Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
ISSN: 1309 - 6575

Karma Testlerin Eşitlenmesinde MTK Eşitleme Yöntemlerinin Eşitlik Özelliği Korunumu Ölçütüne Göre Karşılaştırılması*

Comparing Different IRT Equating Methods Based on Test Equity Property for Mixed-Format Test Equating

Neşe ÖZTÜRK GÜBEŞ ** Hülya KELECİOĞLU***

Öz
Bu araştırmada, karma testlerin eşitlenmesinde madde tepki kuramına (MTK) dayalı gerçek-puan eşitleme ve gözlenen-puan eşitleme yöntemleri test eşitlemenin eşitlik özelliği korunumu ölçütüne dayalı olarak karşılaştırılmıştır. Araştırma, eşdeğer olmayan gruplar ortak test deseni kullanılarak yürütülmüştür. Araştırma verilerini, TIMSS 2011 8. Sınıf Türkiye örnekleminde yer alan öğrencilerin 5. ve 6. kitapçıkta yer alan matematik testlerine verdiği cevaplar oluşturmaktadır. Araştırma sonucunda, MTK gerçek-puan eşitleme yönteminin birinci-sıra eşitlik özelliğini korumada daha iyi performans gösterdiği, MTK gözlenen-puan eşitleme yönteminin ise ikinci-sıra eşitlik özelliğini korumada daha iyi performans gösterdiği bulunmuştur.

Anahtar Kelimeler: Madde Tepki Kuramı, gerçek-puan eşitleme, gözlenen-puan eşitleme, eşitlik özelliği.

Abstract
In this study, Item Response Theory (IRT) true-score equating and observed-score equating methods were compared based on preserving test equity property. The data used in this study were 8th grade mathematics test item responses which obtained from booklet 5 and 6 of Trends in International Mathematics and Science Study) (TIMSS) 2011 Turkey sample. Results showed that while IRT true-score equating outperformed IRT observed-score equating in terms preserving first-order equity, however IRT observed-score equating method IRT true- score equating in terms of second order equity property.

Key Words: Item Response Theory, true-score equating, observed-score equating, equity property.

GİRİŞ

Birçok geniş ölçekli sınavda çoktan seçmeli ve açık uçlu maddelerden oluşan karma testler kullanılarak ölçme ve değerlendirme yapılmaktadır. Örneğin Uluslararası Matematik ve Fen Eğilimleri Araştırması (TIMSS), Uluslararası Öğrenci Değerlendirme Programı (PISA) ve TOEFL gibi uluslararası düzeyde uygulanan sınavlar bunlardan bazılarıdır. Çoktan seçmeli ve açık uçlu maddelerin kendine göre bir takım üstünlükleri ve sınırlılıkları bulunmaktadır. Çoktan seçmeli testlerin üstünlükleri arasında kısa bir zaman diliminde çok sayıda soru sorularak geniş bir kapsam alanının ölçülebilmesi ve objektif puanlanarak güvenilir puanlar elde edilebilmesi gösterilebilir (Kim ve Walker, 2012; Livingston, 2009). Ancak çoktan seçmeli testler çoktan seçmeli öğretime yol açtığı, yazma ve üst düzey düşünme gibi becerilerin ölçülmesinde yetersiz kaldığı gerekçeleri ile eleştirilmektedir (Haladyna, 1997).

* Bu araştırma IV. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi'nde sözlü bildiri olarak sunulmuştur.

** Yrd. Doç. Dr. Neşe ÖZTÜRK GÜBEŞ, Mehmet Akif Ersoy Üniversitesi, Eğitim Fakültesi, Burdur-Türkiye, nozturk@mehmetakif.edu.tr
*** Prof. Dr. Hülya KELECİOĞLU, Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, hulyaebb@hacettepe.edu.tr

Üst düzey düşünme becerilerini ölçmek için açık uçlu maddelerin kullanılması önerilmektedir. Açık uçlu maddeler, öğrencinin verilen seçeneklerden doğru cevabı seçmek yerine kendi cevabını oluşturmasını gerektirir. Böylece çoktan seçmeli maddelerde var olan şansla doğru cevap verme olasılığı açık uçlu maddelerde oluşmaz. Ancak açık uçlu maddeleri objektif ve güvenilir olarak puanlamak çoktan seçmeli maddelere kıyasla oldukça zordur (Kirkpatrick, 2005). Çoktan seçmeli ve açık uçlu maddelerin sıraladığımız bu birtakım sınırlılıklarından dolayı TIMSS, PISA, TOEFL gibi birçok sınavda çoktan seçmeli ve açık uçlu maddeler birlikte kullanılmakta farklı madde türlerinin birbirine olan üstünlüğü bir avantaja dönüştürülmektedir.

