PISA 2012 MATEMATİK OKURYAZARLIĞI TESTİNDE FARKLI ÖLÇEK DÖNÜŞTÜRME YÖNTEMLERİNİN KARŞILAŞTIRILMASI[1]

COMPARISON OF DIFFERENT SCALE LINKING METHODS IN PISA 2012 MATHEMATICS LITERACY TEST

Şeyma UYAR[2] Burcu AKSEKİOĞLU[3] Neşe ÖZTÜRK GÜBEŞ[4]

Başvuru Tarihi: 24.07.2017 Yayına Kabul Tarihi: 12.03.2018 DOI: 10.21764/ maeuefd.330613

Özet: Bu çalışmada farklı ölçek dönüştürme yöntemlerini PISA 2012 matematik okuryazarlığı verileri üzerinde karşılaştırmak amaçlanmıştır. Bu amaçla seçilen iki kitapçıktan elde edilen puanlar madde tepki kuramına dayalı ölçek dönüştürme (ortalama-ortalama, ortalama-standart sapma, Stocking-Lord, Haebara) ve test eşitleme yöntemleri (MTK gerçek-puan eşitleme, MTK gözlenen-puan eşitleme) kullanılarak eşitlenmiş ve farklı yöntemlerden elde edilen sonuçlar incelenmiştir. Çalışma, 4 ve 11 numaralı kitapçıklardaki matematik testlerine verilen cevaplar kullanılarak yürütülmüştür. Bu nedenle araştırmanın çalışma grubunu Türkiye örnekleminde 4 numaralı kitapçığı cevaplayan 348 ve 11 numaralı kitapçığı cevaplayan 368 olmak üzere toplam 716 öğrenci oluşturmaktadır. Çalışmada test eşitleme için "denk olmayan gruplarda ortak madde deseni" kullanılmıştır. Verilerin analizinin ilk aşamasında madde tepki kuramının tek boyutluluk varsayımı test edilmiştir. Ardından PARSCALE 4.1 programı ile madde ve yetenek parametreleri kestirilmiştir. Parametre kestiriminde iki-parametreli lojistik model ve genelleştirilmiş kısmi kredi modeli kullanılmıştır. Daha sonra STUIRT programı ile dört farklı yöntem kullanılarak ölçek dönüştürme işlemi yapılmıştır. Son aşamada ise her iki formdan elde edilen test puanları POLYEQUATE programı ile eşitlenmiştir. Farklı yöntemlerden elde edilen hata miktarları ise ağırlıklandırılmış hata kareleri ortalaması (WMSE) ile hesaplanmıştır. Çalışma sonucunda, en az hata miktarına sahip yöntemin gerçek-puan eşitlemede Stocking-Lord, gözlenen-puan eşitlemede ise Haebara yönteminin olduğu bulunmuştur. En yüksek eşitleme hatasını ise ortalama-standart sapma yönteminin verdiği tespit edilmiştir.

Anahtar Sözcükler: *Test eşitleme, karma test, ölçek dönüştürme yöntemleri, eşitleme hatası*

Abstract: In this study, the objective was to compare different scale linking methods over the PISA 2012 mathematics literacy data. For this purpose, scores obtained from two selected booklets were equated using scale linking (mean-mean, mean-sigma, Stocking-Lord, Haebara) and test equating methods (IRT true-score equating, IRT observed-score equating) based on the item response theory, and results obtained from different methods were analyzed. The study was conducted using answers given to mathematics tests in booklet-4 and booklet-11. Therefore, the sample consists of 716 students in Turkey; 348 of these participants are the takers of booklet-4, 368 of them are the takers of booklet-11. In order to equate test forms, "the common-item nonequivalent groups" design was used in this research. In the first stage of data analysis, unidimensionality assumption of the item response theory was analyzed. Then PARSCALE 4.1 was used to estimate item and ability parameters. Generalized partial credit and two-parameter logistic model were used to estimate parameters. Afterwards, STUIRT program was used for scale linking for four different methods. In the last step test scores obtained from different forms were equated by using POLYEQUATE program. Equating error obtained from different methods calculated with weighted mean squares error (WMSE) index. Results showed that Stocking-Lord method had the smallest equating error in true-score equating and Haebara method had the smallest equating error in observed-score equating. The amount of maximum error has been established that of the mean-sigma method.

Keywords: *Test equating, mixed test, scale linking methods, equating error*

[1]Bu çalışma 1-3 Eylül 2016 tarihleri arasında Akdeniz Üniversitesi'nde düzenlenen V. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi'nde sözlü bildiri olarak sunulmuştur.

[2] Dr. Öğr. Üyesi, Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, syuksel@mehmetakif.edu.tr, Orcid No: 0000-0002-8315-2637

[3] Arş. Gör., Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, baksekioglu@mehmetakif.edu.tr, Orcid No: 0000-0001-5635-0753

[4] Dr. Öğr. Üyesi, Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, nozturk@mehmetakif.edu.tr, Orcid No: 0000-0003-0179-1986

Giriş

Ölçme alanında tartışılan önemli sorunlardan biri, iki farklı ölçme aracı ya da yöntemiyle değerlendirme yapıldığında, sonuçların denk olduğuna dair kanıt aramaktır. Bir başka deyişle aynı özellik iki farklı testle ölçülmek istenilirse, geçerlik ve güvenirlikleri yüksek de olsa buradaki en büyük sorun, bir testten elde edilen puanın diğeriyle karşılaştırılabilir olup olmadığıdır (Skaggs & Lissitz, 1982). Test geliştiriciler içerik ve istatistiksel özellikler bakımından birbirine paralel formlar oluşturmaya çalışsa da, test formları güçlük düzeyi bakımından bir miktar farklılık gösterebilmektedir (Kolen & Brennan, 1995; Tanguma, 2000). Bu nedenle, başarıyı ölçmek için birden fazla formun kullanıldığı testlerde bir grup avantajlı iken, diğer grup dezavantajlı duruma düşebilmektedir (Felan, 2002). Birinci testten 20 puan alan bir kişinin puanının, ikinci testten 20 puan alan birinin puanına eşit olduğunu söyleyebilmek için test formlarının aynı güçlük düzeyinde olduğunu kanıtlamak gerekmektedir (Skaggs & Lissitz, 1982; Tanguma, 2000; Zhu, 1998). Bu problemi ortadan kaldırmak için farklı formlardan elde edilen puanların ortak bir ölçeğe yerleştirilerek test eşitleme çalışmasının yapılması gerekmektedir (Cook & Eignor, 1991; Skaggs & Lissitz, 1982; Tanguma, 2000).

Test eşitleme, test formlarından alınan puanları eşitlemede kullanılan istatistiksel bir süreçtir. Eşitleme işleminde bir formun birim sistemi diğer bir formun birim sistemine dönüştürülür. Böylece iki farklı formdan alınan puanlar doğrudan eşit kabul edilmekte ve farklı formları alan iki kişinin performansı doğrudan karşılaştırılabilmektedir (Angoff, 1984; Crocker & Algina, 1986; Tsai, Hanson, Kolen & Forsyth, 2001). Ayrıca eşitleme çalışmaları, test sürecinin daha adil olmasını sağlayarak karar verme süreçlerinde de daha faydalı olabilmektedir (Felan, 2002; Kolen & Brennan, 1995).

Test eşitlemede veri toplamak için çeşitli desenler kullanılmaktadır. Eşitleme desenleri, bireylerin tabi tutulduğu test formu sayısı ile karşılaştırma yapılan grup sayısına göre farklılık göstermektedir (Holland, Dorans & Petersen, 2007). Literatürde yaygın olarak kullanılan desenler tek grup deseni, dengelenmiş tek grup deseni, eşdeğer gruplar deseni ve denk olmayan gruplarda ortak madde desenidir. Denk olmayan gruplarda ortak madde deseni, alanyazında en yaygın olarak kullanılan desendir (Kolen & Brennan, 2004). Bu desende grupların eşdeğer olması beklenmez ve gruplar, ortak maddelerin yer aldığı iki farklı test formundan birini yanıtlar. Böylece iki grubun testin tamamına verdiği yanıtlarla ortak maddelere verdiği yanıtlar arasındaki farklılıklar kontrol edilebilmektedir (Kolen, 1988). Desenin en önemli avantajı uygulamasının esnek olmasıdır (Petersen, Kolen & Hoover 1989; Sinharay & Holland, 2010). Uygulamada eşitleme çalışması

yapılacak iki grup denk olmak zorunda değildir. Bu nedenle grupların ortalama ve standart sapmalarının da farklılaşmasının bir önemi yoktur (Embretson & Reise, 2000).

Test eşitleme yöntemleri dayandığı kuramsal temele göre klasik test kuramına (KTK) ve madde tepki kuramına (MTK) dayalı yöntemler olmak üzere iki başlık altında incelenir. KTK'ya dayalı yöntemlerde eşitleme sonuçları genel olarak gruba bağımlıdır ve bazı eşitleme varsayımlarını karşılamak güçtür. Klasik test kuramına dayalı eşitlenmenin bu gibi zayıf noktalarından dolayı madde tepki kuramına dayalı test eşitleme önem kazanmıştır (Hambleton, 1989). MTK'da madde parametreleri, testin uygulandığı gruptan bağımsız olarak elde edilebilmektedir. Bireylerin yeteneği de, maddelere verdikleri cevaplardan bağımsız olarak kestirilebilmektedir. Bireyin, farklı güçlük düzeyine sahip test formlarından herhangi birini alması, yetenek kestirimini değiştirmemektedir (Hambleton & Swaminathan, 1985). MTK ve KTK ile eşitleme yöntemlerini karşılaştıran çalışmalar incelendiğinde, genellikle MTK eşitleme yöntemlerinin daha kararlı sonuçlar verdiği görülmektedir (Han, Kolen & Pohlman, 1997; Yang & Houang, 1996). Bu nedenle MTK test eşitlemede yaygın olarak kullanılmaktadır (Kim & Lee, 2006).

