Ankara University, Journal of Faculty of Educational Sciences, Year: 2015, Vol: 48, No: 1, 199-214

Comparing Different Test Equating Methods Based on Equity Property*

Neşe ÖZTÜRK GÜBEŞ** **Hülya KELECİOĞLU*****

ABSTRACT. The purpose of this study was to compare the performance of three equipercentile and two item response theory (IRT) equating methods based on equity property. In this study, the social science test scores of Booklet A and Booklet C from 2009 the Assessment of Student Achievement were equated under equivalent groups design. The study group consisted of 15.173 and 14.365 9th grade students which took Booklet A and Booklet C. The equity property was evaluated based on first-order equity (FOE) which requires the same conditional means after equating and second-order equity (SOE) which requires the same conditional error of measurement. Results showed that IRT true score equating method best preserved FOE property and IRT observed score equating method best preserved SOE property.

Key Words: Test equating, equity property, item response theory.

* This study was presented as a poster on April 2 -3, 2014 in 76th Annual Meeting of National Council on Measurement in Education, Philadelphia.

** Assistant Professor, Mehmet Akif Ersoy University, Faculty of Education Department of Educational Sciences, Burdur-Turkey, nozturk@mehmetakif.edu.tr.

*** Prof. Dr., Hacettepe University, Faculty of Education Department of Educational Sciences, Ankara-Turkey, hulyaebb@hacettepe.edu.tr.

SUMMARY

Purpose and Significance: Assessing equity property of equating methods can be valuable evidence for evaluating how well equating methods adjust difficulty difference between forms. The purpose of this study was to compare the performance of three equipercentile and two item response theory (IRT) equating methods based on equity property with using real data.

Method: The data used in this study were obtained from the Assessment of Student Achievement which was a national assessment that used for assessing secondary grade students' achievement in Turkish, mathematics, science, social science and English domains in Turkey. In this study, we chose to use the data from social science tests in Booklet A and Booklet C. The equating was conducted under equivalent groups design. The sample size for Form A was 15,173 and for Form C was 14,365. In this study, the raw scores on the Form C were equated to Form A using three equipercentile equating methods (unsmoothed, presmoothed (log-linear method with C=5), and postsmoothed (s=0.01) equipercentile equating) and two IRT equating methods (IRT true- and observed-score equating). To evaluate how well equity properties were preserved, conditional means and conditional SEMs were obtained assuming 3PL IRT model. Then, FOE and SOE properties were evaluated by calculating D_1 index for FOE, D_2 index for SOE.

Results: The results showed that while IRT true-score equating method performed best in terms of preserving FOE ($D_1 = 0.038$), IRT observed-score equating method performed best in preserving SOE ($D_2 = 0.041$) property. Equipercentile equating methods performed similar in preserving both equity properties.

Discussion and Conclusion*:* Overall, our findings were consistent with previous researches (Tong & Kolen, 2005; Lee et al., 2010; Lee & Brennan, 2012). As Lee and Brennan (2012) indicated equity property of equating methods are influenced by underlying psychometric model assumed. When IRT framework is used the IRT true-score equating method may have more advantage compare to other methods for preserving FOE because the true - score distributions are based on the IRT model. As He (2011) noted if the main purpose is to prevent individual examinees from being advantaged or disadvantaged by being administered one test form versus another, then satisfying FOE might be important. In this case, IRT true-score equating method might be suggested. However, if the primary interest is to improve measurement precision and decreasing measurement errors, then satisfying SOE equity is the main concern. In this case, IRT observed score equating method might be preferred.

Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, Yıl: 2015, Cilt: 48, Sayı: 1, 299-214

Farklı Test Eşitleme Yöntemlerinin Eşitlik Özelliği Ölçütüne Göre Karşılaştırılması*

Neşe ÖZTÜRK GÜBEŞ** **Hülya KELECİOĞLU*****

ÖZ. Bu araştırmanın amacı üç farklı eşit yüzdelikli eşitleme ve iki farklı Madde Tepki Kuramı'na (MTK) dayalı test eşitleme yöntemlerini test eşitlemenin eşitlik özelliği ölçütüne göre karşılaştırmaktır. Araştırmada, 2009 Öğrenci Başarılarının Belirlenmesi Sınavı'nda A ve C kitapçığında yer alan sosyal bilimler testi puanları eşdeğer gruplar deseni altında eşitlenmiştir. Araştırmanın çalışma grubunu A kitapçığını alan 15.173 ve C kitapçığını alan 14.365 dokuzuncu sınıf öğrenci oluşturmuştur. Eşitlik özelliği eşitleme yapıldıktan sonra alternatif formlardan elde edilen puanların koşullu ortalamalarının eşit olmasını gerektiren birinci-sıra eşitlik (BSE) ve koşullu ölçmenin standart hatasının eşit olmasını gerektiren ikinci-sıra eşitlik (İSE) özelliği ölçütlerine göre değerlendirilmiştir. Araştırmanın sonucunda, BSE özelliğini en iyi koruyan yöntemin MTK gerçek puan eşitleme, İSE özelliğini en iyi koruyan yöntemin MTK gözlenen puan eşitleme yöntemi olduğu görülmüştür.

Anahtar Kelimeler: Test eşitleme, eşitlik özelliği, madde tepki kuramı

---

* Bu araştırma, 2-3 Nisan 2014 tarihleri arasında Philadelphia'da gerçekleşen "Annual Meeting of National Council on Measurement in Education" da poster olarak sunulmuştur.

** Yrd. Doç. Dr., Mehmet Akif Ersoy Üniversitesi, Eğitim Fakültesi Eğitim Bilimleri Bölümü, Burdur-Türkiye, neseozturk07@gmail.com.tr .

