgut, M. F., & Baykul, Y. (2014). *Eğitimde ölçme ve değerlendirme* (6. Baskı). Ankara: Pegem Akademi.

Yelboğa, A. ve Tavşancıl, E. (2010). Klasik test ve genellenebilirlik kuramına göre güvenirliğin bir iş performansı ölçeği üzerinde incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri, 10*(3), 1825-1854.

Yıldıztekin, B. (2014). *Klasik test teoremi ve genellenebilirlik kuramından puanlayıcılar arası tutarlılığın farklı yöntemlere göre karşılaştırılması* (Yüksek Lisans Tezi, Hacettepe Üniversitesi, Ankara). https://tez.yok.gov.tr/UlusalTezMerkezi/ adresinden edinilmiştir.

EXTENDED ABSTRACT

Introduction

The term social skills refers to the ability to communicate with others in an acceptable way in a social environment (Combs & Salaby, 1977). Inadequacies in these skills cause people to experience challenges in communicating in social and educational life. One of the special groups with special needs that are highly affected by this situation is those with autism spectrum disorder (Özdemir, Diken, Diken & Şekercioğlu, 2013). In the studies conducted in special education, the evaluations of the autistic people are usually done by scales of their families and teachers, because they are not competent enough to evaluate themselves. In such evaluations, rater-based errors can be encountered such as carelessness, personal bias, etc. (Turgut and Baykul, 2014). For this reason, it is very important to test the reliability between the raters while doing evaluation in the field of special education so that reliable results can be obtained.

Method

In this study, interrater reliability was compared based on classical test theory (CTT) and generalizability theory (GT) according to the scores which were obtained from five raters' ratings with Autism Social Skills Profile (ASSP). Levels of reliability coefficients obtained from CTT and different designs in GT formed by five raters' jointly and alternatively ratings were determined and which theory presented more information was tried to be specified. The research group consisted of 50 children and youths with autism who were being trained in a special education and rehabilitation center in Ankara and five raters rated them through social reciprocity sub-scale under Autism Social Skills Profile. The raters scored independently. The obtained data were analyzed in GT with the crossed design *pxixr* (person, item, rater) in which people were scored by all raters through all items and with the nested design *(r:pxi)* in which people were scored by different raters through all items. Additionally, Cronbach Alpha (α) coefficient for internal consistency, Kendall's concordance coefficient for interrater reliability and correlation coefficients of five raters' scores were calculated in CTT and it was investigated whether there was a difference among the means of raters' scores with F test. SPSS 20.0 and EduG 6.1 computer programmes were used in the analyses.

Results and Discussion

According to the obtained Cronbach Alpha coefficients (α = 0.95 - α = 0.98), it can be interpreted that the scores given by the raters are internally consistent. The Kendall's coefficient of concordance was found to be lower than the expected level (at least 0.80). This finding is consistent with the low level Kendall's coefficient of concordance findings of the studies conducted by Deliceoğlu and Çıkrıkçı-Demirtaşlı (2012) and Öztürk (2011) with grading scales. In addition, Pearson moments product correlation coefficients were found between 0.362 and 0.901. The significant difference obtained from the one-factor variance analysis supports the interpretation that the scores of the raters obtained by Kendall's coefficient of concordance are not parallel to each other.

Two different designs were used for the GT. In the *pxixr* design; it was found that the variance source having the highest total variance explanatory percentage is the main effect person (*p*). The variance component of the common effect; person x rater (*pr*) is the second variance source that brings the greatest contribution to the total variance. The high level of this value indicates that the relative status of the persons varies by raters. For this reason, it can be interpreted that the concordance between the raters is low. In the nested *(r:p)xi* design, the variance component of the variable person (*p*) has the highest percentage contributing to the total variance among five variance components. The variance value of the variable in which the variables of the rater and person (*r:p*) are nested is the second component that brings the greatest contribution to the total variance. The high level of this value indicates that the person-rater interaction differentiates, and that scores differ from one rater to another. Comparing the variance and total variance explanatory percentages obtained from the G studies conducted in both designs, it was generally found that the variance

values estimated according to the findings obtained from the crossed and nested designs are parallel with each other. This finding supports the findings of Alkan (2013) and Nalbantoğlu (2009).

G and Phi coefficients estimated in the crossed design in different rater scenarios are estimated much higher than those estimated in the nested design. In addition, G coefficients are found higher than the Phi coefficients in all scenarios regardless of the crossed or nested design. In similar studies (Nalbantoğlu, 2009; Alkan, 2013), it can be observed that the G coefficients are estimated higher than the Phi coefficients. Since the absolute error variance by definition has a value higher than the relative error variance, Phi coefficient has a lower value than G coefficient. According to Shavelson and Webb (1991), G and Phi coefficients should be at least 0.80 in order to provide "high" reliability. According to D studies, the number of raters on both designs must be at least four in order to achieve this value. In this case, it can be said that the most economical application with high reliability considering the time and labour force can be done with four raters.

When the obtained results are examined, in CTT multiple analyses are required in order to determine the inter-rater reliability. GT, on the other hand, ensures reliability studies with a single analysis taking into account multiple error sources at the same time. In addition, while the reliability is calculated only for the relative evaluations in the CTT; reliability coefficients for both relative and absolute evaluations are estimated in GT. Moreover, GT makes reliability estimations for different scenarios with decision-making studies and in this way, it can produce scenarios including higher reliability and lower errors for further studies. For these reasons, it can be said that GT is more advantageous theory than CTT.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(2);216-238

Eğitimde Ölçme ve Değerlendirme Kongrelerinde Sunulan Bildirilerin Doküman Analizi Yöntemi ile İncelenmesi*

The Investigation of the Papers Presented in Measurement and Evaluation in Education and Psychology Congresses with Document Analysis

Mahmut Sami KOYUNCU ** Mehmet ŞATA *** İsmail KARAKAYA ****

Öz

Bu araştırmanın amacı Türkiye'de Eğitimde ve Psikolojide Ölçme ve Değerlendirme (EPOD) Kongrelerinde sunulan bildirilerin eğilimlerinin nasıl olduğunun belirlenmesidir. Bu kapsamda 2008, 2010, 2012 ve 2014 yıllarında Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongrelerinde sunulmuş olan bildirilerin konu temaları, anahtar kelimeleri, yazarların kurumları ve unvanları, araştırma türü, araştırma modeli/deseni, örneklem büyüklüğü, kullanılan veri türü ve veri toplama aracı incelenmiştir. Bu araştırma, Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresinde yapılan çalışmaların eğilimlerinin nasıl olduğunun betimlenmesi ve var olan durumun yorumlanması amaçlandığı için, araştırma deseni nitel araştırma türlerinden durum çalışmasıdır. Veri toplama yöntemi olarak doküman analizi kullanılmıştır. EPOD kongrelerinde sunulan bildirilerin genel olarak "Ölçek Geliştirme ve Uyarlama" , "Eğitim araştırmalarında kullanılan istatistiksel yönelimler" ve "Geniş ölçekli sınavlar, seçme ve yerleştirme sınavları, yüksek riskli sınavlar" konu temalarında yoğunlaştığı belirlenmiştir. "ÖSYM Uygulamaları" ve "Etik" temalarında çok fazla çalışma yapılmadığı belirlenmiştir. Sunulan bildiri türlerinin daha çok nicel araştırma olduğu, buna bağlı olarak tarama ve betimsel araştırma modellerinin kullanıldığı belirlenmiştir. Bildirilerde en çok "ölçme ve değerlendirme", "güvenirlik", "geçerlik" ve anahtar kelimelerinin kullanıldığı tespit edilmiştir.

Anahtar Kelimeler: Bildiri, doküman analizi, EPOD, ölçme ve değerlendirme.

Abstract

This study aims to determine the trends of the papers presented in Measurement and Evaluation in Education and Psychology (MEEP) Congresses in Turkey. For that purpose, authors` title and where they work, the subjects, key words, research methods, sample size, data type and data collection tool of the papers presented in MEEP Congresses in 2008, 2010, 2012 and 2014 were investigated. In the study, since describing how the trends in the papers presented orally in MEEP congresses and interpreting the existing situation are aimed, the case study which is one of the qualitative research designs was selected as a research design. Data analysis was done using document analysis. The subjects of the papers were generally centered on the following topics: "Scale Development and Implementation", "statistics used in educational research", "standardized tests, selection and placement tests, high-stakes tests". However, "student selection and placement center's implementations" and "ethics" are the subjects that were studied rarely. It is determined that the presented papers were generally in quantitative research methods, especially survey and descriptive research models were used. It is also determined that the most frequently used key words were reliability, validity and measurement and evaluation in these papers.

Keywords: Papers, document analysis, MEEP, measurement and evaluation.

*Bu çalışma V. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi'nde (1-3 Eylül 2016, Antalya) sözlü bildiri olarak sunulmuştur.

** Arş. Gör., Gazi Üniversitesi, Gazi Eğitim Fakültesi, Ankara-Türkiye, e-posta: ms_koyuncu@hotmail.com, ORCID ID: https://orcid.org/0000-0002-6651-4851

*** Arş. Gör., Ağrı İbrahim Çeçen Üniversitesi, Eğitim Fakültesi, Ağrı-Türkiye, e-posta: mehmetwsata@gmail.com, ORCID ID: https://orcid.org/0000-0003-2683-4997

**** Doç. Dr., Gazi Üniversitesi, Gazi Eğitim Fakültesi, Ankara-Türkiye, e-posta: ikarakaya2002@gmail.com, ORCID ID: https://orcid.org/0000-0003-4308-6919

Geliş Tarihi: 12.08.2017
DOI: 10.21031/epod.334292 *Kabul Tarihi: 12.06.2018*

GİRİŞ

Eğitim sisteminin temel ögeleri arasında yer alan öğrenme sürecinin en genel amacı, eğitimin tanımında yer alan istendik davranışları öğrencilere kazandırmaktır. Öğrenme sürecinde veya sonunda öğrencinin akademik başarı düzeyini belirlemek ve bu düzeyin istenilen oranda olup olmadığına karar vermek için ölçme ve değerlendirme büyük öneme sahiptir (Bayram, 2011). Ölçme ve değerlendirme etkinliği ile eğitim sistemin kontrolü sağlanmakta ve ilgili kişilere gerekli geri bildirim verilebilmektedir.