MTK ile eşitleme sürecinde öncelikle veriye uyum gösteren model belirlenerek her iki form için madde ve yetenek parametreleri kestirilir. Kestirilen parametreler farklı ölçeklerde yer aldığı için doğrudan karşılaştırılamaz. Bu nedenle çeşitli kalibrasyon yöntemleri kullanılarak elde edilen parametreler ortak bir ölçeğe yerleştirilir. Bunlar, parametrelerin tek bir analizle kestirildiği "eşzamanlı kalibrasyon" ve parametrelerin ayrı ayrı kestirildiği "ayrı kalibrasyon" yöntemleridir. Eş zamanlı kalibrasyonda her iki formdaki parametreler tek bir analizle aynı anda kestirilir (Hanson & Beguin, 2002). Bu nedenle bu yöntemde doğrusal bir ölçek dönüştürme işlemine gerek kalmaz (Kim & Kolen, 2006; Tsai ve diğ., 2001). Ayrı kestirimde ise kalibrasyon işlemi iki aşamada gerçekleştirilir. Öncelikle iki farklı test formu için madde ve yetenek parametreleri ayrı ayrı kestirilir. Ardından kestirilen parametreler için lineer bir dönüşüm işlemi gerçekleştirilir (Cook & Eignor, 1991). Bu dönüşüm kestirilen parametreleri ortak bir ölçeğe yerleştirmeyi sağlayacak doğrusal dönüşüm denklemi ile hesaplanır (Hanson & Beguin, 2002; Kim & Cohen, 2002). Yetenek parametresi için dönüşüm denklemi aşağıda verilmiştir (Cook & Eignor, 1991).

$$\theta^{*} = A\theta + B$$

θ : Kişinin yetenek düzeyi

θ^{*} : Kişinin girmediği sınavdan alacağı kuramsal yetenek düzeyi

A : Eşitleme denkleminin eğimi

B : Eşitleme denkleminin sabiti

Doğrusal dönüşüm denklemi aynı şekilde ayırıcılık (α) ve güçlük (b) parametreleri için de şu şekilde hesaplanır:

$$\alpha_i^* = \frac{1}{A}\alpha_i$$

$$b_i^* = Ab_i + B$$

$$c_i^* = c_i$$

α_i : Ayırıcılık parametresi

α_i^* : Dönüştürülmüş ayırıcılık parametresi

b_i : Güçlük parametresi

b_i^* : Dönüştürülmüş güçlük parametresi

c_i : Şans parametresi

c_i^* : Dönüştürülmüş şans parametresi

A : Eşitleme denkleminin eğimi

B : Eşitleme denkleminin sabiti

Ayrı kalibrasyon yöntemlerinin amacı, parametreleri ortak bir ölçeğe yerleştirmek için gerekli olan eşitleme katsayılarını (A ve B) elde etmektir (Embretson & Reise, 2000). Literatürde bu yöntemler moment yöntemleri ve karakteristik eğri dönüştürme yöntemleri olarak üzere iki kategoride ele alınır. Moment yöntemleri ortalama-ortalama (mean-mean) ve ortalama-standart sapma (mean-sigma) yöntemleridir. Ortalama-ortalama yöntemi, A ve B katsayılarını ortak maddelerin ayırıcılık ve güçlük parametrelerinin ortalamasını kullanarak hesaplar (Loyd & Hoover, 1980). Ortalama-standart sapma yöntemi ise A ve B katsayılarını ortak maddelerin güçlük parametrelerinin ortalama ve standart sapmasını kullanarak hesaplar (Marco, 1977). Ayrı kalibrasyon yöntemlerinin bir diğeri de karakteristik eğri dönüştürme yöntemleridir. Stocking-Lord ve Haebara olmak üzere iki başlık altında incelenen bu yöntemlerde ölçek dönüşüm sabitleri kestirilerek madde karakteristik eğrileri arasındaki farklar hesaplanır. Haebara yöntemi madde karakteristik eğrileri arasındaki farkı, her bir ortak maddenin madde karakteristik eğrileri arasındaki farkın karelerinin toplamını alarak hesaplarken (Haebara, 1980), Stocking-Lord yöntemi farkı, her bir ortak maddenin madde karakteristik eğrileri arasındaki farkın toplamının karesini alarak hesaplar.

Moment yöntemleri bir formdan alınan puanları diğer formun birim sistemine dönüştürmedeki kolaylığı ile bilinirken, karakteristik eğri yöntemleri istatistiksel olarak kararlılık ve üstünlüğü ile ön plana çıkmaktadır (Kim & Kolen, 2006). Nitekim birçok çalışmada karakteristik eğri yöntemlerinin moment yöntemlerine göre daha kararlı sonuçlar verdiği belirtilmiştir (Baker & Al-Karni, 1991; Hanson & Beguin, 2002; Kim & Cohen, 2002; Kim & Kolen, 2006; Ogasawara, 2001).

Eşitleme sürecinde elde edilen parametreler için ölçek dönüştürme işlemi gerçekleştirildikten sonra eşitleme işlemi uygulanır. MTK'ya dayalı eşitleme, gerçek-puan ve gözlenen-puan eşitleme yöntemi olmak üzere iki başlıkta incelenir. MTK gerçek-puan eşitleme yönteminde, eski ve yeni formda yetenek düzeyi ile ilişkili gerçek puanların 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 (Han ve diğ., 1997; Kolen & Brennan, 2004; Lord & Wingersky, 1984). Han ve diğ. (1997), gerçek veriler üzerinde gerçekleştirdikleri çalışmada gerçek-puan eşitleme yönteminin gözlenen-puan eşitlemeden daha kararlı sonuçlar verdiğini belirtmişlerdir. Lord ve Wingersky (1984), ortak maddeli test deseninde gerçekleştirdikleri çalışmada gerçek ve gözlenen-puan eşitlemenin birbirine benzer sonuçlar verdiğini tespit etmişlerdir. Kolen (1981) ise en az hata veren yöntemin eşitleme koşullarına göre değiştiğini belirtmiştir.

Bu araştırmada farklı ölçek dönüştürme yöntemleri Uluslararası Öğrenci Değerlendirme Programı (PISA) 2012 verileri üzerinde karşılaştırılmıştır. PISA, Ekonomik İşbirliği ve Kalkınma Teşkilatı (OECD) tarafından düzenlenen, uluslararası karşılaştırmaların yapıldığı ve eğitsel kararların alındığı dünyanın en kapsamlı eğitim araştırmalarından biridir. Üç yılda bir yapılan bu uygulamayla 15 yaş grubu öğrencilerin matematik, okuma ve fen alanlarında okulda öğrendikleri bilgileri günlük hayattaki problem durumlarında ne derece kullanabildikleri ölçülmektedir (MEB, 2013). PISA uygulamasına katılan öğrenciler, uygulama kapsamında yer alan tüm maddeleri cevaplamamaktadır. Her öğrenci, seçkisiz yöntemle belirlenen ve içerisinde farklı maddelerin yer aldığı 13 kitapçıktan birini cevaplamaktadır. Test türlerinin birbirilerine göre avantaj ya da dezavantaj durumu dikkate alındığında uluslararası sınavlarda bazı madde türlerinin birlikte kullanılması gündeme gelmektedir (Gültekin, 2014; Kubiszyn & Borich, 2013). PISA 2012 uygulamasında ağırlıklı alan matematiktir ve matematik testinde bilişsel alanda yöneltilen sorular hem çoktan seçmeli hem de açık uçlu şekilde hazırlanmaktadır. Bu şekilde farklı madde türlerinin bir arada kullanıldığı karma formattaki testlerle öğrenciyi daha üst düzey bilişsel basamakta ölçme imkanı bulunmaktadır (Kim & Lee, 2006). Çoktan seçmeli maddeler ikili puanlanırken, açık uçlu maddelerde çoklu puanlama yapılmaktadır (MEB, 2013; OECD, 2009).