PISA ve TIMSS gibi uluslararası geniş ölçekli sınavlarda test güvenliğini sağlamak, aynı soruları sormadan ölçmedeki değişimi takip etmek ya da her bir öğrencinin bütün sorulara cevap vermesini gerektirmeden daha fazla sayıda soruyu örneklemek için farklı test kitapçıkları kullanılmaktadır (Xu, 2009). Bu sınavlarda, test maddeleri bloklara bölünmektedir. Bu bloklar madde sayısı, kapsam ve madde formatı açısından her kitapçık birbirine eşdeğer olacak şekilde farklı kitapçıklara dağıtılmaktadır. Öğrenciler, herhangi bir kitapçığa tesadüfî olarak atanmaktadır. Her ne kadar kitapçıklar aynı yapıyı ölçmek için benzer kapsam ve istatistiksel özelliklere sahip olacak şekilde hazırlansa da test puanlarının aynı ölçekte ifade edilmesi bir başka deyişle eşitlenmesi gerekir. Eşitleme yapıldıktan sonra öğrencilerin hangi kitapçığı aldığının bir öneminin olmaması; geniş ölçekli sınavlar için bütün sonuçların öğrencilerin tümünün her bir uygulamada bütün soruları almış gibi oluşması gerekir (Xu, 2009). Dolayısıyla, her uygulamada birden fazla test kitapçığının kullanıldığı uluslararası sınavlar için test eşitleme önemli bir süreçtir.

Kolen ve Brennan (2004) eşitlemeyi; benzer içerik ve güçlük düzeyine sahip test formları arasındaki farklılıkları düzenleyerek, bu formlardan elde edilen puanların birbiri yerine kullanılmasını sağlayan istatistiksel bir süreç olarak tanımlamıştır. Bir eşitleme sürecine başlamadan önce ilk olarak veri toplamak için gerekli olan eşitleme desenine karar vermek gerekir. Kolen ve Brennan (2004) uygulamada veri toplamada kullanılan üç temel deseni şöyle sıralamışlardır; tek grup deseni (single group design), eşdeğer gruplar deseni (random groups design) ve eşdeğer olmayan gruplar ortak test (non-equivalent groups anchor test (NEAT)) deseni. Bu araştırma NEAT deseni kullanılarak yürütülmüştür. NEAT deseninde her grup bir test formunu alır. Test formlarını alan grupların eşdeğer yetenek dağılımına sahip olması varsayımı yoktur. Puan farklılıklarına formlar arasındaki farklılıklar sebep olduğu kadar grup farklılıklarının etkileri de karışır. Grupların farklılığı her iki formda yer alan ortak madde seti ile giderilmeye çalışılarak ortak maddeler aracılığı ile formlar birbirine eşitlenir (He, 2011).

NEAT deseni için klasik test kuramına ve madde tepki kuramına dayalı birçok test eşitleme yöntemi geliştirilmiştir. Tucker yöntemi (Gulliksen, 1950), Levine gerçek ve gözlenen puan yöntemi (Levine, 1955) ve Braun-Holland doğrusal yöntemi (Braun ve Holland, 1982), frekans kestirimi (Angoff, 1971) ve zincirleme (chained) eşit yüzdelikli eşitleme yöntemleri klasik test kuramına dayalı eşitleme yöntemlerinden bazılarıdır. MTK'na dayalı eşitleme yöntemleri, genel olarak MTK gerçek-puan eşitleme ve MTK gözlenen puan eşitleme olmak üzere iki başlık altında toplanabilir. MTK gerçek puan eşitleme yönteminde, bir test formuna ait θ yetenek düzeyi ile ilişkili gerçek puanın diğer formun θ yetenek düzeyi ile ilişkili gerçek puanına eşdeğer olduğu kabul edilir. MTK gözlenen-puan eşitleme yönteminde ise MTK modelleri kullanılarak her bir formun gözlenen puan dağılımları kestirildikten sonra eşit yüzdelikli eşitleme yöntemi ile puanlar birbirine eşitlenir (Kolen ve Brennan, 2004).