Ölçme, ölçülen niteliklerin aralarındaki büyüklük ve küçüklük gibi ilişkileri koruyarak bu niteliklerin sayı ve sembollerle ifade edilmesidir (Tan, 2012). Genel anlamda ölçme, herhangi bir niteliği, o niteliğe uygun araçlarla gözlemek ve sonucu araca uygun bir birimle ifade etmektir. Değerlendirme ise Tan (2012)'a göre ölçümlerin ölçüt veya ölçütlerle kıyaslanarak bir karara varma işidir. Bu süreç içerisinde, bireylerin davranışlarındaki değişimin ölçülmesi ve ölçme sonuçlarının değerlendirilmesi bireylerin gelişim düzeylerini belirlemek açısından önem kazanmaktadır. Değerlendirme ile öğrencilerin öğrenme süreçleri izlenerek gerektiğinde kullanılan etkinlikler, yöntemler ve teknikler değiştirilmelidir (Demirel ve Şahinel, 2006).

Öğretme-öğrenme sürecinin önemli değişkenlerinden biri de yapılan ölçme-değerlendirme etkinlikleridir. Öğretim faaliyetlerinin etkililiğinin belirlenmesi süreç içerisinde yapılan çeşitli ölçme-değerlendirme çalışmaları ile belirlenir. Eğitim ortamında bir öğrencinin güçlü ya da zayıf yanlarının ya da öğrenme eksikliklerinin belirlenmesi; sınıfın bir ders ya da üniteyle ilgili olarak hangi konuları öğrenemediği, hangi davranışlarının geliştirilmesi gerektiği; sınıftaki çeşitli gruplar arasında (kız-erkek gibi) çeşitli değişkenler açısından fark olup olmadığı; sınıf düzeyindeki gerekli kazanımların kazanılıp kazanılmadığı ve başarı oranları; öğrencilerin tek tek veya sınıf olarak duyuşsal özelliklerinin durumu ve süreç içindeki değişimi; sınıflar arası başarı, okulun hedeflerine ulaşılıp ulaşılmadığının belirlenmesi gibi durumlarda ölçme ve değerlendirme gereklidir (Erkuş, 2014).

Doküman analizi çoğunlukla diğer araştırma yöntemlerini tamamlayıcı nitelikte iken, aynı zamanda tek başına bir yöntem olarak da kullanılabilmektedir. Doküman analizi hem basılı hem de elektronik belgeleri gözden geçirmek ve değerlendirmek için sistematik bir prosedürdür (Bowen, 2009). O'Leary (2017) ise doküman analizini birincil araştırma verisi kaynağı olarak çeşitli yazılı metin biçimlerinin toplanması, incelenmesi, sorgulanması ve analiz edilmesini amaçlayan bir araştırma aracı olarak açıklamaktadır.

Dokümanlar, bir araştırma girişiminin bir parçası olarak çeşitli amaçlara hizmet edebilirler. Belgesel materyalin beş özel fonksiyonunu vardır. İlk olarak dokümanlar araştırmada katılımcıların içinde bulunduğu bağlam hakkında veri sağlayabilir. Belgeler, geçmiş olaylara tanıklık ederek, olayların arka plan bilgisini ve tarihsel iç görüsünü sağlayabilirler. Bu tür bilgiler, araştırmacıların çok özel konuların tarihsel köklerini anlamasına yardımcı olabilir ve şu anda incelenmekte olan olaylara etki eden koşulları gösterebilirler. İkinci olarak, belgelerde yer alan bilgiler, sorulması gereken bazı sorulara ve araştırmanın bir parçası olarak gözlemlenmesi gereken durumlara işaret edebilir. Üçüncü olarak, dokümanlar ek araştırma verileri sağlarlar. Dokümanlardan elde edilen bilgiler, bir veri tabanı için değerli ek bilgiler sağlayabilir. Dördüncüsü, dokümanlar değişimi ve gelişimi izlemenin bir yolunu sağlarlar. Belirli bir belgeye ait çeşitli taslaklara erişilebildiğinde, araştırmacı değişiklikleri belirlemek için bunları karşılaştırabilir. Beşinci olarak ise, dokümanlar, bulguları doğrulamak veya diğer kaynaklardan elde edilen kanıtları doğrulamak için bir yol olarak kullanılmaktadır. Özetle, dokümanlar arka plan ve içerik, sorulacak ek sorular, ek veriler ve diğer veri kaynaklarından elde edilen bulguların doğrulanmasını sağlar (Bowen, 2009).

Doküman analizinin avantajları ve sınırlılıkları bulunmaktadır. Doküman analizinin avantajlarına bakıldığında etkili bir yöntem olduğu görülmektedir. Doküman analizi, diğer araştırma yöntemlerinden daha az zaman almaktadır ve bu nedenle daha verimli olmaktadır. Kısacası, veri toplama yerine veri seçimi gerektirmektedir. İkinci avantajı ise kullanılabilirliktir. Yani, birçok belge, özellikle İnternet'in ortaya çıkmasından bu yana kamusal alanda olup, yazarların izni olmadan edinilebilmesidir. Bu doküman analizini nitel araştırmacılar için cazip bir seçenek haline

getirmektedir. Diğer bir avantajı da maliyet etkililiğidir. Yani, doküman analizi, diğer araştırma yöntemlerinden daha az maliyetlidir ve yeni verilerin toplanması mümkün olmadığında genellikle tercih edilen yöntemlerden biridir. Veriler (belgelerde) zaten toplanmıştır; geriye kalan ise, değerlendirilecek belgelerin içeriği ve kalitesidir. Diğer bir avantajı ise kararlılığıdır. Başka bir değişle araştırmacı tarafından birden çok kez okunabilmekte ve incelenebilmektedir. Ayrıca araştırma sürecinde değişmeden kalır ve araştırmacının etkisi yoktur (Bowen, 2009).

Doküman analizinin avantajlarının yanında bazı sınırlılıkları da mevcuttur. Bunlardan birincisi ayrıntının yeterince verilmemiş olabilmesidir. Başka bir değişle, dokümanlar araştırmadan ziyade başka amaçlara hizmet etmek için oluşturulmaktadır; yani bunlar bir araştırma gündeminden bağımsız olarak oluşturulmuştur. Sonuç olarak, genellikle bir araştırma sorusuna cevap vermek için yeterli ayrıntı sağlayamayabilirler. İkinci sınırlılığı ise ulaşılabilirliktir. Yani, belgelere erişim kasıtlı olarak engellenmiş olabilir. Diğer bir sınırlılığı ise önyargılı seçiciliktir. Belgelerin eksik bir şekilde toplanması 'önyargılı seçiciliği' göstermektedir (Yin, 1994).

Son yıllarda dergilerde yayımlanan makalelerin ve araştırma raporlarının metodolojisinin bir parçası olan doküman analizi sayısında bir artış olmuştur. Alan yazına bakıldığında eğitim alanında yapılan tezlerin ve makalelerin incelendiği çalışmaların var olduğu (Erdem, 2011; Küçükoğlu ve Ozan, 2013; Tavşancıl, vd., 2010; Yaşar ve Papatğa, 2015) görülmektedir. Eğitim bilimlerinin alt alanı olan eğitimde ölçme ve değerlendirme alanında doküman analizi yöntemi kullanılarak yüksek lisans ve doktora tez çalışmalarının incelendiği araştırmaların (Ayva, Ceyhan ve Doğan, 2015; Şenyurt ve Özkan, 2017) var olduğu, ayrıca ölçme ve değerlendirme ile ilişkili olan ölçek geliştirme ve uyarlama ile ilgili makalelerin incelendiği araştırmaların (Acar-Güvendir ve Özer-Özkan, 2015; Tavşancıl, Güler ve Ayan, 2014, Worthington & Whittaker, 2006)) olduğu görülmekte fakat ilgili alanda sunulan bildirilerin incelendiği araştırmaların olmadığı görülmektedir. Bir alanda sunulan bildiriler; ilgili alanın eğilimine, güncel çalışmalarına ve mevcut problemlere çözüm önerileri üretmeye yönelik olduğundan dolayı, sunulan bildirilen bir arada incelenmesi önem arz etmektedir. Bu bağlamda, Eğitimde ve Psikolojide Ölçme ve Değerlendirme Derneği tarafından düzenli olarak iki yılda bir gerçekleştirilen Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongre'lerinde sunulan bildiri çalışmalarının doküman analizi yöntemi ile incelenmesi önemli görülmüştür. Bu nedenle ölçme ve değerlendirme alanındaki bu boşluğu kapatmak amacıyla bu alanda yapılan bildiriler nitel araştırma desenlerinden durum çalışması ile desenlenmiş ve mevcut veri seti doküman analizi ile incelenmiştir.

Araştırmanın Amacı

Bu araştırmanın genel amacı Türkiye'de Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongrelerinde sunulan bildirilerin eğilimlerinin nasıl olduğunun belirlenmesidir. Bu kapsamda 2008, 2010, 2012 ve 2014 yıllarında Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongrelerinde sunulmuş olan bildirilerin konu temaları, anahtar kelimeleri, yazarların kurumları ve unvanları, araştırma türü, araştırma modeli/deseni, örneklem büyüklüğü, kullanılan veri türü ve veri toplama aracı incelenmiştir. Bu araştırma kapsamında aşağıdaki sorulara yanıt aranmıştır:

1. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongrelerinde sunulan bildirilerin kongrelere göre konu temalarının eğilimi nasıldır?
2. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongrelerinde sunulan bildirilerin kongrelere göre yöntem bölümü nasıldır?"
    a) araştırma türünün dağılımı nasıldır?
    b) araştırma modeli/deseninin dağılımı nasıldır?
    c) örneklem büyüklüğü nedir?
    d) kullanılan veri türü nedir?
    e) kullanılan veri toplama araçları nelerdir?

3. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongrelerinde sunulan çalışmalarda hangi anahtar kelimeler kullanılmıştır?
4. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongrelerinde bildiri sunan kişilerin unvanları nedir?
5. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongrelerinde bildiri sunan kişilerin kurumlarına göre dağılımı nasıldır?

YÖNTEM

Araştırmanın Modeli

Bu araştırma, Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresinde yapılan çalışmaların eğilimlerinin nasıl olduğunun betimlenmesi ve var olan durumun yorumlanması amaçlandığı için, nitel bir araştırma olup, nitel araştırma türlerinden durum çalışması ile desenlenmiştir. Veri toplama yöntemi olarak doküman analizi kullanılmıştır.

Evren

Araştırmanın evrenini 2008, 2010, 2012 ve 2014 yıllarında gerçekleştirilen Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongrelerinde sunulan bildiriler oluşturmaktadır. İlk olarak 2008 yılında yapılan I. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi Ankara Üniversitesi'nde, 2010 yılında yapılan II. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi Mersin Üniversitesi'nde, 2012 yılında yapılan III. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi Bolu Abant İzzet Baysal Üniversitesi'nde ve 2014 yılında yapılan IV. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi ise Hacettepe Üniversitesi ev sahipliğinde gerçekleştirilmiştir. Araştırma doğrudan evren üzerinden yapılmış olup her hangi bir örnekleme yöntemi kullanılmamıştır. Araştırma kapsamında incelenen bildirilere ilişkin betimsel istatistikler Tablo 1'de verilmiştir.