Uluslararası önem taşıyan bu uygulamada kullanılan kitapçıklarda farklı maddelerin yer alması alınan puanların karşılaştırılmasını zorlaştırmaktadır. Bu karşılaştırmanın adil olması açısından öğrencilerin aldıkları puanların birbirine eşitlenmesi gerekmektedir. Ancak test eşitleme için yöntem seçimi kritik öneme sahiptir. Çünkü yanlış yöntem seçimi yüksek eşitleme hatalarının elde edilmesine neden olabilmektedir. Bu nedenle farklı yöntemlerden elde edilen sonuçların

karşılaştırılması gerekmektedir. Alanyazın incelendiğinde ölçek dönüştürme yöntemlerinin genellikle simülatif veriler üzerinde karşılaştırıldığı görülmektedir (Kim & Lee, 2006; Lee & Ban, 2010; Öztürk-Gübeş & Kelecioğlu, 2016; Tate, 2000; Uysal, 2014). Simülasyon çalışmaları, test eşitleme çalışmaları için araştırmacıya çok sayıda faktörü inceleme imkanı vermesi yönüyle değerlidir. Ancak, sonuçlar gerçek veriyle desteklendiğinde daha kullanışlı olmaktadır (Harris & Crouse, 1993). Literatürde ölçek dönüştürme yöntemlerinin gerçek veriler üzerinde karşılaştırıldığı çalışmalar sınırlı sayıdadır. Bu nedenle çalışmanın gelecek dönemlerde karma formatta hazırlanan testlerle gerçekleştirilecek eşitleme çalışmalarına katkı sağlayacağı düşünülmektedir.

Bu çalışmada karma maddelerden oluşan iki kitapçık farklı ölçek dönüştürme yöntemleri ile eşitlenmiş ve en az hata veren yöntem gerçek veriler üzerinde incelenmiştir. Araştırma kapsamında aşağıdaki problemlere yanıt aranmıştır:

PISA 2012 Matematik okuryazarlığı testinde Kitapçık-11 (Form X)'den elde edilen puanların Kitapçık-4 (Form Y)'ten elde edilen puanlara eşitlenmesinde;

1. Ortalama-ortalama yönteminden elde edilen eşitlenmiş puanlar nasıldır?
2. Ortalama-standart sapma yönteminden elde edilen eşitlenmiş puanlar nasıldır?
3. Stocking-Lord yönteminden elde edilen eşitlenmiş puanlar nasıldır?
4. Haebara yönteminden elde edilen eşitlenmiş puanlar nasıldır?
5. En az eşitleme hatasına sahip ölçek dönüştürme ve test eşitleme yöntemi hangisidir?

Yöntem

Araştırma Modeli

Bu çalışmada PISA 2012 Matematik okuryazarlığı testinden seçilen iki kitapçık karakteristik eğri ve moment yöntemlerine dayalı ölçek dönüştürme yöntemleri ile eşitlenmiş ve farklı yöntemlerden elde edilen hata miktarları incelenmiştir. Araştırmada var olan yöntem ve teknikler sınandığı için çalışma, betimsel araştırma niteliği taşımaktadır. Betimsel araştırmalar, hâlihazırda var olan olguların ne olduğunun betimlenip açıklanarak ortaya konulmasını amaçlar. Bu tür araştırmalarda, üzerinde çalışılan doğal ve toplumsal olguları kontrol etme etkinliği yoktur ve araştırmacı bu olgulara müdahale etmez. Olgu, var olan haliyle incelenir (Sönmez & Alacapınar, 2013).

Çalışma Grubu

Araştırmanın çalışma grubunu, Türkiye örnekleminde PISA 2012'nin 4 numaralı kitapçığını cevaplayan 348 ve 11 numaralı kitapçığını cevaplayan 368 olmak üzere toplam 716 öğrenci oluşturmaktadır.

Araştırma Verileri ve Eşitleme Deseni

Araştırma verilerini PISA 2012 Türkiye uygulamasına katılan öğrencilerin Kitapçık-4 ve Kitapçık-11'de yer alan matematik okuryazarlığı testine verdiği cevaplar oluşturmaktadır. Bu kitapçıkların seçilme sebebi, madde tepki kuramı ile test eşitleme yapabilmek için gerekli olan tek boyutluluk varsayımının sağlanmış olmasıdır. Çalışmada test eşitleme için "denk olmayan gruplarda ortak madde deseni" kullanılmıştır. Araştırmaya ilişkin kullanılan eşitleme deseni Tablo 1'de verilmiştir.

Tablo 1
Araştırmaya İlişkin Denk Olmayan Gruplarda Ortak Madde Deseni

Grup	Kitapçık-4	Ortak	Kitapçık-11	Toplam
1	23(ip)	12 (3 kp+9 ip)		35
2		12 (3 kp+9 ip)	23 (ip)	35

Not :kp = kısmi puanlanan madde; ip = ikili puanlanan madde

Tablo 1'e göre her iki kitapçıkta da 12'si ortak olmak üzere toplam 35'er madde yer almaktadır. Ortak maddelerin üçü kısmi (0-1-2) puanlanırken, dokuz madde ikili (0-1) puanlanmaktadır. Ortak maddelerin dışında kitapçıklarda ikili puanlanan 23 ayrı madde yer almaktadır ve bir kitapçıktan alınabilecek en yüksek puan 38'dir.

Verilerin Analizi

Verilerin analizinde öncelikle betimsel istatistikler hesaplanmıştır. Ardından PARSCALE 4.1 (Muraki & Bock, 2003) programı ile madde ve yetenek parametreleri kestirilmiştir. Parametre kestiriminde ikili puanlanan maddeler için iki-parametreli lojistik model (2PLM), çoklu puanlanan maddeler için ise genelleştirilmiş kısmi kredi modeli (GPCM) kullanılmıştır. İki-parametreli lojistik model (2PLM), θ yetenek düzeyindeki bir bireyin bir maddeyi doğru cevaplama olasılığını madde güçlük ve ayırıcılık parametresini kullanarak hesaplar ve bu modelde şans parametresi (c) sıfıra sabitlenir (Hambleton, Swaminathan & Rogers, 1991). Genelleştirilmiş kısmi kredi modeli (GPCM) ise çoklu puanlanan maddelerde parametre kestirimi için geliştirilmiştir. Denklemi Muraki (1997) tarafından oluşturulan model ayırıcılık ve kategori sınır parametrelerini kullanmaktadır (Ostini & Nering, 2006). Parametre kestiriminin ardından STUIRT (Kim & Kolen, 2004) programı ile dört

farklı yöntem (ortalama-ortalama, ortalama-standart sapma, Stocking-Lord, Haebara) için ölçek dönüştürme işlemi gerçekleştirilmiştir. Son aşamada ise her iki kitapçıktan elde edilen puanlar MTK-gerçek puan eşitleme ve MTK gözlenen-puan eşitleme yöntemleri kullanılarak POLYEQUATE (Kolen, 2004) programı ile eşitlenmiş, farklı yöntemlerden elde edilen hata miktarlarını belirlemek için ise WMSE (Ağırlıklandırılmış Hata Kareleri Ortalaması) katsayısı hesaplanmıştır. Elde edilen WMSE değerleri DTM (Difference That Matters) katsayısı ile karşılaştırılmıştır. DTM, Dorans ve Feigenbaum (1994) tarafından önerilen, eşitleme sonuçlarının önemli olup olmadığını değerlendiren bir ölçüttür ve sonuçları yorumlamak amacıyla kullanılır (Kolen & Brennan, 2014). Elde edilen sonuçları standartlaştırmak için DTM değeri eski formun standart sapmasına bölünerek standartlaştırılmış DTM değeri (SDTM) elde edilir (Kim, 2006; Kolen & Brennan, 2014). Bu araştırmada farklı ölçek dönüştürme yöntemleri kullanılarak elde edilen eşitleme sonuçlarının karşılaştırılmasında SDTM ölçütü kullanılmıştır.

Betimsel istatistikler. Verilerin analizinde öncelikle betimsel istatistikler hesaplanmış ve Tablo 2'de verilmiştir.

Tablo 2
Kitapçıklara İlişkin Betimsel İstatistikler

Test İstatistikleri	Kitapçık-4	Kitapçık-11
Madde Sayısı	35	35
Kişi Sayısı	348	368
Aritmetik Ortalama	14.37	13.74
Ortanca	13.00	12.00
Tepe Değer	9.00	8.00
En Düşük	1.00	1.00
En Yüksek	36.00	35.00
Standart Sapma	8.24	7.52
Varyans	67.96	56.49
Çarpıklık Katsayısı	0.638	0.766
Basıklık Katsayısı	-0.607	-0.167

Tablo 2'deki değerler incelendiğinde 4 numaralı kitapçık 348 öğrenciye uygulanmıştır ve 35 sorudan alınan en düşük puan 1, en yüksek puan 36'dır. Testin ortalaması 14.37 olarak bulunmuştur. 368 öğrencinin cevapladığı 11 numaralı kitapçık incelendiğinde, 35 maddeden en düşük 1, en yüksek 35 puan alınmıştır. Dağılıma ilişkin aritmetik ortalama 13.74 olarak elde edilmiştir. Genel olarak puan dağılımlarına bakıldığında merkezi eğilim ölçülerinin birbirine yakın olması, çarpıklık ve basıklık katsayılarının [+1,-1] aralığında olması nedeniyle puanların normale yakın bir dağılım gösterdiği söylenebilir (Büyüköztürk, Çokluk & Köklü, 2013).

Kitapçıklardan elde edilen puan ortalamaları arasında anlamlı bir farklılık olup olmadığını test etmek için bağımsız gruplar t-testi yapılmış ve sonuçlar Tablo 3'te verilmiştir.

Tablo 3
Ortalamalar Arası Farkın Manidarlığına İlişkin Yapılan Bağımsız Gruplar T-Testi Sonuçları

Kitapçık	N	X	S	df	t	p
4	348	14.37	8.24	698.727	1.074	0.283
11	368	13.74	7.52			

p>.05

Analiz sonucunda her iki kitapçıktan alınan puan ortalamaları arasında anlamlı bir fark bulunmamıştır, t(698.727) = 1.074, p>.05.