Eşitleme çalışmalarından elde edilen sonuçları değerlendirmek için birçok ölçüt geliştirilmiştir. Bu ölçütlerden biri eşitleme sonuçlarının eşitlik özelliğini sağlayıp

sağlamadığının belirlenmesidir. Eşitlemenin eşitlik özelliği ilk kez Lord (1980) tarafından tanımlanmıştır. Lord' un eşitlik özelliği ancak bir bireyin X ya da Y formunu almasının fark oluşturmadığı durumda sağlanır. Lord' un eşitlik özelliği iki test formu birebir birbirinin aynısı olmadığı sürece sağlanamayacaktır. Böyle bir durumda da zaten eşitleme yapmak gereksizdir (Kolen ve Brennan, 2004). Bu sebeple Divgi (1981) ve Morris (1982) zayıf eşitlik özelliği ya da birinci-sıra eşitlik (first-order equity) özelliği de denilen bir kavramı öne sürmüştür. Birinci-sıra eşitlik özelliği, eşitleme yapıldıktan sonra alternatif formların koşullu ortalamalarının eşit olmasını gerektirmektedir. Morris (1982), ayrıca eşitleme yapıldıktan sonra alternatif formlara ait ölçmenin koşullu standart hatasının (SEM) eşit olması durumunda sağlanan ikinci-sıra eşitlik (second-order equity) özelliğini önermiştir (akt. Harris ve Crouse, 1993). Matematiksel olarak birinci-sıra eşitlik ve ikinci-sıra eşitlik özellikleri sırayla (1) ve (2) numaralı eşitliklerinde olduğu gibi ifade edilebilir:

$$E[eq_Y(X) \mid \tau] = E(Y \mid \tau) \quad (1)$$

$$SEM_{eq_Y(x)\mid \tau} = SEM_{Y\mid \tau} \quad (2)$$

Eşitliklerde yer alan τ örtük yeteneği, $eq_Y(X)$ X formundaki bir puanı Y formunun ölçeğine dönüştüren eşitleme fonksiyonunu göstermektedir.

Birinci-sıra eşitlik özelliği Tong ve Kolen (2005) tarafından geliştirilen D_1 indeksi hesaplanarak değerlendirilebilir:

$$D_1 = \frac{\sqrt{\sum_i q_i\{E[Y \mid \theta_i] - E[e\hat{q}_Y(x) \mid \theta_i]\}^2}}{SD_Y} \quad (3)$$

D_1 eşitliğindeki; $E[Y \mid \theta_i]$ bir θ_i yetenek düzeyi için eski formun koşullu ortalamasıdır. $E[e\hat{q}_Y(x) \mid \theta_i]$ise bir θ_i düzeyi için eşitlenmiş puanların koşullu ortalaması; $q_i\theta_i$'deki quadrature ağırlığıdır. Son olarak, SD_Y Y formunun standart sapmasıdır. D_1 indeksi değeri küçüldükçe birinci-sıra eşitlik özelliği daha iyi korunmaktadır.

İkinci-sıra eşitlik özelliğinin değerlendirilmesine ilişkin D_2 indeksi yine Tong ve Kolen (2005) tarafından geliştirilmiştir:

$$D_2 = \frac{\sqrt{\sum_i q_i(SEM_Y \mid \theta_i - SEM_{e\hat{q}_Y(x)\mid} \theta_i)^2}}{SD_Y} \quad (4)$$

D_2 indeksinde yer alan $SEM_Y \mid \theta_i$ eski formu alan θ_i yetenek düzeyindeki bireylere ait koşullu ölçmenin standart hatası (SEM); $SEM_{\hat{q}_Y(x)} \mid \theta_i$ eşitlenen yeni formu alan θ_i yetenek düzeyindeki bireylere ait koşullu ölçmenin standart hatasıdır. Benzer şekilde D_2 indeksi değeri küçüldükçe ikinci-sıra eşitlik özelliği daha iyi korunmaktadır.

Araştırmanın Amacı

Bu araştırmanın amacı, karma testlerin eşitlenmesinde MTK gerçek-puan eşitleme ve MTK gözlenen-puan eşitleme yöntemlerini birinci-sıra eşitlik özelliği ve ikinci-sıra eşitlik özelliklerine göre karşılaştırmaktadır.

YÖNTEM

Araştırmanın Verileri

Araştırma verilerini, TIMSS 2011 Türkiye örnekleminde sınava giren 8. sınıf öğrencilerinin 5. ve 6. kitapçıkta yer alan matematik testlerine verdiği cevaplar oluşturmaktadır.