*** Prof. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi Eğitim Bilimleri Bölümü, Ankara-Türkiye, hulyaebb@hacettepe.edu.tr.

GİRİŞ

Birçok geniş ölçekli sınavda, güvenlik nedeniyle ya da öğrencilere yılın belirli dönemlerinde birden fazla değerlendirilme şansını tanımak amacıyla bir testin farklı formları kullanılmaktadır. Her ne kadar test geliştirme uzmanları içerik ve istatistiksel özellikler bakımından birbirine olabildiğince paralel test formları geliştirmeye çalışsa da test formları en azından güçlük düzeyi bakımından birbirinden farklılık gösterecektir. Bir sınavın farklı iki formunu alarak sınav olan iki öğrenciden A öğrencisinin B öğrencisinden daha yüksek puan almasının sebeplerinden biri A öğrencisinin daha başarılı olması olabileceği gibi A öğrencisinin aldığı test formun B öğrencisinin aldığı formdan daha kolay olması da olabilir. Dolayısıyla bir testin farklı formlarından elde edilen puanlarını karşılaştırabilmek ve bir testin kolay formunu alan öğrencinin avantajını ya da zor formunu alan bireyin dezavantajını önlemek için test puanlarının eşitlenmesi gerekir. Kolen ve Brennan (2004) test eşitlemeyi farklı test formlarından elde edilen puanların birbiri yerine kullanılabilmesini sağlamak amacıyla yürütülen istatistiksel süreç olarak tanımlamışlardır.

Bir test eşitleme çalışması genellikle üç basamak takip edilerek yürütülür (Harris ve Crouse, 1993). Test eşitleme çalışmasının ilk basamağını veri toplamak için gerekli olan test eşitleme desenin seçilmesi oluşturur. Yaygın olarak kullanılan üç test eşitleme desenini Kolen ve Brennan (2004) tek grup deseni, eşdeğer gruplar deseni ve eşdeğer olmayan gruplar ortak test deseni olarak sıralamışlardır. Tek grup deseninde, bir testin farklı formları aynı gruba uygulanır. Bu desenin olumsuz yanı formların uygulanma sırasının birbirinden etkilenmesidir. Eşdeğer gruplar deseninde ise her bir test formu yetenek dağılımları eşdeğer olan iki ayrı gruba uygulanır. Bu desende öğrenciler alacakları forma tesadüfî olarak atanırlar. Öğrencileri gruplara tesadüfî atama yollarından biri test formlarını paketlerken sarmal bir sürecin uygulanmasıdır. Test formları ilk öğrenci Form X'i, ikinci Form Y'i, üçüncü Form X…'i alacak şekilde dağıtılır. Bu sarmal süreç sayesinde tesadüfî olarak eşdeğer iki grup oluşturulur. Üçüncü test eşitleme deseni, eşdeğer olmayan gruplar ortak test desenidir. Bu desende test formlarında ortak maddeler vardır ve her bir öğrenci sadece bir test formunu alır. Farklı test formlarını alan grupların eşdeğer olması gerekmez. Gruplardan ve formlardan kaynaklanan farklılıklar her iki formda da ortak olan maddeler aracılığı ile düzenlenir (Kolen ve Brennan, 2004).

Uygun test eşitleme desenine karar verilip, veri toplandıktan sonra ikinci aşama test formlarının eşitleneceği eşitleme yöntemi ya da yöntemlerinin seçilmesidir. Test eşitleme yöntemleri geleneksel eşitleme yöntemleri ve Madde Tepki Kuramı (MTK)'na dayalı yöntemler olmak

üzere iki ana başlıkta toplanabilir. Ortalama eşitleme, doğrusal eşitleme ve eşit yüzdelikli eşitleme geleneksel eşitleme yöntemleridir. Ortalama eşitlemede formların sadece ortalamaları eşitlenirken doğrusal eşitlemede ortalamaları ve standart sapmaları eşitlenir. Eşit yüzdelikli eşitlemede ise ortalama ve standart sapmalarının yanında formların dağılımları da (çarpıklık, basıklık vb.) eşitlenir (Kolen, 1988). Eşit yüzdelikli eşitleme yönteminde aynı yüzdelik sıraya sahip olan puanların eşdeğer olduğu kabul edilir (Kolen ve Brennan, 2004). Gerçek test uygulamalarında puan dağılımlarında oluşan düzensizlikler, eşit yüzdelikli eşitlemede birtakım problemlere (eşitleme sonuçlarının genellenememesi gibi) neden olmaktadır. Bu problemle başa çıkmanın bir yolu puan dağılımlarının düzgünleştirilmesidir (Livingston, 2004). Düzgünleştirme işlemi puanlar eşitlenmeden önce puan dağılımına uygulanırsa ön-düzgünleştirme (presmoothing), eşitleme yapıldıktan sonra eşitlenmiş puanlara uygulanırsa son-düzgünleştirme (postsmoothing) adını alır (Kolen ve Brennan, 2004).

MTK'ya dayalı eşitleme yöntemlerini Kolen ve Brennan (2004) MTK gerçek puan eşitleme ve gözlenen puan eşitleme olmak üzere iki sınıfa ayırmıştır. MTK gerçek puan eşitleme yönteminde, her iki formda aynı Θ yetenek düzeyine denk gelen puanların eşit olduğu kabul edilir. MTK gözlenen puan eşitlemede, her bir formun gözlenen puan dağılımı MTK modelleri kullanılarak kestirildikten sonra eşit yüzdelikli eşitleme yöntemleri kullanılarak puanlar eşitlenir. Bu araştırma, literatürde en çok kullanılan eşit yüzdelikli eşitleme ve MTK'na dayalı eşitleme yöntemleri kullanılarak yürütülmüştür (Andrews, 2011; He, 2011; Lee, Lee ve Brennan, 2010; Tong ve Kolen, 2005).