Madde tepki kuramı varsayımları. Madde tepki kuramının tek boyutluluk ve yerel bağımsızlık olmak üzere iki temel varsayımı vardır.

Tek boyutluluk. MTK'nın temel varsayımlardan ilki tek boyutluluktur. Madde tepki kuramı bireyin test performansının altında tek bir gizil yeteneğin olduğunu varsayar (de Ayala, 2009). Bu varsayımın karşılanması için test performansını etkileyen baskın bir faktörün olması yeterli görülmektedir (Hambleton & Swaminathan, 1985). Bu nedenle tek boyutluluk varsayımını test etmek için FAKTOR 10.4 (Lorenzo-Seva & Ferrando, 2006) programı ile açımlayıcı faktör analizi yapılmıştır. Araştırma verilerinde kısmi puanlanan maddeler de yer aldığı için faktör analizi polikorik korelasyon matrisi oluşturularak gerçekleştirilmiştir.

Tablo 4
Kitapçıklara İlişkin Açımlayıcı Faktör Analizi Sonuçları

	Kitapçık-4			**Kitapçık-11**		
Faktör	Özdeğer	Varyans Yüzdesi	Yığılmalı Yüzde	Özdeğer	Varyans Yüzdesi	Yığılmalı Yüzde
1	6,28484	17,9570	17,9570	7 ,71605	22,0460	22,0460
2	1,77273	5,0650	23,2200	1,77323	5,0660	27,1120

Tablo 4 incelendiğinde 4 numaralı kitapçık için birinci faktörün özdeğeri ikinci faktörün yaklaşık 4 katına eşittir. 11 numaralı kitapçık incelendiğinde ise birinci faktörün özdeğerinin ikinci faktörün 4 katından fazla olduğu görülmektedir.

Açımlayıcı faktör analizi ile elde edilen tek boyutlu yapının doğruluğunu sınamak için LISREL 8.7 (Jöreskog & Sorbön, 1986) programı ile doğrulayıcı faktör analizi yapılmış ve elde edilen değerler Tablo 5'te verilmiştir.

Tablo 5.

Doğrulayıcı Faktör Analizi Sonucu Elde Edilen Uyum İndeksleri

Uyum İndeksi	Mükemmel Uyum Değerleri	İyi Uyum Değerleri	Model Değerleri	
			Kitapçık-4	Kitapçık-11
x^2/sd	x^2/sd ≤ 3	x^2/sd ≤ 5	1.42	1.46
RMSEA	RMSEA ≤ 0.05	RMSEA < 0.08	0.04	0.04
GFI	GFI ≥ 0.95	GFI ≥ 0.90	0.88	0.88
AGFI	AGFI ≥ 0.95	AGFI ≥ 0.90	0.86	0.87
NFI	NFI ≥ 0.95	NFI ≥ 0.90	0.94	0.91
NNFI	NNFI ≥ 0.95	NNFI ≥ 0.90	0.98	0.97
CFI	CFI ≥ 0.95	CFI ≥ 0.90	0.98	0.97
RMR	RMR ≤ 0.05	RMR ≤ 0.08	0.01	0.01
SRMR	SRMR ≤ 0.05	SRMR ≤ 0.08	0.05	0.05

(Kaynak: Sümer, 2000; Brown, 2006; akt. Çokluk, Şekercioğlu & Büyüköztürk, 2014)

Tablo 5'e göre Ki-kare/Serbestlik Derecesi (x^2/sd), oranının 3'ten küçük, Yaklaşık Hataların Ortalama Karakökü (RMSEA) değerlerinin ise 0.05'ten küçük olduğu görülmektedir. Normlaştırılmış Uyum İndeksi (NFI) iyi uyuma işaret ederken, Uyum İyiliği İndeksi (GFI) ile Düzeltilmiş Uyum İyiliği İndeksi (AGFI) zayıf uyuma karşılık gelmektedir. Bunun yanında Normlaştırılmamış Uyum İndeksi (NNFI) ve Karşılaştırmalı Uyum İndeksi (CFI) ise mükemmel uyuma işaret etmektedir. Artık Ortalamaların Karakökü (RMR) ve Standardize Edilmiş Artık Ortalamaların Karakökü (SRMR) değerleri incelendiğinde her iki kitapçık için de 0'a yakın değerler elde edilmiştir. Sonuç olarak tek boyutlu model için elde edilen uyum indeksleri incelendiğinde modelin veri ile iyi uyum gösterdiği görülmüştür. Bu nedenle tek boyutluluk varsayımının karşılandığı söylenebilir.

Yerel bağımsızlık. Yerel bağımsızlık, yetenek değişkeni sabit tutulduğunda herhangi iki maddeye verilen cevapların istatistiksel olarak birbirinden bağımsız olması demektir (Lord, 1980; akt. Hambleton & Swaminathan, 1985). Yerel bağımsızlık tek boyutlulukla ilişkili bir kavramdır. Genel olarak tek boyutluluğun karşılanması durumunda yerel bağımsızlığın da sağlandığı kabul edilmektedir (Hambleton & Swaminathan, 1985). Bu nedenle bu çalışmada tek boyutluluk varsayımı sağlandığı için yerel bağımsızlık varsayımının da karşılandığı kabul edilmiştir.

Çalışmada tek boyutluluk varsayımı incelendikten sonra her iki kitapçık için madde ve yetenek parametreleri kestirilmiştir. Yetenek parametre kestiriminde EAP (Expected A Posteriori) yöntemi kullanılmıştır. Parametre kestiriminin ardından ölçek dönüştürme ve eşitleme işlemleri gerçekleştirilmiştir. Son aşamada ise farklı yöntemlerden elde edilen hata miktarını belirlemek için ağırlıklandırılmış hata kareleri ortalaması (WMSE) hesaplanmıştır. WMSE katsayısının matematiksel ifadesi şu şekildedir:

$$WMSE = \sum_{i=1}^{k-1} f_i (X_E - X_{crit})^2 \,/\, \sum_{i=1}^{k} f_i \, S^2 Y$$

k : Y testindeki madde sayısı

S^2 : Y testindeki ham puanların varyansı

X_{crit} : Y testindeki i. ham puan

X_E : X testindeki i. ham puana eşit olan puan

f_i : Y testindeki i. ham puanın frekansı

Bu çalışmada 11 numaralı kitapçıktan alınan puanlar 4 numaralı kitapçıktan alınan puanlara eşitlenmiştir. Bu nedenle eşitlikteki Y testi 4 numaralı kitapçığı (referans test), X testi ise 11 numaralı kitapçığı (dönüştürülen test) temsil etmektedir. WMSE katsayısının azalması düşük eşitleme hatasına karşılık gelmektedir.

Bulgular

Birinci Alt Probleme İlişkin Bulgular

PISA 2012 Matematik okuryazarlığı testinde Kitapçık-11'den elde edilen puanların Kitapçık-4'ten elde edilen puanlara eşitlenmesinde ortalama-ortalama yönteminden elde edilen eşitlenmiş puanlar nasıldır?

Bu alt problemin çözümü için öncelikle iki kitapçıktan elde edilen ham puanlar üzerinde ortalama-ortalama yöntemine göre ölçek dönüştürme işlemi gerçekleştirilmiş, ardından MTK gerçek-puan ve MTK gözlenen-puan eşitleme yöntemleri kullanılarak 11 numaralı kitapçıktan elde edilen puanlar 4 numaralı kitapçıktan elde edilen puanlara eşitlenmiştir. Ham puanlar, eşitlenmiş puanlar ve ham puanlar ile eşitlenmiş puanlar arasındaki fark Tablo 6'da verilmiştir.

Tablo 6

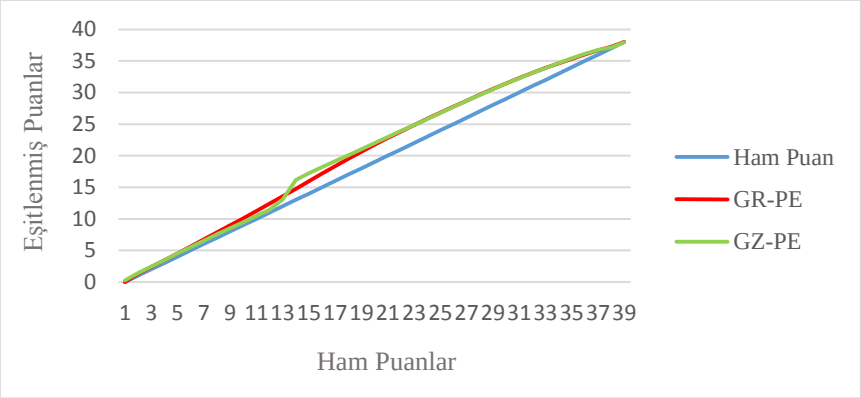
Ortalama-Ortalama Yönteminden Elde Edilen Eşitlenmiş Puanlar