Araştırmanın çalışma grubunu, TIMSS 2011 Türkiye örnekleminde yer alan Kitapçık 5'e cevap veren 490 ve Kitapçık-6'ya cevap veren 494 öğrenci oluşturmaktadır.

Veri Toplama Aracı

Bu araştırma, eşdeğer olmayan gruplar ortak test (NEAT) deseni kullanılarak yürütülmüştür. Kitapçık-6'da yer alan matematik testi (X Formu) puanları MTK gerçek-puan eşitleme ve MTK gözlenen-puan eşitleme yöntemleri kullanılarak Kitapçık-5'teki matematik testi (Y Formuna) puanlarına eşitlenmiştir. Kitapçık- 5 ve 6'da yer alan maddeler madde formatına göre dağılımları Tablo 1'de görülmektedir.

Tablo 1. *Kitapçık-5 ve Kitapçık-6'da Yer Alan Maddelerin Dağılımı*

	Kitapçık-5	Ortak Maddeler	Kitapçık-6
Çoktan Seçmeli Madde	10	8	10
Açık Uçlu Madde (1-0)	3	9	3,1[a]
Açık Uçlu (0-1-2)	1	1[b]	1
Toplam	14	18	15

[a]*Testlerdeki madde sayısını eşitlemek için bu madde araştırmaya dâhil edilmemiştir. .*
[b]*1-0 şeklinde puanlanan iki açık uçlu maddenin puanlarının birleştirilmesiyle elde edildiği için dâhil edilmemiştir.*

Araştırma, Kitapçık-5 ve Kitapçık-6'da yer alan 10 çoktan seçmeli ve 4 açık uçlu madde ile ortak olan 8 çoktan seçmeli ve 9 açık uçlu madde kullanılarak yürütülmüştür. Araştırma kapsamında eşitlenen her bir test toplam 31 maddeden oluşmaktadır. Bir test formundan elde edilebilecek en yüksek puan 32'dir.

Verilerin Analizi

Verilerin analizi dört basamakta tamamlanmıştır. Birinci aşamada MTK modelleri kullanılarak madde parametreleri kestirilmiştir. Çoktan seçmeli maddeler için 3-parametreli lojistik model (3PL), 1-0 şeklinde puanlanan açık uçlu maddeler için 2-parametreli lojistik model (2PL) ve 0,1,2 şeklide kısmi puanlanan açık uçlu maddeler için ise genelleştirilmiş kısmi puan modeli (GPCM) kullanılmıştır. Madde parametreleri PARSCALE (Muraki & Bock, 2003) bilgisayar programı kullanılarak her bir form için ayrı ayrı kestirilmiştir. Veri analizinin ikinci aşamasında, her bir forma ait kestirilen madde parametrelerini aynı ölçekte ifade etmek için ölçek dönüştürme işlemi uygulanmıştır. Ölçek dönüştürme, Haebara (Haebara, 1980) yöntemi kullanılarak STUIRT (Kim ve Kolen, 2004) bilgisayar programında gerçekleştirilmiştir. Veri analizinin üçüncü aşamasında, POLYEQUATE (Kolen, 2004a) bilgisayar programı kullanılarak MTK gerçek-puan eşitleme ve MTK gözlenen-puan eşitleme yöntemleri ile test puanları birbirine eşitlenmiştir. Veri analizinin son aşamasında ise elde edilen eşitleme sonuçlarını birinci-sıra eşitlik özelliğine göre değerlendirmek için D_1, ikinci-sıra eşitlik özelliğine göre değerlendirmek için D_2 indeksi hesaplanmıştır. D_1 ve D_2 indekslerini hesaplamak için gerekli olan koşullu ortalama ve standart sapmalar POLYCSEM (Kolen, 2004b) bilgisayar programı aracılığı ile elde edilmiştir.

BULGULAR

Araştırmada kullanılan Kitapçık-5 ve Kitapçık-6 matematik testlerinin betimsel istatistikleri Tablo 2'de verilmiştir.

Tablo 2. *Matematik Testlerine Ait Betimsel İstatistikler*

	N	$\bar{X}$	SD	Çarpıklık K.	Basıklık K.