Test eşitlemenin son aşaması yürütülen sürecin kontrol edilmesidir. Literatürde test eşitleme sonuçlarının değerlendirmek için birçok ölçüt geliştirilmiştir. Bunlardan bir kısmını Harris ve Crouse (1993) RMS (kareler ortalamasının karekökü (root mean square)), RMSD (kareler farkının ortalamasının karekökü (root mean square difference)), MSD (işaretli farkların ortalaması (mean signed difference)), MAD (mutlak farkların ortalaması (mean absolute difference)), MSE (hata karelerinin ortalaması (mean squared error)), yanlılık, standart hatalar olarak sıralamışlardır. Eşitleme sonuçlarını değerlendirmede kullanılan bir diğer ölçüt eşitlik özelliğidir. 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. Diğer bir ifade ile Lord'un eşitlik özelliği iki test formu birbirinin aynısı olmadığı sürece sağlanamayacaktır. Böyle bir durumda ise testleri eşitlemeye gerek kalmayacaktır (Kolen ve Brennan, 2004). Bu nedenle Morris (1982) eşitlik özelliğinin daha esnek bir tanımı olan birinci-sıra eşitlik (BSE) ya da zayıf

eşitlik olarak da adlandırılan bir kavramı öne sürmüştür. BSE özelliği, eşitleme yapıldıktan sonra alternatif formların koşullu ortalamalarının eşit olmasını gerektirir. Morris (1982), ayrıca eşitleme yapıldıktan sonra alternatif formlara ait ölçmenin koşullu standart hatasının (standart error of measurement, SEM) eşit olması durumunda sağlanan ikinci-sıra eşitlik (İSE) özelliğini de önermiştir. BSE ve İSE eşitlik özellikleri Tong ve Kolen (2005) tarafından geliştirilen D_1 ve D_2 indeksleri hesaplanarak değerlendirilebilir:

$$D_1 = \frac{\sqrt{\sum_i q_i\{E[Y \mid \Theta_i] - E[e\hat{q}_Y(x) \mid \Theta_i]\}^2}}{SD_Y} \quad (1)$$

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

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 karelemelerin (quadrature) ağırlığıdır. Son olarak, SD_Y Y formunun standart sapmasıdır. D_2 indeksinde yer alan $SEM_Y \mid \Theta_i$ eski formu alan Θ_i yetenek düzeyindeki bireylere ait ölçmenin koşullu standart hatası (SEM); $SEM_{\hat{q}_Y(x)} \mid \Theta_i$ ise eşitlenen yeni formu alan Θ_i yetenek düzeyindeki bireylere ait ölçmenin koşullu standart hatasıdır. D_1 ve D_2 değerleri küçüldükçe BSE ve İSE özellikleri daha iyi korunmaktadır.

Ülkemizde, öğrenci başarısının belirlenmesinde uluslararası düzeyde yapılan PISA, TIMSS ve PIRLS sınavlarının yanında ulusal düzeyde Milli Eğitim Bakanlığı (MEB) tarafından ilköğretim ve ortaöğretim düzeylerinde okul öğrenmelerinin izlenmesi amacıyla Öğrenci Başarılarının Belirlenmesi Sınavı (ÖBBS) yapılmaktadır. ÖBBS, ilköğretim (4, 5, 6, 7 ve 8. sınıflarda) ve ortaöğretim (9. ve 10. sınıflarda) düzeyinde üçer yıllık aralıklarla yapılan bir sınavdır. İlköğretim düzeyinde ilk uygulama 2002 yılında, ortaöğretim düzeyinde ilk uygulama 2006 yılında gerçekleştirilmiştir. ÖBSS'nin ortaöğretim uygulaması ortak zorunlu ve seçmeli sayılabilecek Türk edebiyatı, dil ve anlatım, matematik, fen bilimleri (fizik, kimya ve biyoloji), sosyal bilimler (tarih ve coğrafya) ve İngilizce dersleriyle sınırlı tutulmuştur (Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı [EARGED], 2010).

ÖBBS'nin uygulanması sırasında kopyanın önüne geçmek ve aynı soruları sormadan daha fazla kapsam alanını örnekleyebilmek amacıyla A, B, C ve D olmak üzere dört farklı kitapçık kullanılmaktadır. Bu kitapçıklardan A ile C ve B ile D kapsam ve güçlük açısından birbirlerine paralel olacak şekilde hazırlanmıştır. EARGED (2010), birbirine paralel

olarak hazırlanan kitapçıklardan elde edilen puanları eşdeğer kabul edip testlerin eşitlenmesine gerek duymamıştır. Ancak, her ne kadar testler birbirine paralel olarak hazırlanmaya çalışılsa da en azından güçlük düzeyi olarak farklılık gösterecektir ve farklı test formlarını alan öğrencilerin puanlarını karşılaştırabilmek için test puanlarının eşitlenmesi gerekir. Bu araştırmanın amacı, ÖBBS 9. sınıf A ve C kitapçıklarında yer alan sosyal bilimler testi puanlarını eşit yüzdelikli ve MTK eşitleme yöntemleri ile eşitleyerek farklı eşitleme yöntemlerini test eşitlemenin eşitlik özelliği ölçütüne göre karşılaştırmaktır.