Tablo 1. Kongrelere göre Bildirilerin Betimsel İstatistikleri

Yıl	Kongreler	Frekans	Yüzde (%)
2008	I. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi	60	24.2
2010	II. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi	42	16.9
2012	III. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi	52	21.0
2014	IV. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi	94	37.9
	Toplam	248	100

Tablo 1 incelendiğinde çalışmada toplam 248 bildiri incelenmiştir, bu bildirilerin (n=60) %24.2'si 2008 yılında gerçekleştirilen I. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresinde, (n=42) %16.9'u 2010 yılında gerçekleştirilen II. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresinde, (n=52) %21.0'ı 2012 yılında gerçekleştirilen III. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresinde ve (n=94) %37.9'u ise 2014 yılında gerçekleştirilen IV. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresinde sunulmuştur.

Verilerin Toplanması ve Analizi

Araştırmada Eğitimde ve psikolojide ölçme ve değerlendirme kongrelerinde sunulan bildirilere ulaşabilmek için, ilgili kongreleri düzenleyen akademisyenler ile iletişime geçilmiş bunun sonucunda 2008, 2010 ve 2012 yıllarındaki kongrelerde sunulan bildirilerin tam metinlerine 2014 yılındaki kongrede ise özet metinlerine ulaşılmıştır. Ayrıca ilgili kongreleri düzenleyen Eğitimde ve

Psikolojide Ölçme ve Değerlendirme Derneği (EPODDER) yönetimi ile iletişime geçilmiş ve gerekli izinler alınmıştır. Bu kapsamda toplam olarak 248 bildiriye ulaşılmış ve araştırmacılar tarafından incelenmeye alınmıştır.

Veri toplama aracı olarak, araştırmacılar tarafından geliştirilen Ek 1'deki "Bildiri İnceleme Formu" kullanılmıştır. Bu form üç kısımdan oluşmaktadır. Birinci kısımda bildirinin künyesi, ikinci kısımda bildirinin teması, üçüncü kısımda bildirinin yöntemi kısımlarından oluşmaktadır. Veri toplama aracının hazırlanmasında ilk olarak ilgili alan yazında daha önce başka araştırmacılar tarafından hazırlanılan formlar incelenmiştir. Daha sonra araştırmacılarla iletişime geçilip dönüt alınmış ve en son olarak ölçme ve değerlendirme uzmanı olan 7 kişinin görüşü sonucunda form son halini almıştır.

Araştırma kapsamında doküman analiziyle incelenen bildirilerden elde edilen verinin analizinde betimsel istatistikler ve kelime frekans sorgusu kullanılmıştır. Kelime frekans sorgusu analizi bir metinde geçen kelimelerin frekanslarını hesaplamakta ve frekans ağırlıklarına göre görsel sonuçlar oluşturmaktadır. Kelime/değişken sayısının çok fazla olduğu durumlarda kelime/değişken frekanslarının tablo halinde verilmesi zor olmaktadır, böyle durumlarda kelime frekans sorgusu analizinde frekansa göre ağırlıklandırılmış görsel şekillerin verilmesi bulguların yorumlanmasını kolaylaştırmaktadır.

Bildirilerin konu teması incelenirken EPOD kongrelerinin sitesinde yer alan ve kongreye bildiri gönderilebilecek konuları ifade eden konu temalarından yararlanılmıştır. Bu konu temalarına ek olarak araştırmacılar tarafından eksikliği hissedilen bazı konu temaları eklenmiştir. Ayrıca bu temalar dışında kalan konular "diğer" konu teması altında ifade edilmiştir.

İnandırıcılık (Güvenirlik) ve Aktarılabilirlik (Geçerlik) Çalışması

Araştırma kapsamında kullanılan EPOD bildiri sınıflama formu için kodlayıcılar arası puanlama güvenirliğini belirlemek amacıyla rasgele seçilen 6 bildiri her araştırmacı tarafından veri toplama aracı doğrultusunda oluşturulan forma göre ayrı ayrı bağımsız olarak kodlanmış ve kodlayıcılar arasındaki farklılıkların olup olmadığını belirlemek için uzlaşma katsayısı hesaplanmıştır.

Güvenirlik = Uzlaşma sayısı / (Uzlaşma sayısı + Uzlaşmama sayısı)

Güvenirlik = 0.933 olarak bulunmuştur.

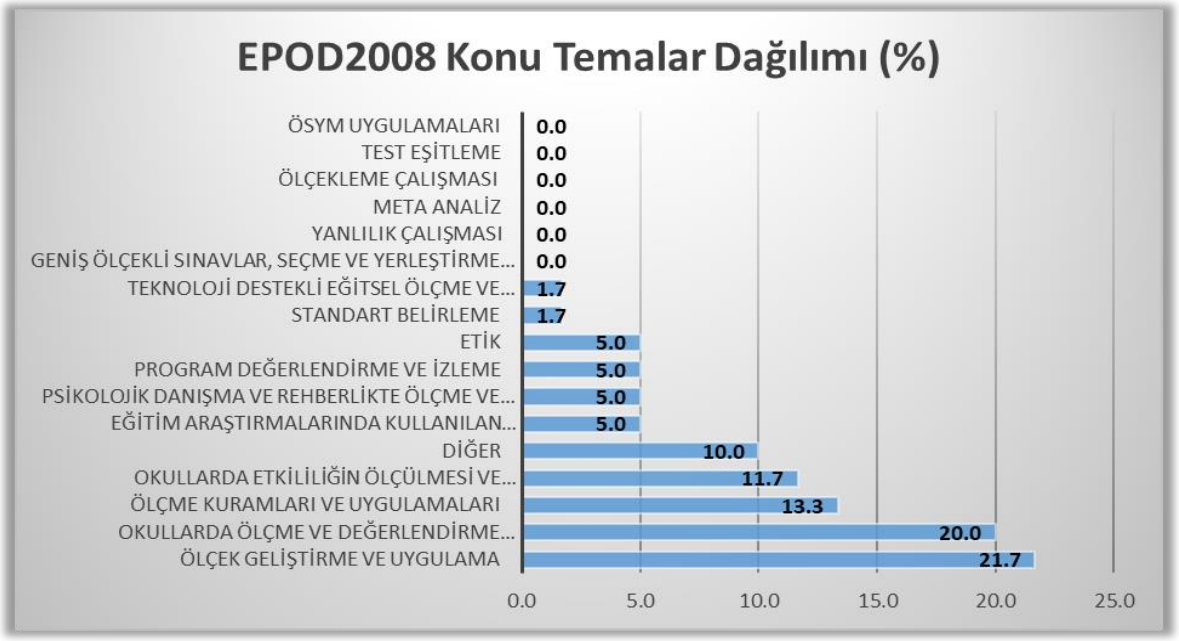
Elde edilen sonuç ile kodlayıcılar arası tutarlılık anlamındaki güvenirliğinin sağlandığını görülmüştür.

BULGULAR

Bu bölümde, araştırma sürecinde toplanan verilerden elde edilen bulgu ve yorumlara yer verilmiştir. Her bir alt problem sırasıyla ele alınmıştır.

Birinci Araştırma Sorusuna Ait Bulgular

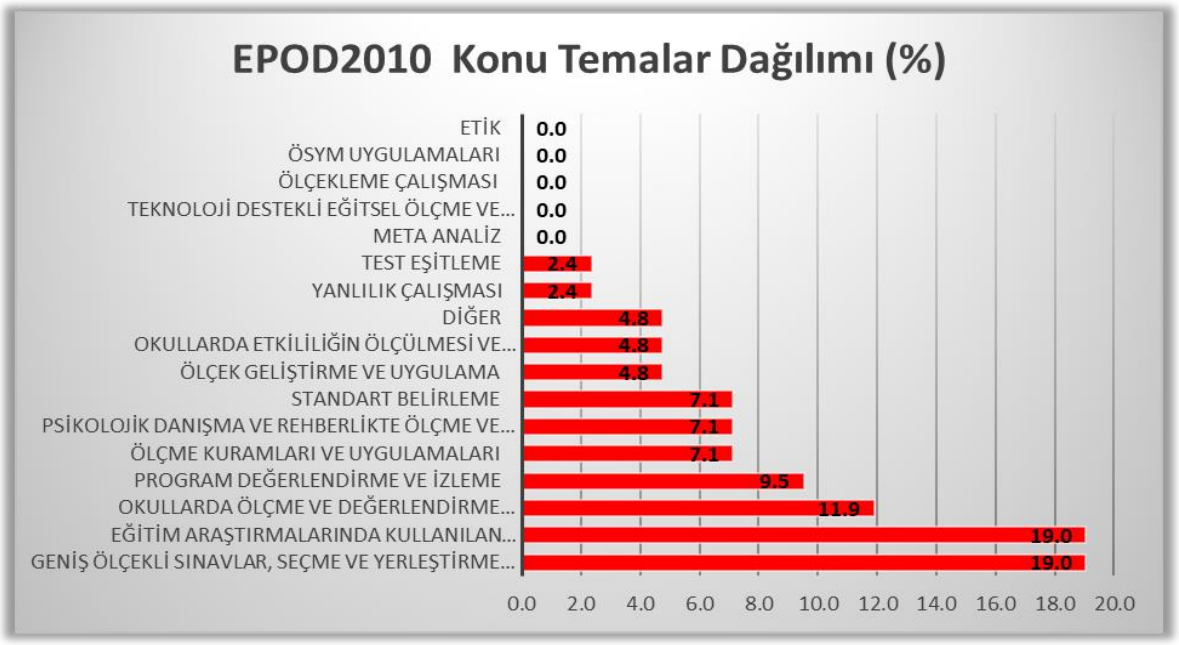
Araştırmanın birinci alt amacı olarak "Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongrelerinde sunulan bildirilerin kongrelere göre konu temalarının eğilimi nasıldır?" sorusuna yanıt aranmıştır. Şekil 1'de 2008 EPOD kongresinde sunulan bildirilerin konu temalarına göre yüzde olarak dağılımı yer almaktadır.



Şekil 1. EPOD 2008 Kongresine göre Konu Temalarının Yüzde Dağılımı

Şekil 1 incelendiğinde EPOD 2008 kongresinde en fazla %21.7 (n=13) "Ölçek geliştirme ve uyarlama" konu temasında bildiri sunulmuştur. Ölçek geliştirme ve uyarlama temasında sunulan bildirilerin çoğunluğu ise ölçek geliştirme (n=10) çalışmasıdır. EPOD 2008 kongresinde "ÖSYM Uygulamaları", "Test eşitleme", "Ölçekleme çalışması", "Meta Analiz", "Yanlılık Çalışması" ve "Geniş ölçekli sınavlar, seçme ve yerleştirme sınavları, yüksek riskli sınavlar" temalarında hiçbir bildiri sunulmamıştır. Ayrıca en az %1.7 (n=1) "Teknoloji Destekli Eğitsel Ölçme ve Değerlendirme" ve "Standart Belirleme" temalarında bildiri sunulmuştur.