Kitapçık 5	490	14.37	7.93	0.442	-0.873
Kitapçık 6	494	13.49	7.74	0.578	-0.649

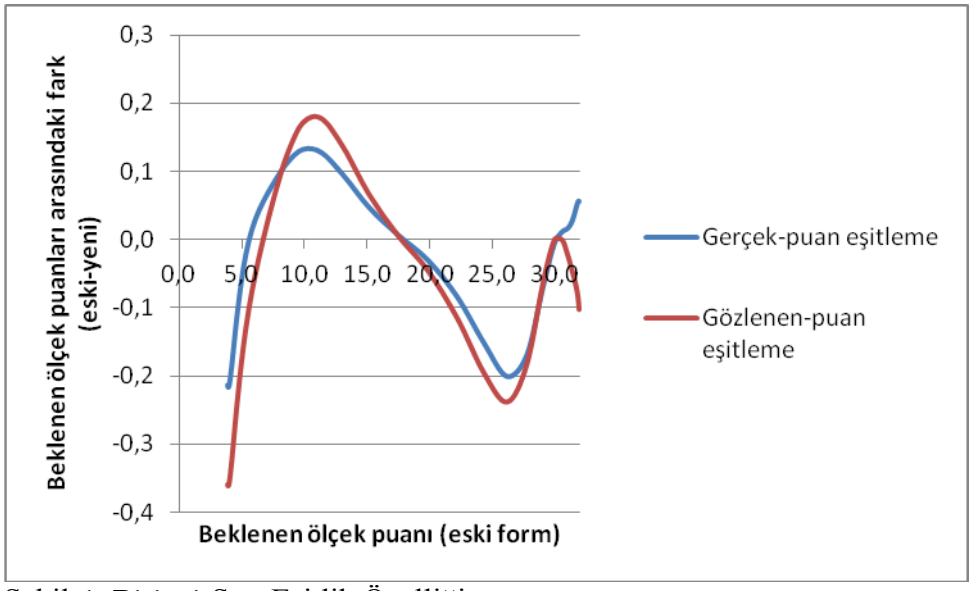
Tablo 2'de de görüleceği üzere testlerin ortalamalarına dayalı olarak Kitapçık-6'da yer alan matematik testinin $(\bar{X} = 13.49)$ Kitapçık-5'te yer alan matematik testinden $(\bar{Y} = 14.37)$ daha zor olduğunu söyleyebiliriz.

Araştırma kapsamında MTK gerçek-puan eşitleme ve MTK gözlenen-puan eşitleme sonuçlarını değerlendirmek için D_1 ve D_2 indeksleri hesaplanmıştır. Elde edilen bulgular Tablo 3'te verilmiştir.

Tablo 3. *Hesaplanan D_1 ve D_2 İndeks Değerleri*

Değerlendirme Ölçütü	MTK Gerçek-Puan Eşitleme	MTK Gözlenen-Puan Eşitleme
D_1	0.0125	0.0171
D_2	0.0216	0.0159

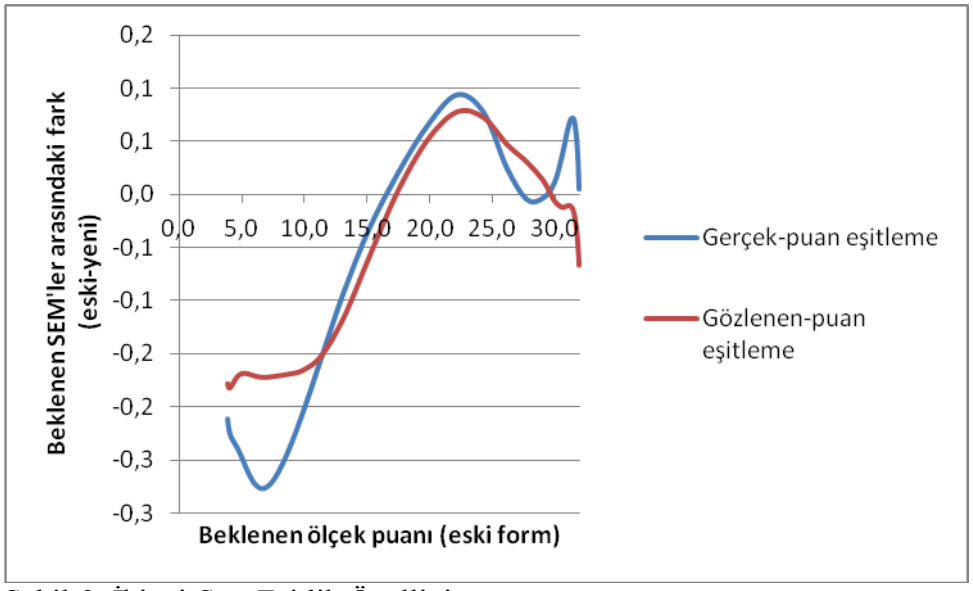
Tablo 3'te yer alan bilgilere dayalı olarak MTK gerçek-puan eşitlemenin birinci-sıra eşitlik özelliğini korumada $(D_1 = 0.0125)$ MTK gözlenen-puan eşitlemeden daha iyi performans gösterdiğini, ikinci-sıra eşitlik özelliğini korumada ise MTK gözlenen-puan eşitlemenin $(D_2 = 0.0159)$ daha iyi olduğunu söyleyebiliriz. Birinci-sıra eşitlik özelliğini puan ölçeği boyunca değerlendirmek için eski formun beklenen ölçek puanına karşılık çizilen beklenen ölçek puanları arasındaki farkın grafiği Şekil 1'de görülmektedir.