YÖNTEM

Çalışma Grubu

Bu araştırmanın çalışma grubunu, 2009 Öğrenci Başarılarının Belirlenmesi Sınavı (ÖBBS)'nda A kitapçığını (A formu) alan 15.173 ve C kitapçığını (C formu) alan 14.365 9. sınıf ortaöğretim öğrencisi oluşturmaktadır. A ve C kitapçıklarını alan öğrencilerin sosyal bilimler testleri puan dağılımlarına ilişkin betimsel istatistikler Tablo 1'de görülmektedir. Tablo 1'de verilen sosyal bilimler testi puan ortalamalarına dayalı olarak, A formunun ($\hat{\mu}_A = 7.737$) C formundan ($\hat{\mu}_C = 8.091$) daha zor olduğu söylenebilir. A formuna ait çarpıklık katsayısının pozitif, C formuna ait çarpıklık katsayısının negatif olması da bu bulguyu destekler niteliktedir. Basıklık katsayılarının negatif olmasına dayalı olarak ise her iki forma ait puan dağılımlarının normal dağılıma göre daha basık olduğu söylenebilir.

Tablo 1. *A ve C kitapçıklarındaki sosyal bilimler testi puanlarına ilişkin betimsel istatistikler*

	N	K	$\hat{\mu}$	$\hat{\sigma}$	$\widehat{sk}$	$\widehat{ku}$
A Formu	15173	15	7.737	3.481	0.169	-0.875
C Formu	14365	15	8.091	3.405	-0.061	-0.867

Not: N: Kişi sayısı; K: Madde sayısı; $\hat{\mu}$: ortalama; $\hat{\sigma}$: standart sapma; $\widehat{sk}$: çarpıklık katsayısı; $\widehat{ku}$: basıklık katsayısı.

Araştırma Verileri ve Eşitleme Deseni

Araştırmada, 2009 ÖBBS 9. sınıflarda uygulanan A ve C kitapçıklarında yer alan sosyal bilimler testine öğrencilerin verdiği cevaplar kullanılmıştır. Bu araştırma için sosyal bilimler testlerinin seçilme nedeni, testlerin eşitleme yapabilmek için gerekli olan tek boyutluluk varsayımını

sağlamış olmasıdır. Sosyal bilimler testi 9. Sınıf düzeyinde toplam 15 çoktan seçmeli maddeden oluşmaktadır. Kitapçıklar ilk öğrenci A, ikinci B, üçüncü C, dördüncü D… kitapçığını alacak şekilde dağıtılmıştır. Uygulanan bu sarmal sürece dayalı olarak A ve C kitapçıklarını alan grupların tesadüfî olarak eşdeğer olduğu kabul edilmiş ve bu araştırmada testler eşdeğer gruplar deseni kullanılarak birbirine eşitlenmiştir.

Verilerin Analizi

Bu araştırmada A kitapçığında yer alan sosyal bilimler testi (A formu) temel alınmış ve C kitapçığında yer alan sosyal bilimler testi (C formu) A formuna eşitlenmiştir. Testler, üç eşit yüzdelikli eşitleme (düzgünleştirilmemiş (unsmoothed), ön-düzgünleştirilmiş (presmoothed) ve son-düzgünleştirilmiş (postsmoothed)) ve iki MTK'ya dayalı test eşitleme (MTK gerçek puan ve gözlenen puan eşitleme) yöntemlerini kullanılarak birbirine eşitlenmiştir.

Eşit yüzdelikli eşitleme RAGE-REQUATE (Cui ve Kolen, 2005) bilgisayar programı kullanılarak yürütülmüştür. Uygulamada en çok kullanılan ön-düzgünleştirme yöntemlerinden biri log-lineer düzgünleştirme yöntemidir. Bu yöntemde ne kadar düzgünleştirme yapılacağını kontrol eden polinomun derecesinin (C) seçimi oldukça önemlidir. C parametresi genellikle 1'den 12'ye kadar değerler alır ve düzgünleştirilmiş modellerin gözlenen modele uyumlu olup olmadığı incelenerek seçilir (Liu, 2011). Kolen ve Brennan (2004), C parametresinin ki-kare (X^2) uyum iyiliği istatistiğine dayalı olarak seçilebileceğini belirtmiştir. Dolayısıyla, bu araştırmada ön-düzgünleştirme için gerekli olan C parametre değerine karar verirken X^2 uyum indeksi değerleri incelenmiştir. Tablo 2 ve Tablo 3'te A ve C formlarına ait log-lineer ön-düzgünleştirme yapıldıktan sonra elde edilen moment ve uyum indeks değerleri görülmektedir.

Tablo 2 ve Tablo 3'teki bilgiler incelendiğinde, her iki forma ait X^2 uyum indeksi değerlerinin $C \geq 4$ olduğu koşullarda istatistiksel olarak anlamlı olmadığı bir diğer deyişle model-veri uyumunun sağlandığı görülmektedir. Ki-kare fark testi, A formu için C=5 ve C=4 arasındaki ki-kare farkının istatistiksel olarak anlamlı olmadığını göstermektedir ($X^2_{4\text{-}5}(1) = 0.701$, $p>0.05$). C formu için ise C=5 ve C=4 arasındaki ki-kare farkı istatistiksel olarak anlamlı iken C=5 ve C=6 arasındaki ki-kare farkı anlamlı değildir ($X^2_{5\text{-}6}(1)=0.72$, $p>0.05$). Ön-düzgünleştirme, her iki forma aynı anda uygulanacağından dolayı ortak bir C parametresinin seçilmesi gerekmektedir. Dolayısıyla her iki form için model veri uyumunun sağlandığı ve ki-kare fark istatistiğinin istatistiksel olarak anlamlı olmadığı C=5 değeri ön-düzgünleştirmenin derecesi olarak seçilmiştir.