Şekil 2'de 2010 EPOD kongresinde sunulan bildirilerin konu temalarına göre yüzde olarak dağılımı yer almaktadır.



Şekil 2. EPOD 2010 Kongresine göre Konu Temalarının Yüzde Dağılımı

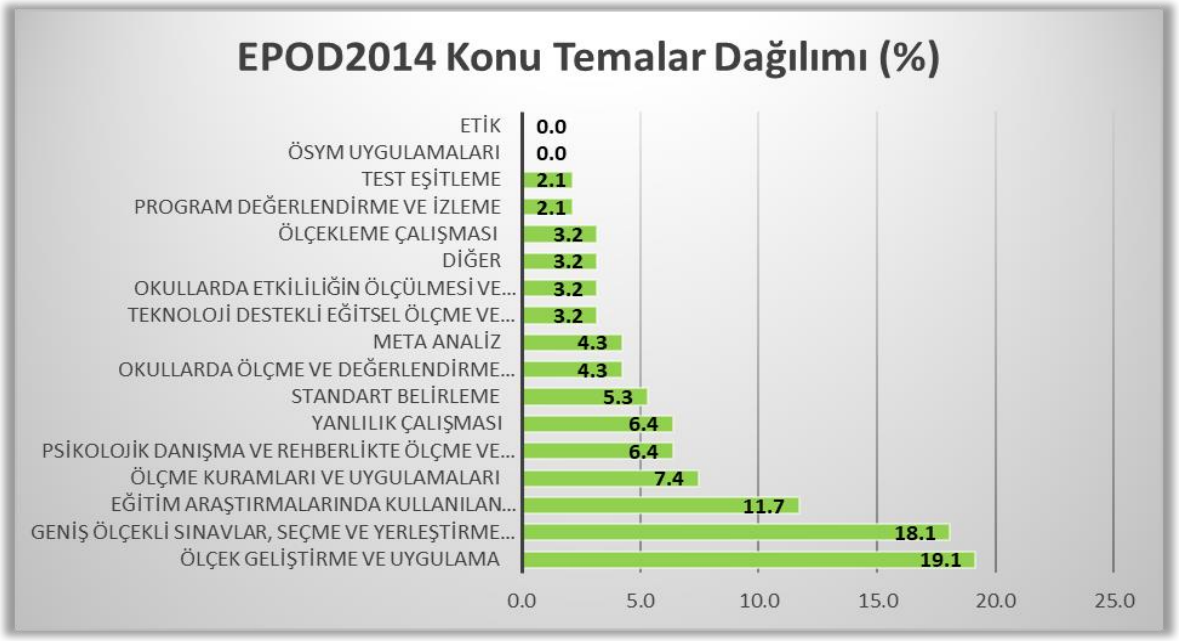
Şekil 2 incelendiğinde EPOD 2010 kongresinde en fazla "Eğitim araştırmalarında kullanılan istatistiksel yönelimler" ve "Geniş ölçekli sınavlar, seçme ve yerleştirme sınavları, yüksek riskli sınavlar" konu temasında bildiri (%19 (n=8)'u) sunulmuştur. Eğitim araştırmalarında kullanılan istatistiksel yönelimler temasında yapılan çalışmaların yapısal eşitlik modelle ve regresyon analizi konularında, geniş ölçekli sınavlar, seçme ve yerleştirme sınavları, yüksek riskli sınavlar temasında yapılan çalışmaların ise Uluslararası Matematik ve Fen Eğilimleri Araştırması ve Seviye Belirleme sınavlarıyla ilgili konulara da yoğunlaştığı belirlenmiştir. EPOD 2010 kongresinde "Etik", "ÖSYM Uygulamaları", "Ölçekleme çalışması" ve "Meta Analiz" konu temalarında hiçbir bildiri sunulmamıştır. Ayrıca en az %2.4 (n=1) "Test Eşitleme" ve "Yanlılık Çalışması" temalarında bildiri sunulmuştur.

Şekil 3'te 2012 EPOD kongresinde sunulan bildirilerin konu temalarına göre yüzde olarak dağılımı yer almaktadır.

Şekil 3. EPOD 2012 Kongresine göre Konu Temalarının Yüzde Dağılımı

Şekil 3. incelendiğinde EPOD 2012 kongresinde en fazla %19.2 (n=10)'u "Eğitim araştırmalarında kullanılan istatistiksel yönelimler" konu temasında bildiri sunulmuştur. Eğitim araştırmalarında kullanılan istatistiksel yönelimler temasında yapılan çalışmaların yapısal eşitlik modelleri ile ilgili konularda yoğunlaştığı belirlenmiştir. EPOD 2012 kongresinde "Etik", "Teknoloji Destekli Eğitsel Ölçme ve Değerlendirme" ve "Meta Analiz" temalarında hiçbir bildiri sunulmamıştır. Ayrıca en az %1.9 (n=1) "Test Eşitleme" ve "ÖSYM uygulamaları" temalarında bildiri sunulmuştur.

Şekil 4'te 2014 EPOD kongresinde sunulan bildirilerin konu temalarına göre yüzde olarak dağılımı yer almaktadır.



Şekil 4. EPOD 2014 Kongresine göre Konu Temalarının Yüzde Dağılımı

Şekil 4 incelendiğinde EPOD 2014 kongresinde en fazla %19.1 (n=18)'u "Ölçek Geliştirme ve Uyarlama" konu temasında bildiri sunulmuştur. Ölçek geliştirme ve uyarlama temasında sunulan bildirilerin çoğunluğu ise ölçek geliştirme (n=11) çalışmasıdır. EPOD 2014 kongresinde "Etik" ve "ÖSYM Uygulamaları" temalarında hiçbir bildiri sunulmamıştır. Ayrıca en az %2.1 (n=2) "Test Eşitleme" ve "Program Değerlendirme ve İzleme" temalarında bildiri sunulmuştur.

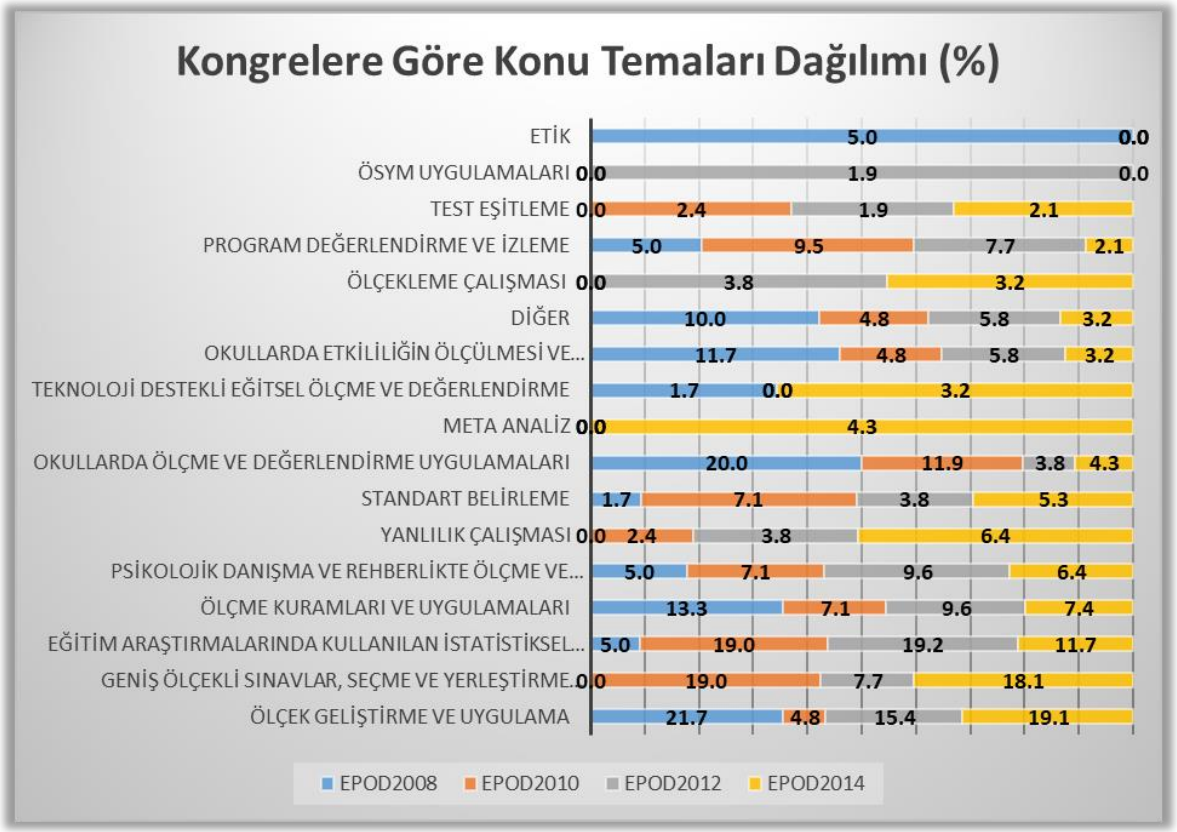
İlk olarak 2008 yılında yapılan ve daha sonra her iki yılda bir düzenlenen EPOD kongreleri bağımsız olarak ayrı ayrı konu temalarına göre incelendikten sonra 2008, 2010, 2012 ve 2014 yıllarında düzenlenen toplam dört kongre konu temasına göre bir arada incelenmiştir. Tablo 2'de tüm (2008, 2010, 2012 ve 2014) EPOD kongresinde sunulan bildirilerin konu temalarına göre dağılımı yer almaktadır.