MTK Gerçek-Puan Eşitleme						MTK Gözlenen-Puan Eşitleme					
HP	EP	F	HP	EP	F	HP	EP	F	HP	EP	F
0	0	0	**20**	22.79	-2.79	**0**	0.27	-0.27	**20**	22.95	-2.95
1	1.32	-0.32	**21**	23.83	-2.83	**1**	1.41	-0.41	**21**	23.90	-2.90
2	2.37	-0.37	**22**	24.84	-2.84	**2**	2.42	-0.42	**22**	24.85	-2.85
3	3.44	-0.44	**23**	25.83	-2.83	**3**	3.46	-0.46	**23**	25.81	-2.81
4	4.52	-0.52	**24**	26.80	-2.80	**4**	4.50	-0.50	**24**	26.76	-2.76
5	5.62	-0.62	**25**	27.76	-2.76	**5**	5.53	-0.53	**25**	27.71	-2.71
6	6.73	-0.73	**26**	28.70	-2.70	**6**	6.55	-0.55	**26**	28.66	-2.66
7	7.84	-0.84	**27**	29.63	-2.63	**7**	7.54	-0.54	**27**	29.59	-2.59
8	8.96	-0.96	**28**	30.54	-2.54	**8**	8.51	-0.51	**28**	30.51	-2.51
9	10.09	-1.09	**29**	31.43	-2.43	**9**	9.48	-0.48	**29**	31.40	-2.40
10	11.23	-1.23	**30**	32.28	-2.28	**10**	10.46	-0.46	**30**	32.25	-2.25
11	12.39	-1.39	**31**	33.10	-2.10	**11**	11.45	-0.45	**31**	33.07	-2.07
12	13.56	-1.56	**32**	33.88	-1.88	**12**	13.08	-1.08	**32**	33.87	-1.87
13	14.76	-1.76	**33**	34.63	-1.63	**13**	16.16	-3.16	**33**	34.66	-1.66

14	15.96	-1.96	**34**	35.34	-1.34	**14**	17.22	-3.22	**34**	35.41	-1.41
15	17.16	-2.16	**35**	36.03	-1.03	**15**	18.18	-3.18	**35**	36.13	-1.13
16	18.34	-2.34	**36**	36.68	-0.68	**16**	19.14	-3.14	**36**	36.72	-0.72
17	19.50	-2.50	**37**	37.25	-0.25	**17**	20.09	-3.09	**37**	37.16	-0.16
18	20.63	-2.63	**38**	38.00	0.00	**18**	21.04	-3.04	**38**	37.94	0.06
19	21.73	-2.73				**19**	21.99	-2.99			

Not: HP = Ham Puan; EP = Eşitlenmiş Puan; F = Fark

Tablo 6'da verilen MTK gerçek-puan eşitleme sonuçları incelendiğinde puan ölçeği boyunca (0 ve 38 ham puan hariç) bütün eşitlenmiş puanların ham puanlardan büyük olduğu görülmektedir. Bu bulguya dayalı olarak Kitapçık-11'de yer alan matematik testinin puan ölçeği boyunca Kitapçık 4'te yer alan matematik testinden daha zor olduğu söylenebilir. Benzer durum gözlenen-puan eşitleme sonuçlarında da görülmektedir. Tablo 6 incelendiğinde puan ölçeği boyunca (38 ham puan hariç), eşitlenmiş puanların ham puanlardan büyük olduğu görülmektedir. Bu bulguya dayalı olarak aynı şekilde Kitapçık-11'de yer alan matematik testinin puan ölçeği boyunca Kitapçık 4'te yer alan matematik testinden daha zor olduğu söylenebilir. Tablo 6'ya göre gerçek-puan eşitleme yöntemi için elde edilen farklar -2.84 ile 0 arasında değişmekle birlikte farkların 0-22 puanları arasında artarken, 23-38 puanları arasında azaldığı görülmektedir. Gözlenen-puan eşitlemede ise farkların -3.22 ile 0.06 arasında değiştiği ve puan ölçeği boyunca yer yer artıp azaldığı görülmektedir. Ham puanlar ile eşitlenmiş puanlar arasındaki ilişki Şekil 1'de verilmiştir.



*GR-PE:Gerçek-puan eşitleme *GZ-PE:Gözlenen-puan eşitleme

Şekil 1. Ham puanlarla ortalama-ortalama yönteminden elde edilen eşitlenmiş puanlar arasındaki ilişki

Şekil 1 incelendiğinde alt puanlarda ham puanlarla eşitlenmiş puanların benzerlik gösterdiği, dağılımın ortalarında aradaki farkın arttığı ve üst puanlarda ise farkın giderek azalarak dağılımın uç noktalarda birleştiği görülmektedir. Bunun yanında gerçek ve gözlenen-puan eşitleme yöntemlerinden elde edilen puanların genel olarak birbiriyle paralellik gösterdiği söylenebilir.

İkinci Alt Probleme İlişkin Bulgular

PISA 2012 Matematik okuryazarlığı testinde Kitapçık-11'den elde edilen puanların Kitapçık-4'ten elde edilen puanlara eşitlenmesinde ortalama-standart sapma yönteminden elde edilen eşitlenmiş puanlar nasıldır?

Bu alt probleme ilişkin ölçek dönüştürme işlemi ortalama-standart sapma yöntemi kullanılarak gerçekleştirilmiştir. Elde edilen ham puanlar, eşitlenmiş puanlar ve ham puanlar ile eşitlenmiş puanlar arasındaki fark Tablo 7'de verilmiştir.

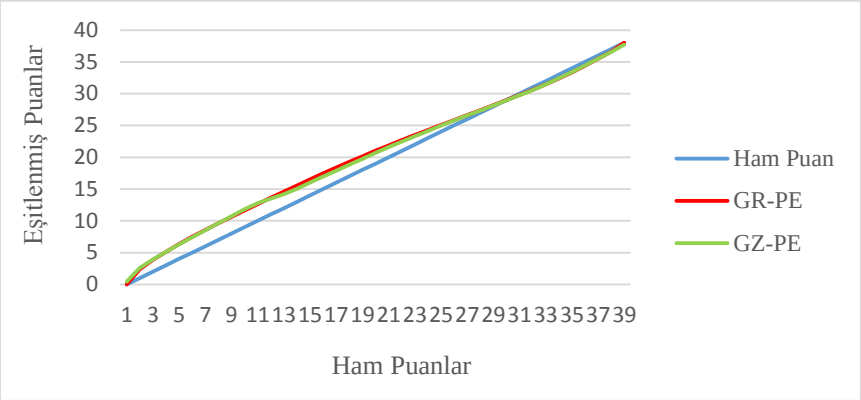
Tablo 7
Ortalama-Standart Sapma Yönteminden Elde Edilen Eşitlenmiş Puanlar

MTK Gerçek-Puan Eşitleme						**MTK Gözlenen-Puan Eşitleme**					
HP	EP	F	HP	EP	F	HP	EP	F	HP	EP	F
0	0	0	**20**	21.86	-1.86	**0**	0.46	-0.46	**20**	21.58	-1.58
1	2.39	-1.39	**21**	22.68	-1.68	**1**	2.60	-1.60	**21**	22.46	-1.46
2	3.85	-1.85	**22**	23.49	-1.49	**2**	3.92	-1.92	**22**	23.30	-1.30
3	5.15	-2.15	**23**	24.29	-1.29	**3**	5.14	-2.14	**23**	24.14	-1.14
4	6.35	-2.35	**24**	25.07	-1.07	**4**	6.30	-2.30	**24**	24.97	-0.97
5	7.49	-2.49	**25**	25.85	-0.85	**5**	7.42	-2.42	**25**	25.78	-0.78
6	8.57	-2.57	**26**	26.63	-0.63	**6**	8.49	-2.49	**26**	26.58	-0.58
7	9.62	-2.62	**27**	27.41	-0.41	**7**	9.61	-2.61	**27**	27.37	-0.37
8	10.64	-2.64	**28**	28.19	-0.19	**8**	10.72	-2.72	**28**	28.15	-0.15
9	11.65	-2.65	**29**	28.98	0.02	**9**	11.80	-2.80	**29**	28.94	0.06
10	12.64	-2.64	**30**	29.79	0.21	**10**	12.78	-2.78	**30**	29.74	0.26
11	13.62	-2.62	**31**	30.61	0.39	**11**	13.49	-2.49	**31**	30.59	0.41
12	14.60	-2.60	**32**	31.47	0.53	**12**	14.18	-2.18	**32**	31.47	0.53
13	15.56	-2.56	**33**	32.37	0.63	**13**	15.06	-2.06	**33**	32.41	0.59
14	16.52	-2.52	**34**	33.34	0.66	**14**	16.00	-2.00	**34**	33.41	0.59
15	17.46	-2.46	**35**	34.37	0.63	**15**	16.95	-1.95	**35**	34.44	0.56
16	18.38	-2.38	**36**	35.48	0.52	**16**	17.89	-1.89	**36**	35.52	0.48
17	19.28	-2.28	**37**	36.62	0.38	**17**	18.83	-1.83	**37**	36.58	0.42
18	20.16	-2.16	**38**	38.00	0.00	**18**	19.76	-1.76	**38**	37.74	0.26
19	21.02	-2.02				**19**	20.67				

Not: HP = Ham Puan; EP = Eşitlenmiş Puan; F = Fark

Tablo 7'deki puan dağılımları incelendiğinde ham puanların 0 ile 38 puanları arasında değer aldığı, eşitlenmiş puanların ise gerçek-puan eşitleme yöntemi için 0-38 arasında, gözlenen-puan eşitleme yöntemi için ise 0-37.74 arasında değiştiği görülmektedir. Ham puanlarla eşitlenmiş puanlar arasındaki ilişki incelendiğinde, her iki eşitleme yönteminde de 0 ile 28 arasındaki ham puanların eşitlenmiş puanlardan küçük, 29-38 arasındaki ham puanların ise eşitlenmiş puanlardan büyük olduğu görülmektedir. Bu nedenle güçlük düzeyinin puan dağılımı boyunca farklılık gösterdiğini söyleyebiliriz. Tablo 7'de verilen ham puanlarla eşitlenmiş puanalar arasındaki farklar incelendiğinde, farkların gerçek-puan eşitleme yönteminde -2.65 ile 0 arasında, gözlenen-puan eşitlemede ise -2.80 ile 0.06 arasında değerler aldığı bulunmuştur. Her iki eşitleme yöntemine göre

elde edilen fark puanlarının puan ölçeği boyunca yer yer artıp azaldığı görülmektedir. Şekil 2'de ham puanlar ile eşitlenmiş puanlar arasındaki ilişki görülmektedir.