Şekil-1. Birinci-Sıra Eşitlik Özelliği

Birinci-sıra eşitlik özelliği mükemmel bir şekilde sağlanmış olsa idi her iki eşitleme yöntemi sonucu elde edilen beklenen ölçek puanları arasındaki fark sıfır olacaktı. Şekil-1'deki grafikte de görüleceği üzere gerçek puan eşitleme yöntemine göre 6, 18, 30 ve 31 puanlarında; gözlenen puan eşitleme yöntemine göre 7, 18, 30 ve 31 puanlarında beklenen ölçek puanları arasındaki fark sıfırdır. Puan ölçeğinin geri kalan kısmında genel olarak gerçek-puan eşitleme yönteminin gözlenen-puan eşitleme yöntemi ile karşılaştırıldığında daha küçük fark puanlarına sahip olduğunu ve birinci-sıra eşitlik özelliğini daha iyi

koruduğunu söyleyebiliriz. İkinci-sıra eşitlik özelliğini puan ölçeği boyunca değerlendirmek üzere beklenen ölçek puanlarına karşılık beklenen SEM'ler arasındaki farkın grafiği çizilmiş ve Şekil-2'de verilmiştir.



Şekil-2. İkinci-Sıra Eşitlik Özelliği

Benzer şekilde her iki eşitleme yöntemi ile elde edilen eşitleme sonuçları ikinci-sıra eşitlik özelliğini mükemmel bir şekilde sağlamış olsa idi beklenen SEM'ler arasındaki farkın sıfır olması gerekirdi. Şekil-2'deki grafik incelendiğinde, gerçek-puan eşitleme yöntemine göre 17, 28, 29, 30 ve 32 puanlarında; gözlenen-puan eşitleme yöntemine göre 18, 29, 30 ve 31 puanlarında aradaki fark sıfırdır. Puan ölçeğinin geri kalan kısmında genel olarak gözlenen-puan eşitleme yönteminin gerçek-puan eşitleme ile karşılaştırıldığında daha düşük fark puanlarına sahip olduğunu ve ikinci-sıra eşitlik özelliğini daha iyi koruduğunu söyleyebiliriz.

SONUÇLAR ve TARTIŞMA

Elde edilen bulgulara dayalı olarak MTK gerçek puan eşitlemenin birinci-sıra eşitlik özelliğini korumada MTK gözlenen-puan eşitlemeden daha iyi performans gösterdiği; MTK gözlenen-puan eşitlemenin ise ikinci-sıra eşitlik özelliğini korumada MTK gerçek-puan eşitlemeden daha iyi performans gösterdiği sonucuna ulaşılmıştır. Elde edilen bulgular literatürde yapılmış olan eşitlik özelliğinin değerlendirildiği diğer araştırmalarla paralellik göstermektedir (Tong & Kolen, 2005; Lee, Lee & Brennan, 2012). MTK gözlenen-puan eşitleme eşit yüzdelikli eşitleme yöntemini kullanarak gözlenen puan dağılımlarını olabildiğince birbirine benzer yapmaya çalışırken, MTK gerçek-puan eşitleme yöntemi gerçek puanlar aracılığı ile puan dağılımlarını eşitlemektedir ve gözlenen puanlara dayalı değildir. Bu durum MTK gözlenen puan eşitlemenin ikinci-sıra eşitlik özelliğini korumada daha iyi performans göstermesinin bir nedeni olarak gösterilebilir. MTK gerçek-puan eşitleme gerçek puanları eşitlediği ve birinci-sıra eşitlik özelliği gerçek puanlar arasındaki ilişkilere dayalı olarak tanımlandığı için MTK gerçek-puan eşitlemenin birinci-sıra eşitlik özelliğini daha iyi koruması ise beklendiktir.