Tablo 2. EPOD Kongrelerine göre Konu Temalarının Dağılımı

TEMALAR	**Kongreler**				**Toplam**	
	2008	**2010**	**2012**	**2014**	**N**	**%**
Ölçek Geliştirme ve Uyarlama	13	2	8	18	41	16.5
Eğitim araştırmalarında kullanılan istatistiksel yönelimler	3	8	10	11	32	12.9
Geniş ölçekli sınavlar, seçme ve yerleştirme sınavları, yüksek riskli sınavlar	0	8	4	17	29	11.7
Ölçme kuramları ve uygulamaları	8	3	5	7	23	9.3
Okullarda ölçme ve değerlendirme uygulamaları	12	5	2	4	23	9.3
Psikolojik danışma ve rehberlikte ölçme ve değerlendirme	3	3	5	6	17	6.9
Okullarda etkililiğin ölçülmesi ve değerlendirilmesi	7	2	3	3	15	6.0
Diğer	6	2	3	3	14	5.6
Program değerlendirme ve izleme	3	4	4	2	13	5.2
Standart Belirleme	1	3	2	5	11	4.4
Yanlılık Çalışması	0	1	2	6	9	3.6
Ölçekleme Çalışması	0	0	2	3	5	2.0
Meta Analiz	0	0	0	4	4	1.6
Teknoloji Destekli Eğitsel Ölçme ve Değerlendirme	1	0	0	3	4	1.6
Test Eşitleme	0	1	1	2	4	1.6
Etik	3	0	0	0	3	1.2
ÖSYM Uygulamaları	0	0	1	0	1	0.4
Toplam	60	42	52	94	248	100

Tablo 2 incelendiğinde EPOD kongrelerinde sunulan bildirilerin genel olarak "Ölçek Geliştirme ve Uyarlama" , "Eğitim araştırmalarında kullanılan istatistiksel yönelimler" ve "Geniş ölçekli sınavlar, seçme ve yerleştirme sınavları, yüksek riskli sınavlar" konu temalarında yoğunlaştığı görülmektedir. Ayrıca "ÖSYM Uygulamaları" ve "Etik" konu temalarında çok fazla çalışma yapılmadığı görülmektedir. Ayrıca belirlenen 16 konu temasından her hangi birisine girmeyen çalışmalar Tablo 2'de "diğer" konu teması olarak belirtilmiş olup, yapılan çalışmaların %5.6 (n=14)'sı "diğer" konu temasında yer almaktadır.

Şekil 5'te tüm EPOD kongresinde sunulan bildirilerin konu temalarına göre yüzde olarak dağılımı yer almaktadır.



Şekil 5. EPOD Kongresine göre Konu Temalarının Yüzde Dağılımı

Şekil 5 incelendiğinde, "Etik" konu temasıyla ilgili bildiri çalışmasının sadece 2008 EPOD kongresinde, "ÖSYM Uygulamaları" konu temasıyla ilgili bildiri çalışmasının sadece 2010 EPOD kongresinde, "Meta Analiz" konu temasıyla ilgili bildiri çalışmasının sadece 2014 EPOD kongresinde, "Teknoloji Destekli Eğitsel Ölçme ve Değerlendirme" konu temasıyla ilgili bildiri çalışmasının sadece 2008 EPOD ve 2014 EPOD kongrelerinde sunulduğu görülmektedir. "Test eşitleme" konu temasında 2010 EPOD kongresinden itibaren, "Ölçekleme Çalışması" konu temasında 2012 EPOD kongresinden itibaren bildiriler sunulmaya başlanmıştır. "Program Değerlendirme ve İzleme" konu teması ile ilgili her kongrede bildiri sunulmuş olup, 2010 EPOD kongresinden itibaren sunulan bildiri yüzdelerinde bir düşüş olduğu belirlenmiştir. "Okullarda etkililiğin ölçülmesi ve değerlendirilmesi" konu temasıyla ilgili yapılan bildiri çalışmalarında 2010 yılı hariç, yıllara göre yüzde olarak azalma olmuştur. Benzer şekilde "Okullarda Ölçme ve Değerlendirme Uygulamaları" konu temasıyla ilgili yapılan bildiri çalışmalarında ise 2014 yılı hariç yıllara göre yüzde olarak azalma meydana gelmiştir. EPOD kongrelerinde "Yanlılık Çalışması" konu temasıyla ilgili bildiri çalışmalarının yıllara göre sürekli bir artış olduğu tespit edilmiştir.

İkinci Araştırma Sorusuna Ait Bulgular

Araştırmanın ikinci alt amacı olarak "Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongrelerinde sunulan bildirilerin kongrelere göre yöntem bölümü nasıldır?" sorusuna yanıt aranmıştır. Bu kapsamda ilk olarak Tablo 3'te Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongrelerinde sunulan bildirilerin kongrelere göre araştırma türünün dağılımının nasıl olduğu yer almaktadır.

Tablo 3. EPOD Kongrelerine göre Bildirinin Türü

		Kongre				**Toplam**	**Yüzde (%)**
		EPOD2008	**EPOD2010**	**EPOD2012**	**EPOD2014**		
Bildirinin Türü	Nicel	27	30	36	37	130	52.4
	Belirtilmemiş	22	2	7	51	82	33.1
	Nitel	11	9	9	4	33	13.3
	Karma	0	1	0	2	3	1.2
Toplam		60	42	52	94	248	100

Tablo 3 incelendiğinde kongrelerde sunulan bildirilerden (n=130) %52.4 nicel araştırma türünde, (n=82) %33.1'i araştırma türü belirtilmeyen, (n=33) %13.3'ü nitel araştırma türünde ve (n=3) %1.2'sinin ise karma araştırma türünde olduğu görülmektedir. Kongrelerde sunulan bildirilerde en çok tercih edilen araştırma türünün nicel, en az tercih edilen ise karma araştırma türünde olduğu belirlenmiştir. Ayrıca EPOD 2014'te bildiri türü belirtmeyenlerin sayısının diğer kongrelere göre daha yüksek olduğu görülmektedir. Bunun sebebinin daha önceden de belirtildiği gibi 2014 EPOD kongrelerinin sadece kısa özetlerine ulaşılabilmiş ve araştırma verileri kısa özetten elde edilmesinden kaynaklanabileceği düşünülmektedir.

Tablo 4'te Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongrelerinde sunulan bildirilerin kongrelere göre araştırma modeli/deseni dağılımı nasıl olduğu yer almaktadır.

Tablo 4. EPOD Kongrelerine göre Araştırma Modeli/Deseni

		Kongre				Toplam	Yüzde (%)
		EPOD2008	EPOD2010	EPOD2012	EPOD2014		
Araştırma Modeli/Desen	Belirtilmemiş	34	14	23	67	138	55.6
	Tarama	12	8	11	11	42	17.0
	Betimsel araştırma	8	7	8	2	25	10.1
	Korelasyonel	0	2	4	6	12	4.8
	Doküman analizi	1	3	1	2	7	2.8
	Temel Araştırma	0	3	2	1	6	2.4
	Meta-Analiz	0	0	0	4	4	1.6
	Deneysel	1	2	1	0	4	1.6
	Durum çalışması	1	1	0	1	3	1.2
	Olgu bilim	2	0	1	0	3	1.2
	Kuramsal araştırma	0	1	0	0	1	0.4
	Eylem araştırması	0	1	0	0	1	0.4
	Örnek Olay	1	0	0	0	1	0.4
	Kuram oluşturma	0	0	1	0	1	0.4
Toplam		60	42	52	94	248	100

Tablo 4 incelendiğinde kongrelerde sunulan bildirilerin (n=138) %55,6'sının araştırma modeli/deseni belirtilmemiştir. Kongrelerde en çok tercih edilen araştırma modelleri/desenleri ise (n=42) %17.0'i "Tarama", (n=25) %10.1'i "Betimsel araştırma" olduğu belirlenirken, en az tercih edilen araştırma türü ise (n=1) %0,4 ile "Kuram oluşturma", "Örnek olay", "Eylem Araştırması" ve "Kuramsal araştırma" olduğu görülmektedir.

Tablo 5' te Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongrelerinde sunulan bildirilerin kongrelere göre örneklem büyüklükleri yer almaktadır.

Tablo 5. EPOD Kongrelerine göre Örneklem Büyüklüğü

		Kongre				Toplam	Yüzde (%)
		EPOD2008	EPOD2010	EPOD2012	EPOD2014		
Örneklem Büyüklüğü	301-1000 arası	11	11	12	25	59	23.8
	101-300 arası	14	9	15	13	51	20.6
	Belirtilmemiş	9	2	5	31	47	19.0
	1000'den fazla	3	8	9	12	32	12.9
	1-30 arası	14	4	3	8	29	11.7
	31-100 arası	9	7	6	5	27	10.9
	Yok	0	1	2	0	3	1.2
Toplam		60	42	52	94	248	100

Tablo 5 incelendiğinde kongrelerde sunulan bildirilerde en çok tercih edilen örneklem büyüklükleri (n=59) %23.8 ile "301-1000 arası" ve (n=51) %20.6 ile "101-300 arası" olduğu görülmektedir. Ayrıca sunulan bildirilerden (n=47) %19.0'ının ise örneklem büyüklükleri belirtilmemiştir. Özellikle EPOD 2014'te örneklem büyüklü belirtmeyenlerin sayısı diğer kongrelere göre daha yüksek olduğu görülmektedir. Bunun sebebinin daha önceden de belirtildiği gibi 2014 EPOD kongrelerinin sadece kısa özetlerine ulaşılabilmiş ve araştırma verileri kısa özetten elde edilmesinden kaynaklanabileceği düşünülmektedir. Kongrelerde sunulan bildirilerde en az tercih edilen örneklem büyüklüğü ise (n=27) %10.9 ile "31-100 arası" olduğu görülmektedir. Sunulan bildirilerin (n=3) %1.2'sinin ise örneklem büyüklüğü olmayan ve tabloda "yok" ile belirtilen kuramsal çalışmalardır.

Tablo 6'da Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongrelerinde sunulan bildirilerin kongrelere göre kullanılan veri türü yer almaktadır.

Tablo 6. EPOD Kongrelerine göre Kullanılan Veri Türü

		Kongre				Toplam	Yüzde (%)
		EPOD2008	EPOD2010	EPOD2012	EPOD2014		
Kullanılan Veri Türü	Uygulama Verisi	40	23	36	43	142	57.3
	Hazır Veri	7	10	13	19	49	19.8
	Belirtilmemiş	10	1	1	23	35	14.1
	Simülasyon Verisi	1	2	1	9	13	5.2
	Yok	1	6	1	0	8	3.2
	Hazır ve Uygulama verisi	1	0	0	0	1	0.4
Toplam		60	42	52	94	248	100

Tablo 6 incelendiğinde kongrelerde sunulan bildirilerde en çok kullanılan veri türü (n=142) %57.3 ile "uygulama verisi" ve (n=49) %19.8 ile "Hazır veri" olduğu görülmektedir. Ayrıca sunulan bildirilerden (n=35) %14.1'ının ise kullanılan veri türü belirtilmemiştir. Kongrelerde sunulan bildirilerde en az kullanılan veri türü ise (n=1) %0.4 ile "Hazır ve Uygulama verisi" nin bir arada kullanıldığı çalışma olduğu görülmektedir. Sunulan bildirilerin (n=8) %3.2'sinin ise veri türü yoktur.

Tablo 7'de Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongrelerinde sunulan bildirilerin kongrelere göre veri toplama araçları yer almaktadır.