*GR-PE:Gerçek-puan eşitleme *GZ-PE:Gözlenen-puan eşitleme

Şekil 2. Ham ilen eşitlenmiş puanlar arasındaki ilişki

Şekil 2 incelendiğinde ham puanlarla eşitlenmiş puanlar arasındaki farkların alt puanlarda fazla iken, üst puanlara doğru azaldığı görülmektedir.

Üçüncü Alt Probleme İlişkin Bulgular

PISA 2012 Matematik okuryazarlığı testinde Kitapçık-11'den elde edilen puanların Kitapçık-4'ten elde edilen puanlara eşitlenmesinde Stocking-Lord yönteminden elde edilen eşitlenmiş puanlar nasıldır?

Bu alt probleme ilişkin ölçek dönüştürme işlemi, Stocking-Lord yöntemi için tekrar edilerek puan dağılımları Tablo 8'de verilmiştir.

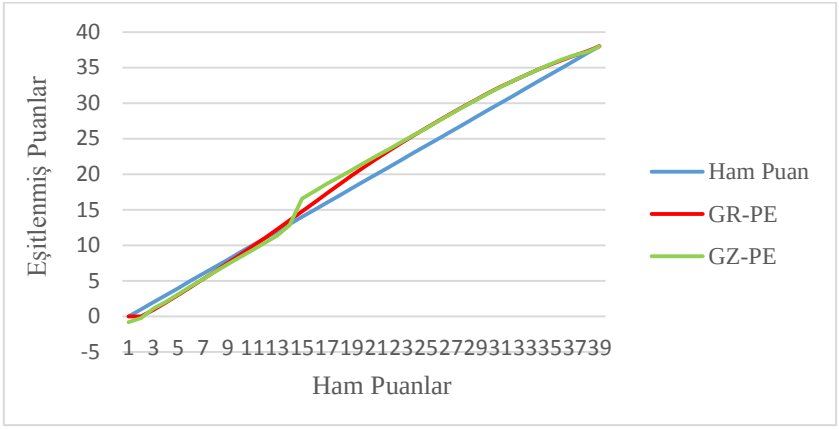
Tablo 8
Stocking-Lord Yönteminden Elde Edilen Eşitlenmiş Puanlar

	MTK Gerçek-Puan Eşitleme						**MTK Gözlenen-Puan Eşitleme**					
HP	EP	F	HP	EP	F		HP	EP	F	HP	EP	F
0	0	0	**20**	22.16	-2.16		**0**	-0.78	0.78	**20**	22.55	-2.55
1	0	1	**21**	23.29	-2.29		**1**	-0.20	1.20	**21**	23.53	-2.53
2	0.93	1.07	**22**	24.38	-2.38		**2**	1.08	0.92	**22**	24.50	-2.50
3	1.95	1.05	**23**	25.44	-2.44		**3**	2.05	0.95	**23**	25.48	-2.48
4	3.02	0.98	**24**	26.49	-2.49		**4**	3.12	0.88	**24**	26.46	-2.46
5	4.12	0.88	**25**	27.51	-2.51		**5**	4.19	0.81	**25**	27.45	-2.45
6	5.23	0.77	**26**	28.52	-2.52		**6**	5.26	0.74	**26**	28.45	-2.45
7	6.37	0.63	**27**	29.51	-2.51		**7**	6.31	0.69	**27**	29.44	-2.44
8	7.51	0.49	**28**	30.46	-2.46		**8**	7.34	0.66	**28**	30.41	-2.41
9	8.68	0.32	**29**	31.39	-2.39		**9**	8.35	0.65	**29**	31.35	-2.35
10	9.85	0.15	**30**	32.27	-2.27		**10**	9.35	0.65	**30**	32.24	-2.24
11	11.05	-0.05	**31**	33.11	-2.11		**11**	10.35	0.65	**31**	33.09	-2.09
12	12.27	-0.27	**32**	33.91	-1.91		**12**	11.35	0.65	**32**	33.90	-1.90
13	13.51	-0.51	**33**	34.67	-1.67		**13**	12.83	0.17	**33**	34.69	-1.69
14	14.77	-0.77	**34**	35.39	-1.39		**14**	16.58	-2.58	**34**	35.45	-1.45
15	16.05	-1.05	**35**	36.07	-1.07		**15**	17.64	-2.64	**35**	36.17	-1.17
16	17.32	-1.32	**36**	36.71	-0.71		**16**	18.63	-2.63	**36**	36.74	-0.74

17	18.58	-1.58	**37**	37.26	-0.26	**17**	19.62	-2.62	**37**	37.18	-0.18
18	19.81	-1.81	**38**	38.00	0.00	**18**	20.60	-2.60	**38**	37.95	0.05
19	21.00	-2.00				**19**	21.58	-2.58			

Not: HP = Ham Puan; EP = Eşitlenmiş Puan; F = Fark

Tablo 8'e göre eşitlenmiş puan dağılımları incelendiğinde gerçek-puan eşitlemede puanlar 0-38 arasında elde edilirken, gözlenen-puan eşitlemede ise elde edilen puanlar -0.78 ile 37.95 arasında değişmektedir. Ham puanlarla eşitlenmiş puanlar arasındaki ilişki incelendiğinde, gerçek-puan eşitlemede 0-10 puanları arasındaki ham puanların eşitlenmiş puanlardan büyük, 11-37 arasındaki puanların ise (38 ham puan hariç) küçük olduğu görülmektedir. Gözlenen-puan eşitlemede ise 0-13 arasındaki ham puanların eşitlenmiş puanlardan büyük, 14-37 arasında ise küçük olduğu görülmektedir. Tablo 8 incelendiğinde ham puanlarla eşitlenmiş puanlar arasındaki farkların gerçek-puan eşitlemede -2.52 ile 0 arasında, gözlenen-puan eşitlemede -2.64 ile 0.05 arasında değiştiği görülmektedir. Her iki eşitleme yönteminde de fark puanlarının puan ölçeği boyunca yer yer artıp azaldığı görülmektedir. Ham puanlar ile eşitlenmiş puanlar arasındaki ilişki Şekil 3'te verilmiştir.



*GR-PE:Gerçek-puan eşitleme *GZ-PE:Gözlenen-puan eşitleme

Şekil 3. Ham puanlarla Stocking-Lord yönteminden elde edilen eşitlenmiş puanlar arasındaki ilişki

Şekil 3'e göre alt puanlarda eşitlenmiş puanlar ham puanlarla paralellik göstermektedir. Dağılımın ortalarında aralarındaki fark artarken, üst puanlara doğru ise farkların giderek azalarak eşitlenmiş puanların ham puanlara yakın değerler verdiği görülmektedir.

Dördüncü Alt Probleme İlişkin Bulgular

PISA 2012 Matematik okuryazarlığı testinde Kitapçık-11'den elde edilen puanların Kitapçık-4'ten elde edilen puanlara eşitlenmesinde Haebara yönteminden elde edilen eşitlenmiş puanlar nasıldır?