Birinci-sıra eşitlik özelliği testin her bireye adil olması üzerine odaklanırken, ikinci-sıra eşitlik özelliği ölçmenin doğruluğu üzerine odaklanmaktadır (He, 2011). Eğer amaç bireylerin bir testin A ya da B formunu almaktan dolayı sahip olacakları avantaj ya da dezavantajı önlemek ise birinci-sıra eşitlik özelliğini korumak daha önemli hale gelir. Böyle

bir durumda birinci-sıra eşitlik özelliğini daha iyi koruyan MTK gerçek-puan eşitleme yöntemi kullanılarak testler eşitlenmesi önerilebilir. Ancak eğer öncelik ölçme hatasını en aza indirmek ise böyle bir durumda ikinci-sıra eşitlik özelliğini daha iyi koruyan MTK gözlenen-puan eşitleme yöntemi kullanılarak testlerin eşitlenmesi önerilebilir.

KAYNAKLAR

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp.508-600).Washington DC: American Council on Education.

Braun, H.I., & Holland, P.W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P.W. Holland and D.B. Rubin (Eds.), *Test equating* (pp. 9-49). New York: Academic.

Divgi, D.R. (1981). Two direct procedures for scaling and equating tests with item response theory. Paper presented at the Annual Meeting of American Educational Research Association, Los Angels.

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

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, 22(3). 144-149.

Haladyna, T.M. (1997). Writing test items to evaluate higher order thinking. Boston: Allyn & Bacon.

Harris, D.J. & Crouse, J.D. (1993). A study of criteria used in equating. *Applied Measurement in Education*, 6, 195-240.

He, Y. (2011). *Evaluating equating properties for mixed-format tests.* Unpublished doctoral dissertation, University of Iowa.

Kim, S. & Kolen, M. J. (2004). *STUIRT* [computer program]. Iowa City, IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Iowa.

Kim, S. ,& Walker, M. (2012). Determining the anchor composition for a mixed format test: Evaluation of subpopulation invariance of linking functions. *Applied Measurement in Education*, 25, 178-195.

Kirkpatrick, R.K. (2005). *The effects of item format in common item equating*. Unpublished doctoral dissertation, University of Iowa.

Kolen, M.J. (2004a). *POLYEQUATE* [computer program].Iowa City,IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Iowa.

Kolen, M.J. (2004b). *POLYCSEM* [computer program]. Iowa City, IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Iowa.

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

Lee, E., Lee, W-C., and Brennan R. L. (2012). *Exploring equity properties in equating using AP examinations*. (College Board Research Report No. 4).

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

Livingston, S.A. (2009). Constructed-Response test questions: Why we use them; How we score them (R & D Connections, No. 11). Princeton, NJ: Educational Testing Service.Retrived from http://www.ets.org/Media/Research/pdf/RD_Connections11.pdf .

Morris, C.N. (1982). On the foundations of test equating. In p. W. Holland & D. B. Rubin (Eds.) *Test equating* (pp. 169-191). New York: Academic.

Muraki, E., & Bock, R. (2003). *PARSCALE 4.1*. Chicago, IL:Scientific Software International.

Tong, Y., & Kolen, M. J. (2005). Assessing equating results on different equating criteria. *Applied Psychological Measurement, 29(6)*, 418-432.

Xu, Y. (2009). *Measuring change in jurisdiction achievement over time: Equating issues in current international assessment programs*. Unpublished doctoral dissertation, University of Toronto.

EXTENDED ABSTRACT

Introduction

Many testing programs have been using mixed format tests, which contain a mixture of multiple-choice (MC) items and constructed-response (CR) items. Some examples of large

scale standardized assessments that use mixed format exams are Test of English as a Foreign Language (TOEFL), Trends in International Mathematics and Science Study (TIMSS), OECD Programme for International Student Assessment (PISA), Advanced Placement (AP), and National Assessment of Educational Progress (NAEP). MC and CR items have their own advantages and limitations. MC items efficiently measure a broad range of content, afford high reliability, achieve objectivity in scoring, and are fast and inexpensive to score. However, CR items measure higher-order thinking skills more effectively than MC items. For example, A MC test for mathematics students can determine whether they can solve many kinds of problems—but it can't determine whether they can construct a mathematical proof. On the other hand, CR tests typically cover a narrower range on content, are more time-consuming and expensive to score, and subjectively scored (Livingston, 2009).