Tablo 2. *A formuna ait log-lineer ön-düzgünleştirmeye ilişkin moment ve uyum istatistikleri*

Derece	$\hat{\mu}$	$\hat{\sigma}$	$\widehat{sk}$	$\widehat{ku}$	X^2	$X^2(df)$	$X_c^2 - X_{c+1}^2$
C=12	7.737231	3.48057	0.169021	2.125149	1.318	3	
C=11	7.737231	3.48057	0.169021	2.125149	1.357	4	0.039
C=10	7.737231	3.48057	0.169021	2.125149	2.978	5	1.621
C= 9	7.737231	3.48057	0.169021	2.125149	2.979	6	0.001
C= 8	7.737231	3.48057	0.169021	2.125149	7.249	7	4.27
C= 7	7.737231	3.48057	0.169021	2.125149	7.373	8	0.124
C= 6	7.737231	3.48057	0.169021	2.125149	8.426	9	1.053
C= 5	7.737231	3.48057	0.169021	2.125149	8.793	10	0.367
C= 4	7.737231	3.48057	0.169021	2.125149	9.494	11	0.701
*C= 3	7.737231	3.48057	0.169021	2.369793	426.775	12	417.281
*C= 2	7.737231	3.48057	-0.04217	2.383521	785.742	13	358.967
*C= 1	7.737231	4.60608	-0.06227	1.795907	4855.521	14	4069.779

Not.*p<0.05; $\hat{\mu}$= ortalama; $\hat{\sigma}$= standart sapma; $\widehat{sk}$=çarpıklık katsayısı; $\widehat{ku}$=basıklık katsayısı; $X^2_{tablo}(11)=19.68$; $X^2_{tablo}(1)=3.84$

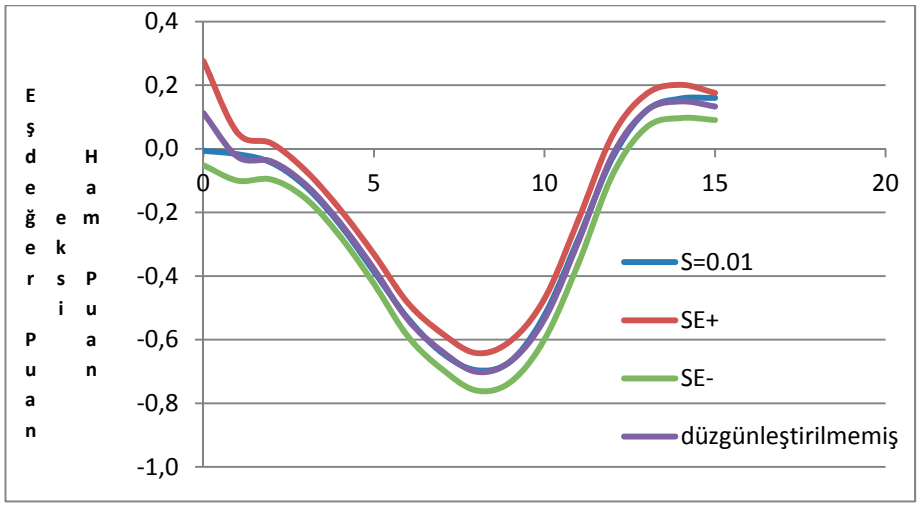
Tablo 3. *C formuna ait log-lineer ön-düzgünleştirmeye ilişkin uyum istatistikleri*

Derece	$\hat{\mu}$	$\hat{\sigma}$	$\widehat{sk}$	$\widehat{ku}$	X^2	$X^2(df)$	$X_c^2 - X_{c+1}^2$
C=12	8.091124	3.404668	-0.06081	2.132875	7.724	3	
C=11	8.091124	3.404668	-0.06081	2.132875	7.725	4	0.001
C=10	8.091124	3.404668	-0.06081	2.132875	9.237	5	1.512
C= 9	8.091124	3.404668	-0.06081	2.132875	9.512	6	0.275
C= 8	8.091124	3.404668	-0.06081	2.132875	12.589	7	3.077
C= 7	8.091124	3.404668	-0.06081	2.132875	12.789	8	0.2
C= 6	8.091124	3.404668	-0.06081	2.132875	13.916	9	1.127
C= 5	8.091124	3.404668	-0.06081	2.132875	14.636	10	0.72
C= 4	8.091124	3.404668	-0.06081	2.132875	18.986	11	4.35
*C= 3	8.091124	3.404668	-0.06081	2.414195	460.582	12	441.596
*C= 2	8.091124	3.404668	-0.10092	2.431532	471.869	13	11.286
*C= 1	8.091124	4.586825	-0.15548	1.823776	4701.998	14	4230.129

Not.*p<0.05; $\hat{\mu}$: ortalama; $\hat{\sigma}$: standart sapma; $\widehat{sk}$: çarpıklık katsayısı; $\widehat{ku}$: basıklık katsayısı; $X^2_{tablo}(11)=19.68$; $X^2_{tablo}(1)=3.84$

Son-düzgünleştirme, literatürde en çok kullanılan yöntemlerden biri olan kübik-spline (cubic spline) yöntemi kullanılarak gerçekleştirilmiştir. Bu yöntemde düzgünleştirmenin derecesi "s" parametresi tarafından kontrol edilir ve bu parametrenin seçiminde kullanılan herhangi bir istatistiksel test bulunmamaktadır. Ancak düzgünleştirme yapıldıktan sonra grafikler ve

betimsel istatistikler incelenerek düzgünleştirmenin derecesine karar verilebilir (Kolen ve Brennan, 2004). Bu araştırmada, son-düzgünleştirme yöntemi için gerekli olan "s" değerine karar vermek için son-düzgünleştirme yapıldıktan sonra elde edilen eşdeğer puanlara ilişkin çizilen grafikler incelenmiştir. Başlangıçta son-düzgünleştirme dört düzgünleştirme parametresi (s=0.01, 0.05, 0.10, 0.20) kullanarak gerçekleştirilmiştir. Şekil-1'de s=0.01 değerinde son düzgünleştirme yapıldıktan sonra elde edilen eşdeğer puanlara ilişkin grafik görülmektedir.