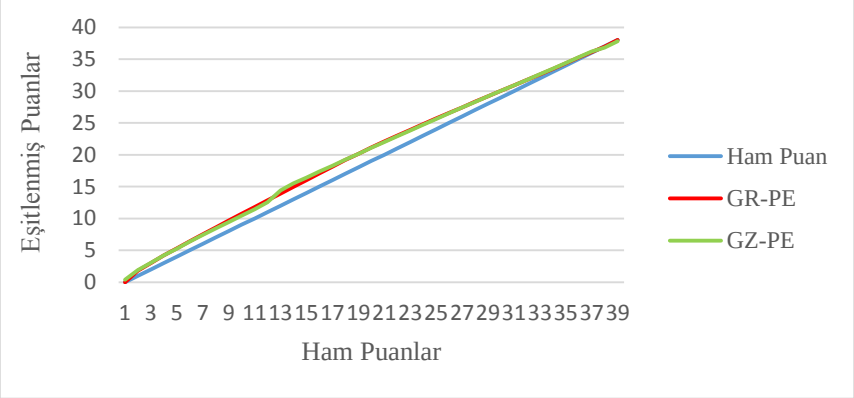
Tablo 9

Haebara Yönteminden Elde Edilen Eşitlenmiş Puanlar

MTK Gerçek-Puan Eşitleme						MTK Gözlenen-Puan Eşitleme					
HP	EP	F	HP	EP	F	HP	EP	F	HP	EP	F
0	0	0	**20**	22.08	-2.08	**0**	0.35	-0.35	**20**	22.00	-2.00
1	1.76	-0.76	**21**	23.01	-2.01	**1**	1.87	-0.87	**21**	22.91	-1.91
2	3.00	-1.00	**22**	23.92	-1.92	**2**	3.05	-1.05	**22**	23.82	-1.82
3	4.18	-1.18	**23**	24.81	-1.81	**3**	4.18	-1.18	**23**	24.72	-1.72
4	5.32	-1.32	**24**	25.70	-1.70	**4**	5.28	-1.28	**24**	25.62	-1.62
5	6.43	-1.43	**25**	26.57	-1.57	**5**	6.34	-1.34	**25**	26.51	-1.51
6	7.52	-1.52	**26**	27.44	-1.44	**6**	7.39	-1.39	**26**	27.40	-1.40
7	8.59	-1.59	**27**	28.30	-1.30	**7**	8.42	-1.42	**27**	28.29	-1.29
8	9.66	-1.66	**28**	29.16	-1.16	**8**	9.43	-1.43	**28**	29.17	-1.17
9	10.71	-1.71	**29**	30.02	-1.02	**9**	10.44	-1.44	**29**	30.04	-1.04
10	11.77	-1.77	**30**	30.88	-0.88	**10**	11.44	-1.44	**30**	30.90	-0.90
11	12.82	-1.82	**31**	31.74	-0.74	**11**	12.54	-1.54	**31**	31.76	-0.76
12	13.89	-1.89	**32**	32.60	-0.60	**12**	14.39	-2.39	**32**	32.63	-0.63
13	14.96	-1.96	**33**	33.47	-0.47	**13**	15.50	-2.50	**33**	33.52	-0.52
14	16.02	-2.02	**34**	34.35	-0.35	**14**	16.43	-2.43	**34**	34.42	-0.42
15	17.08	-2.08	**35**	35.24	-0.24	**15**	17.36	-2.36	**35**	35.32	-0.32
16	18.13	-2.13	**36**	36.14	-0.14	**16**	18.28	-2.28	**36**	36.19	-0.19
17	19.15	-2.15	**37**	36.98	0.02	**17**	19.21	-2.21	**37**	36.84	0.16
18	20.15	-2.15	**38**	38.00	0.00	**18**	20.14	-2.14	**38**	37.83	0.17
19	21.13	-2.13				**19**	21.07	-2.07			

Not: HP = Ham Puan; EP = Eşitlenmiş Puan; F = Fark

Ölçek dönüştürme işlemi son olarak Haebara yöntemi için gerçekleştirilmiştir. Tablo 9'da ham puanlar, eşitlenmiş puanlar ve ham puanlar ile eşitlenmiş puanlar arasındaki farklar görülmektedir. Elde edilen puan dağılımı incelendiğinde, eşitlenmiş puanlar gerçek-puan eşitleme yöntemi kullanıldığında 0-38 puanları arasında değişirken, gözlenen-puan eşitleme yöntemi kullanıldığında ise 0.35-37.83 arasında değer almaktadır. Ham puanlarla eşitlenmiş puanlar arasındaki ilişki incelendiğinde, her iki eşitleme yöntemi için de puan ölçeği boyunca (0, 37 ve 38 hariç) ham puanların eşitlenmiş puanlardan küçük olduğu görülmektedir. Tablo 9'a göre ham puanlarla eşitlenmiş puanlar arasındaki farklar gerçek-puan eşitleme yönteminde -2.15 ile 0 arasında değişmekle birlikte, farkların 0-18 puanları arasında artarken, 19-38 puanları arasında azaldığı görülmektedir. Gözlenen-puan eşitleme yönteminde ise fark puanları -2.50 ile 0 arasında değişmekte, elde edilen farklar 0-13 puanları arasında artarken, 14-37 puanları arasında tekrar azalmaktadır. Ham puanlarla eşitlenmiş puanlar arasındaki ilişki Şekil 4'te verilmiştir.



*GR-PE:Gerçek-puan eşitleme *GZ-PE:Gözlenen-puan eşitleme

Şekil 4. Ham puanlarla Haebara yönteminden elde edilen eşitlenmiş puanlar arasındaki ilişki

Şekil 4'e göre puan ölçeği boyunca eşitlenmiş puanların ham puanlara yakın değerler verdiği görülmektedir.

Beşinci Alt Probleme İlişkin Bulgular

PISA 2012 Matematik okuryazarlığı testinde Kitapçık-11'den elde edilen puanların Kitapçık-4'ten elde edilen puanlara eşitlenmesinde en az eşitleme hatasına sahip ölçek dönüştürme ve test eşitleme yöntemi hangisidir?

Bu alt probleme ilişkin farklı yöntemlerden elde edilen sonuçlar için WMSE katsayısı hesaplanmış ve Tablo 10'da verilmiştir.

Tablo 10
Farklı Yöntemlerden Elde Edilen WMSE Katsayıları

	MTK Gerçek-Puan	**MTK Gözlenen-Puan**
Ortalama-Ortalama	0.0508	0.0625*
Ortalama-Standart Sapma	0.0717*	0.0634*
Stocking-Lord	**0.0298**	0.0453
Haebara	0.0428	**0.0432**

*SDTM değerinden büyük olan değerler

Tablo 10'a göre gerçek-puan eşitleme yöntemi için elde edilen WMSE katsayıları ortalama-ortalama yöntemi için 0.0508, ortalama-standart sapma yönteminde 0.0717, Stocking-Lord yönteminde 0.0298, Haebara yönteminde ise 0.0428 olarak bulunmuştur. Gözlenen-puan eşitleme yöntemi kullanıldığında ise ortalama-ortalama yönteminde hata miktarı 0.0625, ortalama-standart sapmada 0.0634, Stocking-Lord yönteminde 0.0453 ve Haebara yönteminde 0.0432 olarak elde edilmiştir. Elde edilen sonuçlara göre en az hata miktarını, gerçek-puan eşitleme yöntemi kullanıldığında Stocking-Lord yönteminin verdiği, gözlenen-puan eşitleme yöntemi kullanıldığında ise Haebara yönteminin verdiği görülmektedir. En fazla hata oranını ise her iki eşitleme yönteminde de ortalama-standart sapmanın verdiği bulgusuna ulaşılmıştır. Ayrıca, elde edilen sonuçlar incelendiğinde genel olarak MTK gerçek-puan eşitleme yönteminin MTK gözlenen-puan eşitleme yönteminden daha iyi sonuç verdiği görülmüştür.

Çalışmada son olarak farklı yöntemlerden elde edilen sonuçlar standartlaştırılmış DTM değeri ile karşılaştırılmıştır. Bu çalışmada "0.5" DTM ölçütü kullanılmıştır. SDTM değeri ise DTM değerinin eski formun (Kitapçık-4) standart sapmasına bölünmesiyle (0.5/8.24) "0.06" olarak elde edilmiştir. Tablo 10'a göre MTK gerçek-puan eşitleme yönteminde ortalama-standart sapma yöntemi dışında

elde edilen WMSE değerlerinin SDTM değerinden düşük olduğu görülmüştür. MTK gözlenen-puan eşitleme yönteminde ise Stocking-Lord ve Haebara yönteminden elde edilen eşitlenmiş puanlara ait WMSE değeri SDTM değerinden küçüktür. Bu bulgular ışığında, MTK gerçek-puan eşitleme yöntemi seçildiğinde ortalama-standart sapma dışındaki ölçek dönüştürme yöntemlerinin kullanılması ve MTK gözlenen-puan eşitleme yöntemi seçildiğinde ise Stocking-Lord ve Haebara yöntemlerinin kullanılarak puanların eşitlenmesi önerilebilir.