Tablo 7. EPOD Kongrelerine göre Veri Toplama Aracı

		Kongre				Toplam	Yüzde (%)
		EPOD2008	EPOD2010	EPOD2012	EPOD2014		
Veri Toplama Aracı	Geliştirilmiş	33	13	20	25	91	36.7
	Hazır	6	19	19	25	69	27.8
	Belirtilmemiş	10	1	1	23	35	14.1
	Yok	4	8	5	17	34	13.7
	Uyarlanmış	5	0	3	4	12	4.8
	Hazır ve Geliştirilmiş	2	1	3	0	6	2.4
	Uyarlanmış ve Hazır	0	0	1	0	1	0.4
Toplam		60	42	52	94	248	100

Tablo 7 incelendiğinde kongrelerde sunulan bildirilerde en çok kullanılan veri toplama aracı (n=91) %36.7 ile "Geliştirilmiş" ve (n=69) %27.8 ile "Hazır" veri toplama araçları olduğu görülmektedir. Ayrıca sunulan bildirilerden (n=35) %14.1'ının ise kullanılan veri toplama aracı belirtilmemiştir. Bazı çalışmalarda ise birden çok veri toplama aracının bir arada kullanıldığı belirlenmiştir. Kongrelerde sunulan bildirilerde en az kullanılan veri toplama aracı ise (n=1) %0.4 ile "Hazır ve Uyarlanmış" veri toplama aracının bir arada kullanıldığı çalışma olduğu görülmektedir. Sunulan bildirilerin (n=34) %13.7'sinin ise veri toplama aracı yoktur.

Dördüncü Araştırma Sorusuna Ait Bulgular

Araştırmanın dördüncü alt amacı olarak "Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongrelerinde sunulan çalışmalarda hangi anahtar kelimeler kullanılmıştır?" sorusuna yanıt aranmıştır. Tablo 8'de EPOD kongrelerinde kullanılan anahtar kelimelerin frekans sorgusu yer almaktadır. Sunulan bildirilerdeki anahtar kelime sayısı çok fazla olduğundan, anahtar kelime frekansları tablo halinde verilmeyip kelime frekans sorgusu kullanılarak frekansa göre ağırlandırılmış görsellerle sunulmuştur

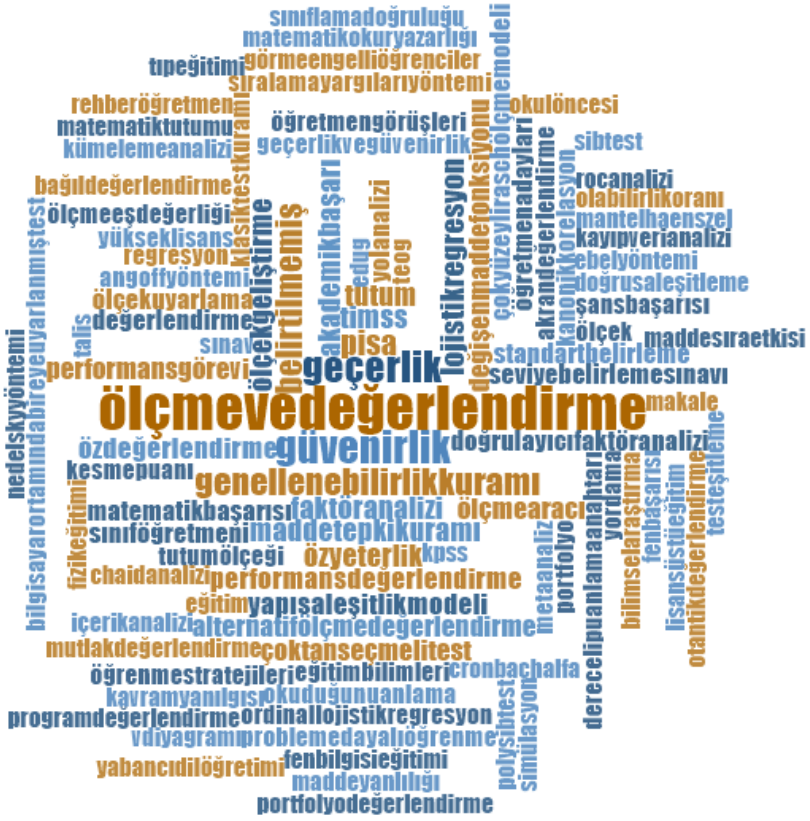
Tablo 8. EPOD Kongreleri "Anahtar Kelime" Kelime Frekans Sorgusu

2008 EPOD Anahtar Kelime	**2010 EPOD Anahtar Kelime**
	
2012 EPOD Anahtar Kelime	**2014 EPOD Anahtar Kelime**
	

Tablo 8 incelendiğinde EPOD 2008'de sunulan bildirilerde en çok kullanımı tercih edilen anahtar kelime "ölçme ve değerlendirme" olduğu bulunmuştur. Anahtar kelime belirtmeyenlerin yani "belirtilmemiş" 'in ise ne yüksek ikinci frekansa sahip olduğu bulunmuştur. Daha sonra ise "geçerlik" ve "güvenirlik" anahtar kelimelerinin ön plana çıktığı belirlenmiştir. Kullanımı en çok tercih edilen anahtar kelimeler EPOD 2010'da sırasıyla "ölçme ve değerlendirme", "faktör analizi", ve "seviye belirleme sınavı" olduğu; EPOD 2012' de sırasıyla "güvenirlik", "geçerlik" ve " ölçme ve değerlendirme"; EPOD 2014'te ise sırasıyla "ölçme ve değerlendirme", "güvenirlik" ve "lojistik regresyon" olduğu belirlenmiştir. Genel olarak bakıldığında ise EPOD 2012 dışındaki diğer

kongrelerde kullanımı en çok tercih edilen anahtar kelimenin "ölçme ve değerlendirme" olduğu, sadece EPOD 2012'de "güvenirlik" olduğu görülmektedir.

Şekil 6'da ise tüm kongrelerin bir arada ele alınmasıyla elde edilen kelime frekans sorgusu yer almaktadır. Ayrıca Şekil 6'da kelime frekans sorgusu sonucunda elde edilen şekillerin büyüklükleri, kelimenin frekans ağırlığı ile doğru orantılıdır.



Şekil 6. Genel EPOD "Anahtar Kelime" Kelime Frekans Sorgusu

Şekil 6 incelendiğinde yapılan tüm kongrelerde kullanımı en çok tercih edilen anahtar kelimelerin sırasıyla %3.74 (n=31) "ölçme ve değerlendirme", %2.17 (n=18) "güvenirlik" ve %2.05 (n=17) "geçerlik" anahtar kelimeleri olduğu belirlenmiştir.

Beşinci Araştırma Sorusuna Ait Bulgular

Araştırmanın beşinci alt amacı olarak "Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongrelerinde bildiri sunan kişilerin unvanlara göre dağılımı nasıldır?" sorusuna yanıt aranmıştır. Yapılan tüm EPOD kongrelerinde bildiri sunan kişilerin unvanlara göre dağılımı Tablo 9'da yer almaktadır.

Tablo 9. EPOD Kongreleri "Unvan" Kelime Frekans Sorgusu

2008 EPOD Unvan	2010 EPOD Unvan
	
2012 EPOD Unvan	**2014 EPOD Unvan**
	

Tablo 9 incelendiğinde bildiri sunan kişilerin EPOD 2008'de en fazla "Yrd.Doç.Dr." ve "Arş. Gör." unvanlarına sahip olduğu, EPOD 2010'da "Arş. Gör." ve "Yrd. Doç. Dr." unvanlarına sahip olduğu, EPOD 2012'de "Arş. Gör." ve "Doç. Dr." unvanlarına sahip olduğu ve EPOD 2014'de ise unvanlarını belirtmeyenleri ifade eden "belirtmemiş" lerin ağırlıkta olduğu daha sonra ise sırasıyla "Arş. Gör." ve "Prof. Dr." unvanlarına sahip olduğu görülmektedir. Tablodaki elde edilen şekillerin büyüklükleri frekans ağırlıklarıyla orantılı olduğundan dolayı, özellikle EPOD 2014'te "belirtilmemiş" frekansının diğer unvanlara göre çok yüksek olduğu görülmektedir. Bunun sebebinin daha önceden de belirtildiği gibi 2014 EPOD kongrelerinin sadece kısa özetlerine ulaşılabilmiş ve araştırma verileri kısa özetten elde edilmesinden kaynaklanabileceği düşünülmektedir.

Tablo 10'da tüm kongrelerin bir arada ele alınmasıyla elde edilen bildiri sunan kişilerin unvanlarına ait frekans tablosu yer almaktadır.

Tablo 10. Tüm EPOD Kongrelerinde Bildiri Sunan Kişilerin Unvanlara göre Frekans Dağılımı

Unvan	Frekans	Yüzde (%)
Belirtilmemiş	186	36.40
Arş. Gör.	94	18.40
Yrd. Doç. Dr.	65	12.72
Doç. Dr.	42	8.22
Prof. Dr.	33	6.46
Uzman	21	4.11
Dr.	15	2.94
Doktora öğrencisi	10	1.96
Öğretmen	10	1.96
Yüksek lisans öğrencisi	9	1.76
Öğr. Gör. Dr.	9	1.76
Öğr. Gör.	7	1.37
Okutman	4	0.78
Arş. Gör. Dr.	2	0.39
Lisans öğrencisi	1	0.20
Müdür	1	0.20
Psikolog	1	0.20
Ölçme uzmanı	1	0.20
Toplam	511	100

Tablo 10 incelendiğinde yapılan tüm kongreler de unvan belirtmeyenlerin yani "belirtilmemiş" kategorisinin en yüksek frekansa (n=186) sahip olduğu belirlenmiştir. Bunun sebebinin daha önceden de belirtildiği gibi 2014 EPOD kongrelerinin sadece kısa özetlerine ulaşılabilmiş ve araştırma verileri kısa özetten elde edilmesinden kaynaklanabileceği düşünülmektedir. Daha sonra kongrelerde bildiri sunan kişilerin sırasıyla "Arş. Gör.", "Yrd. Doç. Dr." ve "Doç. Dr." unvanlarına sahip olduğu bulunmuştur.

Altıncı Araştırma Sorusuna Ait Bulgular

Araştırmanın altıncı alt amacı olarak "Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongrelerinde bildiri sunan kişilerin kurumlarına göre dağılımı nasıldır?" sorusuna yanıt aranmıştır. Yapılan tüm EPOD kongrelerinde bildiri sunan kişilerin kurumlarına göre dağılımı Tablo 11'de yer almaktadır.