Şekil 1. *Son-düzgünleştirilmiş (s=0.01) eşit yüzdelikli eşitleme*

Şekil-1'de verilen grafik incelendiğinde s=0.01 değerinde, eşdeğer puanların yeterince düzgünleştiği ve bütün noktaların standart hata bandının içerisinde yer aldığı görülmektedir. Dolayısıyla son-düzgünleştirilmiş eşit yüzdelikli eşitleme yöntemi için s parametre değeri 0.01 olarak seçilmiştir.

MTK'ya dayalı test eşitleme yapabilmek için öncelikle BILOG-MG 3.0 (Zimowski, Muraki, Mislevy ve Bock, 2003) bilgisayar programında A ve C formlarına ait madde parametreleri kestirilmiş. Madde parametreleri 3-parametreli lojistik (3PL) model ile 40 kareleme (quadrature) noktası tanımlanarak kestirilmiştir. Her iki forma ait parametreler kestirildikten sonra PIE (Hanson ve Zeng, 2004) bilgisayar program aracılığı ile MTK gerçek puan ve gözlenen puan eşitleme yöntemleri ile testler birbirine eşitlenmiştir. Eşitlik özelliği ölçütüne göre eşitleme sonuçlarını değerlendirmek üzere koşullu ortalamalar ve koşullu SEM'ler 3PL model varsayımı altında POLYCSEM (Kolen, 2004) bilgisayar program

kullanılarak hesaplanmıştır. Son olarak, BSE özelliğini değerlendirmek için D_1, İSE özelliğini değerlendirmek için D_2 indeksi hesaplanmıştır.

BULGULAR

Beş farklı eşitleme yöntemine dayalı olarak elde edilen D_1 ve D_2 değerleri Tablo 4'te sunulmuştur. Tablo 4'te verilen bilgilere göre en küçük D_1 değerine (0.038) MTK gerçek puan eşitleme yöntemi sahip olurken en büyük D_1 değerine (0.073) son-düzgünleştirilmiş eşit yüzdelikli eşitleme (EYE) yöntemi sahip olmuştur. Bu bulguya dayalı olarak, BSE özelliğini en iyi MTK gerçek puan eşitleme yöntemi kullanılarak elde edilen eşitleme sonuçlarının koruduğu, en kötü ise son-düzgünleştirilmiş EYE yöntemi ile elde edilen eşitleme sonuçlarının koruduğu söylenebilir. Düzgünleştirilmemiş ve ön-düzgünleştirilmiş EYE yöntemleri aynı ve ikinci büyük D_1 değerine sahip olarak BSE özelliğini korumada benzer performans göstermişlerdir. MTK gözlenen puan eşitleme yöntemi BSE özelliğini korumada her ne kadar MTK gerçek puan eşitleme yönteminden daha kötü performans göstermiş olsa da eşit yüzdelikli eşitleme yöntemlerinden daha iyi performans göstermiştir.

Tablo 4. *Farklı eşitleme yöntemlerine dayalı olarak elde edilen D_1 ve D_2 değerleri*

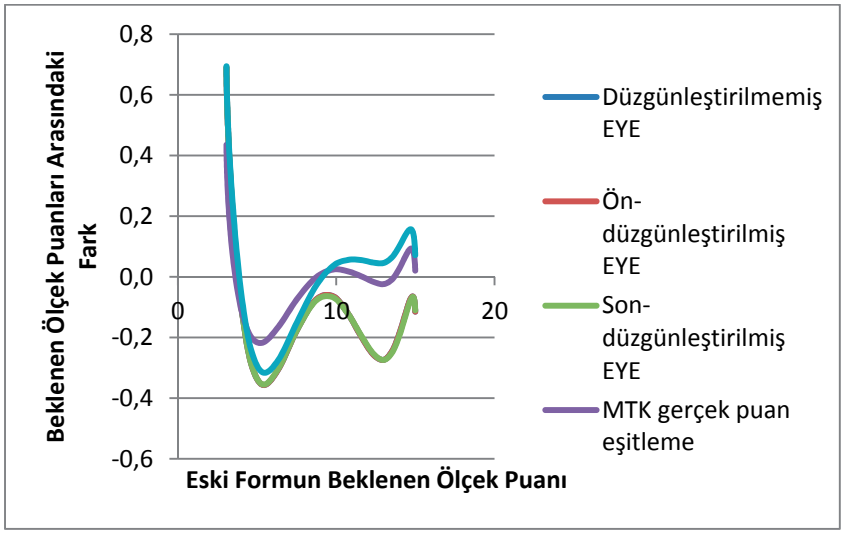
Eşitleme Yöntemleri	D_1	D_2
Düzgünleştirilmemiş EYE	0.072	0.050
Ön-düzgünleştirilmiş EYE (C=5)	0.072	0.049
Son-düzgünleştirilmiş EYE (S=0.01)	0.073	0.049
MTK Gerçek Puan Eşitleme	0.038	0.066
MTK Gözlenen Puan Eşitleme	0.063	0.041

Not: EYE: Eşit yüzdelikli eşitleme, MTK: Madde Tepki Kuramı

Tablo 4'te yer alan bilgilere göre en küçük D_2 değerine (0.041) MTK gözlenen puan eşitleme yöntemi sahip olurken en büyük değere (0.066) MTK gözlenen puan eşitleme yöntemi sahip olmuştur. Bu bulguya dayalı olarak, İSE özelliğini en iyi MTK gözlenen puan eşitleme yöntemi ile elde edilen eşitleme sonuçlarının koruduğu, en kötü ise MTK gerçek puan eşitleme yöntemi ile elde edilen eşitleme sonuçlarının koruduğu söylenebilir. Ön-düzgünleştirilmiş ve son-düzgünleştirilmiş EYE yöntemleri aynı D_2 değerine (0.049) sahip olan yöntemler olmuş, İSE özelliğini korumada da benzer performans göstermişlerdir. Düzgünleştirilmemiş EYE yöntemi ise ikinci büyük D_2 değerine sahip olmuş ve İSE özelliğini korumada

düzgünleştirilmenin uygulandığı EYE yöntemlerinden daha kötü performans göstermiştir. Genel olarak, eşit yüzdelikli eşitleme yöntemlerinin İSE özelliğini korumada MTK gerçek puan eşitlemeden daha iyi performans gösterdiği söylenebilir.