So many test programs use mixed format tests to benefit from their strengths and compensate for their weakness.

In large-scale testing programs, test administration often occurs at different times and/or at different locations, more than one form of a test is needed. The use of different test forms leads another concern: test forms may differ somewhat difficulty. In order to assure test score comparability and interchangeability, we need to equate test scores (Kolen and Brennan, 2004).

There are many different procedures that have been used to evaluate test equating results. One of these methods is assessing how well equating properties are preserved. Lord (1980) proposed Lord's equity property of equating. According to Lord's equity property it must be a matter of indifference to each examinee takes whether Form X or Form Y. "Lord's equity property holds if examinees with given true score have the same distribution of converted scores on Form X as they would on Form Y." (Kolen&Brennan, 2004, p. 10). Lord also showed that it is impossible to satisfy the equity property unless the two forms are essentially identical in which case equating is not necessary or reliability of two forms are 1 which is unrealistic. As a practical solution to this contradiction, Morris (1982) suggested a "weak equity property" or "first-order equity" (FOE). According to FOE, the conditional means are equal for alternate forms after equating (Lee, Lee, and Brennan, 2012). Morris (1982) also proposed the second-order equity property, which holds if the conditional standard errors of measurement (SEM) are equal for the alternated forms after equating.

The purpose of this study is to compare IRT true-score and observed-score equating methods based on first-order equity (FOE) and second-order equity (SOE) properties for mixed format tests.

Method

The data used in this study were 8th grade mathematics test item responses which obtained from TIMSS 2011 Turkey (N=6928) sample. In this study, nonequivalent groups anchor test design was used. The booklet 5 was chosen as base form (old form) and the booklet 6 was chosen as new form. The analyses were conducted with 490 examinees' item responses who took Booklet 5 and 494 examinees' item responses who took Booklet 6.

In this study, data analyses were conducted in four steps. In the first step; 3PL IRT model was used for multiple-choice items, 2PL model was used for 1-point constructed response items and GPCM model were used for 2-point constructed response items. PARSCALE (Muraki & Bock, 2003) program was used for calibrating data. Item parameters for two forms were estimated separately. After item parameters of both forms estimated, the Haebara method was used to transform the estimated parameters and the quadrature points of the posterior distribution on the new form scale to the old form scale. The computer program STUIRT (Kim & Kolen, 2004) was used for the scale

transformation. IRT true-score and observed-score equating were conducted by using POLYEQUATE (Kolen, 2004a) computer program. The item parameters and posterior distribution of the old form obtained from IRT calibration step and the transformed item parameters and posterior distribution of the new form obtained from the scale transformation step were input the program. The first-order and second-order equity properties were assessed under IRT framework. The computer program POLYCSEM (Kolen, 2004b) was used for calculating expected conditional scores and conditional SEMs. After then, first-order equity (D_1) and second-order equity (D_2) indices were calculated.

Results and Discussion

While the IRT true-score method yielded the smallest D_1 value, IRT observed-score equating method yielded the smallest D_2 value. We can say that IRT true-score equating better preserved first-order equity than IRT observed-score equating method. On the other hand, for preserving second-order equity IRT observed –score equating method performed better than IRT true-score equating.

The results were consistent with previous studies (Tong &Kolen, 2005; He, 2011; Lee, et al., 2012). IRT true-score equating outperformed IRT observed-score equating in terms of preserving FOE, however IRT observed-score equating method outperformed IRT true-score equating method in terms of SOE. According to He (2011), the reason is IRT observed-score method incorporates equipercentile equating to make observation distributions as similar as possible, whereas IRT true score equating is not based on observed distributions. Also IRT true score equating equates true scores and first order equity is defined with respect to the true score relationships.

Generally, FOE focuses on test fairness at the level of individual examinee and second order equity investigates measurement precision. He (2011) indicated that if the purpose is to prevent an examine from being advantaged or disadvantaged by being administered one test form versus another, then satisfying FOE is of concern, and IRT true-score equating method can be recommended over observed-score equating method. If the priority is improving measurement error and decreasing measurement errors then preserving SOE becomes more important IRT observed-score equating can be recommended.