Tablo 11. EPOD Kongreleri "Kurum" Kelime Frekans Sorgusu

2008 EPOD Kurum	2010 EPOD Kurum
	
2012 EPOD Kurum	**2014 EPOD Kurum**
	

Tablo 11 incelendiğinde bildiri sunan kişilerin EPOD 2008'de sırasıyla en fazla "Ankara", "Akdeniz" üniversitesi ve "Cito" mensubu olduğu, EPOD 2010'da sırasıyla en fazla "Ankara", "Hacettepe" ve "Mersin" üniversitesi mensubu olduğu, EPOD 2012'de sırasıyla en fazla "Ankara" , "Hacettepe" ve "Sakarya" üniversitesi mensubu olduğu ve EPOD 2014'de ise sırasıyla en fazla "Hacettepe" , "Gazi" ve "Ankara " üniversitesi mensubu olduğu görülmektedir. Ayrıca yapılan tüm EPOD kongrelerinde bildiri sunan kişilerin kurumlarına göre dağılımı ise Şekil 7'de yer almaktadır.

Şekil 7. Tüm EPOD Kongrelerinde Bildiri Sunan Kişilerin Kurum Dağılımı

Şekil 7 incelendiğinde kongrelerde en çok bildiri sunan kişilerin kurumlarının sırasıyla %18.47 (n=94) Ankara, %14.93 (n=76) Hacettepe ve %7.66 (n=39) Gazi üniversitesi olduğu belirlenmiştir. Ankara'daki bu üniversitelerde eğitimde ölçme ve değerlendirme yüksek lisans ve doktora programının olmasının ve eğitimde ölçme ve değerlendirme alanında istihdam edilen akademik personel sayısın diğer üniversitelerden daha fazla olmasının sonucun bu şekilde çıkmasında etkili olduğu düşünülmektedir.

SONUÇLAR ve TARTIŞMA

Bu araştırmada, 2008-2014 yılları arasında yapılmış dört Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresinde sunulan bildiriler çeşitli değişkenlere göre incelenmiştir. Elde edilen bulgular her bir araştırma sorusu doğrultusunda tartışılmıştır.

Bildirilerin konu temalarının dağılımına ilişkin bulgular incelendiğinde, en fazla çalışılan konu başlığının "Ölçek geliştirme ve uygulama" olduğu en az çalışılan konu başlığının ise "ÖSYM uygulamaları" olduğu görülmüştür. Eğitimde Ölçme ve Değerlendirme alanında yapılan yüksek lisans ve doktora tezlerinin doküman analizi ile incelendiği çalışmalarda da benzer bir eğilimin olduğu görülmektedir (Ayva, Ceyhan ve Doğan, 2015; Şenyurt ve Özkan, 2017). Benzer durum eğitim bilimlerinin diğer alanları/bölümleri için de geçerlidir (Tavşancıl, vd., 2010). Ayrıca eğitim bilimleri dışındaki diğer disiplinlerde yapılan çalışmalar incelendiğinde de veri toplamada en fazla ölçek ve başarı testi kullanıldığı görülmektedir (Küçükoğlu ve Ozan, 2013; Yaşar ve Papatğa 2015). Çalışmanın bulguları ve alanyazın dikkate alındığında bu eğilimin sadece eğitimde ölçme ve değerlendirme alanına özgü olmayıp ulusal bir eğilim olduğu söylenebilir. Bildiri çalışmalarında "ÖSYM uygulamalarının" çok az olmasındaki nedenlerden biri olarak Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı'nın akademisyenler ile veri paylaşmama politikasından kaynaklandığı düşünülmektedir. Bulgular bölümündeki Tablo 2 incelendiğinde PISA ve TIMSS gibi sınavların en

çok çalışılan üçüncü konu başlığı olduğu görülmekte bunun temel nedeninin ise bu sınavların kamuoyuna ve bilim insanlarına erişimin açık olmasından kaynaklanmaktadır. Benzer olarak ÖSYM ve MEB böyle bir politikaya sahip olsalar idi araştırmacıların bu kurumların yaptıkları ölçme ve değerlendirme uygulamalarına yönelecekleri düşünülmektedir. Yıllara göre konu dağılımları incelendiğinde ise çoğu konu başlığının yıllar içinde küçük değişimler gösterdiği (örneğin "test eşitleme", "ölçekleme çalışması", diğerleri için şekil 5'e bakınız), bazı konu başlıklarının ise artığı (örneğin "yanlılık çalışması", "meta analiz" gibi) gözlemlenmiştir.

Bildirilerin yöntem bölümlerinin incelenmesi sonucunda, bildirilerin büyük çoğunluğunun nicel yaklaşımı benimsediği çok azının ise karma (nicel ve nitel yaklaşımın birlikte ele alınması) yaklaşımı benimsediği görülmüştür. Eğitimde Ölçme ve Değerlendirme alanında yapılan lisansüstü tez çalışmalarının ve makale çalışmalarının da benzer bir eğilime sahip oldukları görülmüştür(Erdem, 2011; Acar-Güvendir ve Özer-Özkan, 2015; Ayva, Ceyhan ve Doğan, 2015; Şenyurt ve Özkan, 2017). Bildiri çalışmalarında en fazla nicel yaklaşımlardan tarama modelinin tercih edildiği ve bu durumun alanyazın ile benzerlik gösterdiği bulunmuştur(Tavşancıl, vd., 2010; Erdem, 2011; Ayva, Ceyhan ve Doğan, 2015; Şenyurt ve Özkan, 2017). Ayrıca bildiri çalışmalarının büyük bir çoğunluğunda araştırmanın modelinin belirtilmediği görülmektedir. Bildiri çalışmalarında deneysel desenlerin çok az kullanıldığı benzer eğilimin makale ve lisansüstü tez çalışmalarında da olduğu görülmektedir (Tavşancıl, vd., 2010; Erdem, 2011). Bildiri çalışmalarında kullanılan örneklem büyüklükleri incelendiğinde en fazla tercih edilen örneklem büyüklüğünün 101-300 arasında olduğu en az tercih edilen ise 31-100 arasında olduğu görülmektedir. Bu sonuç bildiri çalışmalarının genellikle çalışma grubu üzerinde yapıldığını ve evrene genelleme yapılmadığı gösterir nitelikte olup ölçme ve değerlendirme alanındaki literatür ile uyum içindedir (Tavşancıl, vd., 2010; Ayva, Ceyhan ve Doğan, 2015; Şenyurt ve Özkan, 2017). Kongrelerde sunulan bildiri çalışmalarında kullanılan veri türlerinin dağılımları incelendiğinde en fazla araştırmacıların kendi uygulamaları ile topladığı veri ve ikinci olarak önceden uygulaması yapılmış olan hazır veri olduğu görülmekte ve alanyazın ile uyum göstermektedir. Bildiri çalışmalarında verinin toplanmasında kullanılan ölçme araçlarına ilişkin eğilim incelendiğinde en fazla araştırmacıların kendi geliştirdikleri ölçme araçlarını tercih ettikleri en az ise başka araştırmacıların hazırladıkları ölçme araçlarının kullanıldığı görülmüştür. Elde edilen sonuçlar alanyazın ile ve diğer disiplinlerdeki çalışmalarla uyum göstermektedir (Tavşancıl, vd., 2010; Küçükoğlu ve Ozan, 2013; Ayva, Ceyhan ve Doğan, 2015; Yaşar ve Papatğa, 2015; Şenyurt ve Özkan, 2017).

Kongrelerde sunulan bildiri çalışmalarında kullanılan anahtar kelime dağılımına ilişkin sonuçlar incelendiğinde, en fazla kullanılan anahtar kelimeler sırasıyla ölçme ve değerlendirme, güvenirlik ve geçerlik olduğu gözlemlenmiştir. Benzer sonuçlar Eğitimde ölçme ve Değerlendirme alanında yapılan ve yüksek lisans ve doktora tezlerinde de mevcuttur (Tavşancıl, vd., 2010).

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresinde bildirileri sunan kişilerin unvan dağılımları incelendiğinde, unvanını belirtmeyenlerin en fazla olduğu daha sonra sırasıyla araştırma görevlisi, yardımcı doçent doktor olduğu en az ise profesör doktor olduğu gözlemlenmiştir. Araştırma görevlerinin daha fazla sayıda bildiri yapmasında, makale yayınından önce bildiri çalışmaları yaparak tecrübe kazanılması, son yıllarda lisansüstü eğitim yönetmenliğinde yapılan değişiklilerde tez savunması için bildiri yapma zorunluluğu, alandaki hocalarla tanışma fırsatı gibi etkilerin söz konusu olduğu düşünülmektedir. Benzer olarak yardımcı doçent doktor sayısının da fazla olmasında doçentlik kriterlerinde bildiri yapma şartının bulunması gibi etkilerin olduğu söylenebilir.

Eğitimde ve Psikolojide ölçme ve Değerlendirme alanına özgü yapılan kongrelerde sunulan bildirilerin kurumlara göre dağılımı incelendiğinde, en fazla bildirinin Ankara Üniversitesi'nde çalışan akademisyenler tarafından sunulduğu daha sonra sırasıyla Hacettepe Üniversitesi, Gazi Üniversitesi ve Sakarya Üniversitesi olduğu belirlenmiştir. Eğitimde Ölçme ve Değerlendirme alanında yapılan lisansüstü tezler incelendiğinde benzer bir durumun olduğu bulunmuştur (Tavşancıl, vd., 2010; Şenyurt ve Özkan, 2017). Bu durumun nedenlerinden biri Ankara Üniversitesi ve Hacettepe Üniversitesinde hem yüksek lisans hem de doktora programının bulunması ve akademik personelin diğer üniversitelere göre daha fazla olmasından kaynaklanmaktadır.

Doküman analizi ile bildirilerin incelenmesi sonucunda, bildirilerin belli standartlara sahip olmadığı biçimsel olarak birçok eksik yanının bulunduğu (yazar adının olmadığı, anahtar kelimelerinin bulunmadığı, yöntem kısının eksik olduğu) belirlenmiştir. Bu bağlamda bundan sonra yapılacak olan kongrelerde bildirileri için belli standartların olması gerektiği düşünülmektedir. Eğitimde Ölçme ve Değerlendirme alanındaki eğilime bakıldığında Türkiye'deki eğiliminden çok farklı olmadığı ve birçok çalışmanın birbirini tekrar eder nitelikte olduğu belirlenmiştir. Bu bağlamda bundan sonra yapılacak kongrelerde güncel ve farklı konulardaki temaların belirlenmesi ve daha önce yapılmış çalışmaların benzerlerinin kabul edilmemesi gibi sınırlılıkların getirilmesi ilgili alana katkı getireceği düşünülmektedir.

KAYNAKÇA

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

Ayva, F. G., Ceyhan G. ve Doğan E. G. (2015, Nisan). *Türkiye'de eğitimde ölçme ve değerlendirme alanında yapılan doktora tezlerinin doküman analizi yöntemiyle incelenmesi*. 24. Ulusal Eğitim Bilimleri Kongresi, Niğde, Türkiye.