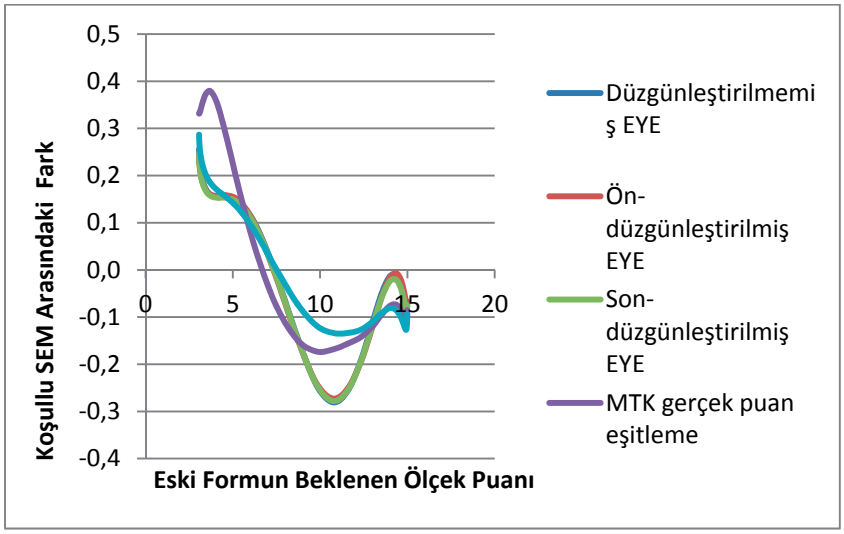
Test eşitleme yöntemlerinin puan dağılımı boyunca BSE özelliğini koruma performansını incelemek için eski formun (A formunun) beklenen ölçek puanlarına karşılık gelen beklenen ölçek puanları arasındaki farka ilişkin grafiği çizilerek Şekil-2'de verilmiştir.



Şekil 2. *Beklenen puanlar arasındaki farka ilişkin grafik.*

Şekil-2'de verilen grafiğe göre puan ölçeği boyunca MTK gerçek puan eşitleme yöntemi ile elde edilen beklenen ölçek puanları arasındaki farkın diğer yöntemlere göre daha düşük olduğu söylenebilir. MTK gözlenen puan eşitleme yöntemi her ne kadar puan dağılımının başlangıcında üç eşit yüzdelikli eşitleme yöntemine yakın fark puanlarına sahip olmuş olsa da puan dağılımının geriye kalan kısmında eşit yüzdelikli eşitleme yöntemlerinden daha düşük fark puanlarına sahip olmuştur. Ayrıca, grafiğe dayalı olarak üç eşit yüzdelikli eşitleme yöntemine ait fark puanlarının puan ölçeği boyunca birbirine çok yakın değerler aldığı söylenebilir.

Test eşitleme yöntemlerinin puan ölçeği boyunca İSE özelliğini korumadaki performansını karşılaştırmak için eski formun beklenen ölçek puanlarına karşılık gelen ölçmenin standart hataları arasındaki farkın grafiği çizilerek Şekil-3'te verilmiştir.



Şekil 3. *Ölçmenin koşullu standart hataları (SEM) arasındaki farka ilişkin grafik*

Şekil-3'te verilen grafiğe dayalı olarak, puan dağılımının başlangıç ve orta noktalarında MTK gözlenen puan eşitleme yönteminin üç eşit yüzdelikli eşitleme yöntemine yakın SEM fark puanlarına sahip olduğu fakat puan dağılımının sonlarına doğru MTK gözlenen puan eşitlemenin daha düşük fark puanlarına sahip olduğu söylenebilir. Puan dağılımının başlangıcında MTK gerçek puan eşitleme yönteminin en yüksek SEM fark puanlarına sahip olduğu grafikten elde edilen bir diğer bilgidir. Benzer şekilde puan dağılımı boyunca üç eşit yüzdelikli eşitlemenin birbirine yakın SEM fark puanlarına sahip olmuştur.

TARTIŞMA ve SONUÇ

Bu araştırma kapsamında üçü eşit yüzdelikli eşitleme, ikisi MTK'ya dayalı eşitleme yöntemi olmak üzere beş farklı test eşitleme yöntemi BSE ve İSE özelliğini korumadaki performanslarına göre karşılaştırılmıştır. Elde edilen bulgular BSE özelliğini en iyi koruyan yöntemin MTK gerçek puan eşitleme yöntemi, İSE özelliğini en iyi koruyan yöntemin MTK gözlenen puan eşitleme yöntemi olduğunu göstermiştir. Eşit yüzdelikli eşitleme yöntemleri her iki eşitlik özelliğini korumada birbirine yakın performans göstermiştir.