Bayram, E. (2011). *Öğretmenlerin ölçme ve değerlendirme yeterliklerinin incelenmesi* (Yüksek Lisans Tezi, Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara). https://tez.yok.gov.tr/UlusalTezMerkezi/ adresinden edinilmiştir.

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal, 9*(2), 27-40. Doi: 10.3316/QRJ0902027

Büyüköztürk, Ş., Çakmak, E., Akgün, Ö. E., Karadeniz, Ş. ve Demirel, F. (2012). *Bilimsel araştırma yöntemleri* (11. Basım). Ankara: Pegem Akademi.

Demirel, Ö ve Şahinel, M. (2006). *Türkçe ve sınıf öğretmenleri için türkçe öğretimi*. Ankara: Pegem Akademi.

Erdem, D. (2011). Türkiye'de 2005–2006 yılları arasında yayımlanan eğitim bilimleri dergilerindeki makalelerin bazı özellikler açısından incelenmesi: Betimsel bir analiz. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 2*(1), 140-147.

Erkuş, A. (2014). *Psikolojide ölçme ve ölçek geliştirme-I*. Ankara: Pegem Akademi.

Küçükoğlu, A. ve Ozan, C. (2013). Sınıf öğretmenliği alanındaki lisansüstü tezlere yönelik bir içerik analizi. *Uluslararası Avrasya Sosyal Bilimler Dergisi, 4*(12), 27-47.

Madge, J. (1965). *The tools of science an analytical description of social scince techniques.* ABD: Anchor.

O'Leary, Z. (2017). *The essential guide to doing your research project.* London:Sage.

Şenyurt, S. ve Özkan, Y. Ö. (2017). Eğitimde ölçme ve değerlendirme alanında yapılan yüksek lisans tezlerinin tematik ve metodolojik açıdan incelenmesi. *İlköğretim Online, 16*(2), 628-653.Doi: 10.17051/ilkonline.2017.304724

Tan, Ş. (2012). *Öğretimde ölçme ve değerlendirme KPSS el kitabı.* Ankara: Pegem Akademi.

Tavşancıl, E., Çokluk, Ö., Çıtak, G. G., Kezer, F., Yıldırım, Ö. Y., Bilican, S., Büyükturan, E. B., Şekercioğlu, G., Yalçın, N., Erdem, E. ve Özmen, D. T. (2010). *Eğitim bilimleri enstitülerinde tamamlanmış lisansüstü tezlerin incelenmesi (2000-2008).* Ankara Üniversitesi Bilimsel Araştırma Projesi Kesin Raporu, Ankara.

Tavşancıl, E., Güler, G. ve Ayan, C. (2014, Haziran). *2002-2012 Yılları arasında Türkiye'de geliştirilen bazı tutum ölçeği geliştirme çalışmalarının ölçek geliştirme süreci açısından incelenmesi*. IV. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi, Ankara, Türkiye.

Yin, R. K. (1994). *Case study research: Design and methods*. California: Sage Publications Inc.

Yaşar, Ş. ve Papatğa, E. (2015). İlkokul matematik derslerine yönelik yapılan lisansüstü tezlerin incelenmesi. *Trakya Üniversitesi Eğitim Fakültesi Dergisi, 5*(2), 113-124.

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research a content analysis and recommendations for best practices. *The Counseling Psychologist, 34*(6), 806-838. doi: 10.1177/0011000006288127

EXTENDED ABSTRACT

Introduction

Document analysis contains analysis of the written materials involving information about a fact or facts which are aimed to investigate. Document analysis is a data collection technique for almost every study. If there is no document, then there is no history (Madge, 1965). Which documents are important and can be used as a data source are closely related to the problem of the study. In the literature, there are studies in which postgraduate and doctorate dissertations together in the measurement and evaluation in education area were examined with document analysis but there are not any studies in which papers were analysed by document analysis. For this reason; to close this gap in the measurement and evaluation area, papers in this field were designed with case study that is one of the qualitative research methods and are examined with available data set document analysis.

General aim of this study is to determine how the trends in papers presented in The Measurement and Evaluation in Education and Psychology (MEEP) Congresses in Turkey are. Within this scope, themes, key words, writers' institutions, type of study, design of study, sample size, data type and data collection tool of papers presented in MEEP Congresses in 2008, 2010, 2012 and 2014 were investigated.

Method

Since the purpose of this study is to depicture how the trends in papers presented orally in MEEP congresses are and to interpret the existing situation, this study is a case study which is one of the qualitative research designs. Research population is composed of the papers presented in MEEP Congresses in 2008, 2010, 2012 and 2014. The research was directly done with the population and none of the sample methods were used. 248 papers were reviewed in the study. 24,2% (n=60) of this papers have been presented in The I. National Measurement and Evaluation in Education and Psychology Congress at Ankara University in 2008; 16,9% (n=42) in The II. National Measurement and Evaluation in Education and Psychology Congress at Mersin University in 2010, 21,0% (n=52) in The III. National Measurement and Evaluation in Education and Psychology Congress at Bolu Abant İzzet Baysal University in 2012 and 37,9% (n=94) in The IV. National Measurement and Evaluation in Education and Psychology Congress at Hacettepe University in 2014. To reach the papers presented in MEEP Congresses, academicians who coordinated these congresses were got in contact with as a result; the whole texts of the papers presented in the congresses in 2008, 2010, 2012 and the abstract texts of the congress in 2014 were obtained. Furthermore, The Administration of the Measurement and Evaluation Congresses in Education and Psychology Association was got contact with and necessary permissions were taken. In this context; 248 papers were totally reached and investigated by the researchers. As a tool of data collection, 'The Paper Examination Form' has been used. This form consists of three parts; the identity of presentation the theme of presentation, and the method of presentation.

Results and Discussion

It has been observed that the papers presented in MEEP congresses has generally been concentrated in the themes of 'Scale Development and Application', 'Statistical Tendencies Used in Educational Studies' and 'Large scale tests, selection and placement tests, high-stake test'. however, not many studies have been done in the themes of 'ÖSYM Practices' and 'Ethic'. Besides, it has been seen that 'Bias Studies' is the only subject theme that has regularly the rise according to years in studies done. Quantitative research type has generally been preferred in all the congresses. 33,1% (n=82) of the studies did not state the type of the research. Moreover, qualitative research and mixed research methods were not too much preferred. 55,6% (n=138) of the papers did not state the research model/design. Due to the fact that congress in 2014 has only short summaries, it is acceptable for the MEEP 2014 congress. When the ones whose research designs/models were not stated were not taken into account, it has been determined that 'Scanning' with the percent of 10,5% (n=26), 'Descriptive Research' with the percent of 10,1 (n=25) were the most preferred research types and 'Grounded Theory', 'Case Study', 'Activity Research', 'Theoretical Research' with the percent of 0,4 (n=1) were the least preferred research types. In the studies, samples sizes 'between 301 and 1000' and

'between 101 and 300' were generally utilized. Those studies mostly used 'Enhanced' data collection tool with the percent of 36,7% (n=91) and 'Prepared' data collection tool with the percent of 27,8% (n=69) and rarely used 'Prepared and Adapted' with 0,4% (n=1) were determined. 13,7% (n=34) of the papers presented do not have a data collection tool. In this sense, the most widely used data type has been stated as "Application Data" with 57,3% (n=142) and "Ready Made Data" with 19,8% (n=49). Moreover, 14,1% (n=35) of the presented papers that used data type has not been stated. The most used key words are "measurement and evaluation", "reliability" and "validity", respectively when all categorizes are taken into account in the analysis. In the analyses done according to titles of people presenting a presentation in the congress the category, 'Unspecified', has been stated as having the highest frequency. Later, it was found that researchers presented papers in congresses had the titles of "Res. Asst.", "Asst. Prof." and "Assoc. Prof. Dr." and those researchers presented papers are mostly from the universities "Ankara", "Hacettepe" and "Gazi", respectively on the basis of institution considering MEEP congresses. Besides, it was defined that some papers did not have writer's name, key words in congresses.

It is suggested that studies presented in MEEP Congresses should have a certain standard.

Ek 1: Bildiri İnceleme Formu

A-BİLDİRİNİN KÜNYESİ
1) Bildirinin Başlığı:
2) Yazar/Yazarlar:
3) Yazar/Yazarların Kurumu:
4) Yazar/Yazarların Unvanları:
5) Kullanılan Anahtar Kelimeler:
6) Hangi Kongrede Sunulduğu: ( )2008 EPOD ( )2010 EPOD ( )2012 EPOD ( )2014 EPOD
B-BİLDİRİNİN TEMASI
( ) Okullarda ölçme ve değerlendirme uygulamaları ( ) Geniş ölçekli sınavlar, seçme ve yerleştirme sınavları, yüksek riskli sınavlar ( ) Ölçme kuramları ve uygulamaları ( ) Teknoloji Destekli Eğitsel Ölçme ve Değerlendirme ( ) Eğitim araştırmalarında kullanılan istatistiksel yönelimler ( ) Program değerlendirme ve izleme ( ) Psikolojik danışma ve rehberlikte ölçme ve değerlendirme ( ) Okullarda etkililiğin ölçülmesi ve değerlendirilmesi ( ) ÖSYM Uygulamaları ( ) Yanlılık Çalışması ( ) Standart Belirleme ( ) Etik ( ) Test Eşitleme ( ) Ölçek Geliştirme ve Uyarlama ( ) Ölçekleme Çalışması ( ) Meta Analiz ( ) Diğer
C-BİLDİRİNİN YÖNTEMİ
7) Bildirinin Türü
( ) Nicel ( ) Nitel ( ) Karma ( ) Belirtilmemiş
8) Araştırma Modeli/deseni

( ) Tarama	( ) Eylem Araştırması
( ) Betimsel Araştırma	( ) Olgu bilim (Fenomoloji)
( ) Korelasyonel	( ) Örnek Olay
( ) Doküman Analizi	( ) Kuramsal Araştırma
( ) Meta-Analiz	( ) Durum Çalışması
( ) Temel Araştırma	( ) Kuram Oluşturma
( ) Deneysel	( ) Belirtilmemiş

9) Kullanılan Veri Türü
( ) Uygulama Verisi ( ) Hazır Veri ( ) Simülasyon Verisi ( ) Yok (Literatür çalışması) ( ) Belirtilmemiş
10) Veri Toplama Aracı
( ) Geliştirilmiş ( ) Uyarlanmış ( ) Hazır ( ) Yok ( ) Belirtilmemiş
11) Örneklem/Evren Büyüklüğü
( ) 1-30 arası ( ) 31-100 arası ( ) 101-300 arası ( ) 301-1000 arası ( ) 1000'den fazla ( ) Yok ( ) Belirtilmemiş