Elde edilen bulgular literatürde yer alan diğer araştırmaların bulgularını destekler niteliktedir. Tong ve Kolen (2005) ön-düzgünleştirilmiş EYE, MTK gerçek puan eşitleme ve MTK gözlenen puan eşitleme yöntemlerini

BSE ve İSE özelliği ölçütlerine göre karşılaştırdıkları araştırmalarında eğer amaç BSE özelliğini korumak ise MTK gerçek puan eşitleme yönteminin, İSE özelliğini korumak ise MTK gözlenen puan eşitleme yönteminin tercih edilmesini önermişlerdir. Lee ve diğerleri (2010), Tong ve Kolen (2005)'in araştırmasının uzantısı olarak yürüttükleri araştırmalarında yedi farklı eşitleme yöntemini (MTK gerçek puan eşitleme, MTK gözlenen puan eşitleme, düzgünleştirilmemiş EYE, ön-düzgünleştirilmiş EYE, son-düzgünleştirilmiş EYE, kernel eşitleme ve süreklendirilmiş log-lineer eşitleme) BSE ve İSE özelliğini koruma performanslarına göre karşılaştırmıştır. Araştırmalarının sonucunda MTK gerçek puan eşitlemenin BSE özelliğini en iyi koruyan yöntem olduğunu, MTK gözlenen puan eşitlemenin BSE özelliğini geleneksel eşitleme yöntemlerinden daha iyi koruduğu fakat MTK gerçek puan eşitlemeden daha kötü koruduğu sonucuna ulaşmışlardır. İSE özelliği ölçüt olarak alındığında ise en iyi performans gösteren yöntemin MTK gözlenen puan eşitleme olduğunu görmüşlerdir.

Lee ve Brennan (2012), eşitlik özelliği ölçütünün değerlendirmede kullanılan psikometrik modelden etkilendiğini belirtmiş ve eğer MTK değerlendirmede varsayılmış ise MTK gerçek puan eşitleme yöntemindeki gerçek puanlar MTK modellerine dayalı olduğu için bu yöntemin diğer yöntemlere nazaran BSE özelliğini korumada daha avantajlı olabileceğini öne sürmüştür. Ayrıca Kim, Brennan ve Kolen (2005) eğer değerlendirmede eşitlemede kullanılan ile aynı psikometrik model kullanılır ise MTK gerçek puan eşitlemenin BSE özelliğini MTK gözlenen puandan daha iyi koruyacağını göstermiştir. Bu araştırmada hem testler eşitlenirken hem de eşitlik özelliği ölçütüne göre değerlendirme yapılırken MTK'ya dayalı modelin kullanılması MTK gerçek puan eşitlemenin diğer yöntemlere göre BSE özelliğini daha iyi korumasının bir nedeni olabilir.

Bolt (1999) eşitlik özelliği ölçütünün yürütülen bir test eşitleme çalışmasının adaletli olup olmadığını testi alan tüm grubun yanında birey düzeyinde de incelemeye fırsat verdiği gerekçesi ile savunmuştur. Eğer amaç birey düzeyinde bir test formunun uygulanmasının avantaj ya da dezavantajını önlemek ise BSE özelliğinin korunumu önemli hale gelir ve böyle bir durumda BSE özelliğini en iyi koruyan MTK gerçek puan eşitleme yöntemi tercih edilebilir. Ancak, eğer öncelikli amaç ölçmenin kesinliğini arttırmak ve ölçme hatalarını azaltmak ise İSE özelliğinin korunumu öne çıkar ve böyle bir durumda MTK gözlenen puan eşitleme yöntemi tercih edilebilir (He, 2011). Bu araştırmanın bulguları sadece ÖBBS 9. sınıf sosyal bilimler testleri ile sınırlıdır. Genelleme yapabilmek için eşitleme yöntemlerinin farklı testler ve farklı değerlendirme ölçütleri kullanılarak karşılaştırılmasına ihtiyaç vardır.

KAYNAKLAR

Andrews, B. J. (2011). Assessing first-and second-order equity for the common item nonequivalent groups design using multidimensional IRT. Unpublished doctoral dissertation, University of Iowa.

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

Cui, Z. & Kolen, M. J. (2005). RAGE-RGEQUATE [Computer Program]. Iowa City, IA: The University of Iowa, Iowa Testing Programs.

Eğitim, Araştırma ve Geliştirme Daire Başkanlığı. (2010). Ortaöğretim ÖBBS raporu 2009.

http://egitek.meb.gov.tr/dosyalar/obbs/OBBS_2009.pdf adresinden 31 Temmuz 2013 tarihinde indirilmiştir.

Hanson, B. & Zeng, L. (2004). PIE: A computer program for IRT equating. (Windows Console Version,

Revised by Cui, May 20, 2004) [Manual]. Unpublished manuscript, College of Education, University of Iowa, Iowa City, Iowa

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, Iowa City.

Kim, D. I., Brennan, R. L. & Kolen, M. J. (2005). A comparison of IRT equating and beta 4 equating. Journal of Educational Measurement, 42(1), 77-99.

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

Kolen, M. J. (1988). Effectiveness of analytic smoothing in equipercentile equating. Journal of Educational Statistics, 9 (1), 25-44.

Kolen, M. J. (2004). POLYCSEM [Computer software]. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa.

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

Lee, E., Lee, W. & Brennan, R. L. (2010). Assessing equating results based on first-order and second- order equity (CASMA Research Report No. 31). Iowa City, IA: Center for advanced Studies in Measurement and Assessment, University of Iowa.

Liu, C. (2011). A comparison of statistics for selecting smoothing paramaters for loglinear presmoothing and cubic spline post smoothing under a random groups design (Unpublished doctoral dissertation). Available from Iowa Research Online. (UMI No. 1013)

Livingston, S. A. (2004). Equating test scores (Without IRT). Educational Testing Service.

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

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.

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

Zimowski, M. F., Muraki, E., Mislevy, R. J. & Bock, R. D. (2003). BILOGMG 3.0 for Windows: Multiple group IRT analysis and test maintenance for binary items [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.