Tartışma, Sonuç ve Öneriler

Bu çalışmada, farklı ölçek dönüştürme yöntemleri kullanılarak MTK gerçek-puan eşitleme ve gözlenen-puan eşitleme yöntemleri ile iki kitapçıktan elde edilen puanlar eşitlenmiş ve farklı yöntemlerden elde edilen sonuçlar gerçek veriler üzerinde karşılaştırılmıştır. Çalışma sonucunda karakteristik eğri dönüştürme yöntemlerinin (Stocking Lord, Haebara) moment yöntemlerinden (ortalama-ortalama, ortalama-standart sapma) daha az hata miktarı verdiği bulunmuştur. Bu bulgu literatürde farklı değerlendirme ölçütlerinin kullanılarak ölçek dönüştürme yöntemlerinin karşılaştırıldığı benzer çalışmalarla (Baker & Al-Karni, 1991; Gök, 2012; Hanson & Beguin, 2002; Kim & Kolen, 2006; Kim & Lee, 2004; Kim & Lee, 2006; Öztürk-Gübeş & Kelecioğlu, 2016; Uysal, 2014) tutarlılık göstermektedir. Elde edilen bu sonuca ilişkin Stocking ve Lord (1983), karakteristik eğri yöntemlerinin uyumsuzluğu gideren bir yapısının olduğunu ve farklı koşullardaki değişime daha dayanıklı olduğunu belirtmişlerdir. Karakteristik eğri dönüştürme yöntemlerinden elde edilen sonuçlar incelendiğinde gerçek-puan eşitlemede en az hata miktarını Stocking Lord yönteminin verdiği, gözlenen-puan eşitlemede ise Haebara yönteminin verdiği bulunmuştur. Literatürde gerçekleştirilen bazı çalışmalarda (Cohen & Kim, 1998; French, 1996; Li, Lissitz & Yang, 1999; Uysal, 2014) Stocking Lord yöntemi daha iyi sonuç verirken, bazılarında ise (Kim & Lee, 2006; Lee & Ban, 2010) Haebara yönteminin daha düşük hata miktarı verdiği tespit edilmiştir. Kim ve Kolen (2007)'in çalışmasında ise her iki yöntemin benzer sonuçlar verdiği bulunmuştur. Moment yöntemlerinden elde edilen sonuçlara göre, en fazla eşitleme hatasını ortalama-standart sapma yönteminin verdiği görülmüştür. Bu bulgu Ogasawara (2000) ve Speron (2009)'un çalışma bulgusu ile örtüşmektedir. Baker ve Al-Karni (1991), bazen ortalama-ortalamanın ortalama-standart sapmadan daha tercih edilebilir olduğunu belirtmişler ve bu durumu, ortalamanın standart sapmadan daha kararlı olması şeklinde yorumlamışlardır.

Çalışmada ayrıca farklı ölçek dönüştürme yöntemleri kullanıldığında MTK gerçek-puan ve MTK gözlenen-puan eşitleme yöntemlerinden elde edilen sonuçlar incelenmiştir. Buna göre genel olarak gerçek-puan eşitleme yönteminin gözlenen-puan eşitleme yönteminden daha düşük hata miktarına

sahip olduğu bulunmuştur. Bu bulgu Han, Kolen ve Pohlmann (1997)'ın gerçekleştirdiği çalışmada sonucu ile örtüşürken Hagge (2010)'nin çalışma bulgusu ile çelişmektedir. Hagge (2010)'nin gerçek veriler üzerinde yürüttüğü çalışmasında gözlenen-puan eşitleme yönteminin daha az hata verme eğiliminde olduğu görülmüştür. Bu sonucun örneklem büyüklüğü ile ilişkili olabileceği düşünülmektedir. Tsai, Hanson, Kolen ve Forsyth (2001) tarafından ikili puanlanan gerçek veriler üzerinde gerçekleştirilen çalışmada, gözlenen-puan eşitleme yönteminin daha iyi sonuç verdiği tespit edilmiştir. Bu durum eşitlenen formlardaki madde türlerinin farklılaşmasından kaynaklanıyor olabilir. Çünkü ikili puanlanan ve karma maddelerin yer aldığı testlerden elde edilen sonuçlar farklılık gösterebilmektedir. Lord ve Wingersky (1984), ortak maddeli test deseninde gerçekleştirdikleri çalışmada, gerçek ve gözlenen-puan eşitlemenin benzer sonuçlar verdiğini tespit etmişlerdir. Kolen (1981) ise çalışmasında bir, iki ve üç parametreli modelden elde edilen sonuçları gerçek veriler üzerinde karşılaştırmıştır. Çalışma sonucunda, bir ve iki parametreli model için gözlenen-puan eşitleme yönteminin daha kararlı sonuçlar verdiğini, üç parametreli modelde ise yöntemlerden elde edilen hata oranlarının farklı örneklemlerde değişiklik gösterdiği bulgusuna ulaşmıştır. Çalışması sonucunda en az hata veren yöntemin eşitleme koşullarına göre değiştiğini belirtmiştir.

Bu çalışmadan elde edilen sonuçlar doğrultusunda karma formattaki testlerde eşdeğer olmayan gruplar test deseni kullanılarak MTK'ya dayalı yapılacak eşitleme çalışmalarında, karakteristik eğri dönüştürme yöntemlerinin kullanılması önerilebilir. İleride yapılacak araştırmalarda ise eş zamanlı kestirim ile ayrı kestirimden elde edilen sonuçlar gerçek ve simülasyon veriler üzerinde karşılaştırılabilir.

MTK test eşitleme yöntemlerini kullanmak için önerilen örneklem büyüklüğü 3PL kullanıldığında 1000'dir (de Ayala, 2009; Hambleton, 1993; Jones, Smith & Talley, 2006; Yen & Fitzpatrick, 2006). Ancak, bu araştırma PISA Türkiye örnekleminden elde edilen veriler kullanılarak yürütüldüğü için örneklem büyüklüğü 716 kişi ile sınırlı kalmıştır. Gelecekte yapılacak olan araştırmalarda daha büyük örneklemler kullanılarak farklı ölçek dönüştürme yöntemlerinin karşılaştırılması önerilebilir.

Kaynakça

Angoff, W. H. (1984). *Scales, norms and equivalent scores*. Princeton, New Jersey: Educational Testing Service.

Baker, F. B. & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, *28* (2), 147- 162.

Büyüköztürk, Ş., Çokluk, Ö. & Köklü, N. (2013). *Sosyal bilimler için istatistik* (12. Baskı). Ankara: Pegem Akademi.

Cohen, A. S. & Kim, S. H. (1998). An investigation of linking methods under the graded response model. *Applied Psychological Measurement, 22*(2), 116-130.

Cook L. & Eignor D. R. (1991). NCME instructional module: IRT equating methods. *Educational Measurement: Issues and Practices, 10*(3), 37-45.

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

Çokluk, Ö., Şekercioğlu, G. & Büyüköztürk, Ş. (2014). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları* (3. Baskı). Ankara: Pegem Yayıncılık.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

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

Felan, G. D. (2002, February). *Test equating: mean, linear, equipercentile and item response theory*. Paper presented at the Annual Meeting of the Southwest Educational Research Association, Austin, Texas.

French, D. J. (1996). *The utility of Stocking-Lord's equating procedure for equating norm-referenced and criterion-referenced tests with both dichotomous and plytomous components*. Unpublished doctorate dissertation, University of Texas, Texas.

Gök, B. (2012). *Denk olmayan gruplarda ortak madde deseni kullanılarak madde tepki kuramına dayalı eşitleme yöntemlerinin karşılaştırılması.* Yayımlanmamış doktora tezi, Hacettepe Üniversitesi, Ankara.

Gültekin, S. (2014). Testlerde kullanılacak madde türleri, hazırlama ilkeleri ve puanlaması. N. Demirtaşlı (Ed.), *Eğitimde ölçme ve değerlendirme* (2. Baskı) içinde. Ankara: Edge Akademi.

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

Hagge, S. L. (2010). *The impact of equating method and format representation of common items on the adequacy of mixed format test equating using nonequivalent groups.* Unpublished doctorate
dissertation, University of Lowa, Lowa City.

Hambleton, R. K. (1989). *Item response theory: Introduction and bibliography.* (Rapor no:196) Amherst: University of Massachusetts.

Hambleton, R. K. (1993). Principles and selected applications of item response theory. R. Linn (Ed.), *Educational measurement* (3. Baskı) içinde. Washington, D.C.: American Council on

Education.

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

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

Han, T., Kolen, M. & Pohlmann, J. (1997). A comparison among IRT true and observed-score equatings and traditional equipercentile equating. *Applied Measurement in Education*, *10*(2), 105-121, doi: 10.1207/s15324818ame10021.

Hanson, B. A. & Beguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using seperate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26* (3), 3-24.

Harris, D. J. & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education*, *6* (3), 195-240.

Holland, P. W., Dorans, N. J. & Petersen, N. S. (2007). Equating test scores. C. R. Rao, S. Sinharay (Eds.), *Handbook of statistics: Pschometrics* (pp. 169-197) içinde. Amsterdam: Elsevier B. V.

Jones, P., Smith, R. V. & Talley, D. (2006). Developing test forms for small-scale achievement testing systems. S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* içinde. Mahwah, N. J.: L. Erlbaum.

Jöreskog, K. G. & Sorbön, D. (1986). *LISREL 8.7*: Prells a program for multivariate data screening and data summarization [Computer software]. Mooresville, Ind: Scientific Software Inc.

Kim, H. K. (2006). *The effect of repeaters on equating: A population invariance approach.* Unpublished doctorate dissertation, The University of Lowa, Lowa City.

Kim, S. & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, *26* (1), 25-41.

Kim, S. & Kolen, M. J. (2004). *STUIRT:* A computer program for scale transformation under unidimentional item response theory models [Computer software]. Lowa City, IA. The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Lowa.

Kim, S. & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education*, *19* (4), 357-381.

Kim, S. & Kolen, M. J. (2007). Effects on scale linking of different definitions of criterion functions for the IRT characteristic curve methods. *Journal of Educational and Behavioral Statistics* *32*(4), 371-397.

Kim, S. & Lee, W. (2004). *IRT scale linking methods for mixed-format tests*. (ACT Research Report 2004-5). Lowa City, IA: Act, Inc.

Kim, S. & Lee, W. (2006). *IRT scale linking methods for mixed-format tests* (ACT Research Report 2004-5). Lowa City, IA: Act, Inc.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18* (1), 1-11.

Kolen, M. J. (1988). An NCME instructional module on traditional equating methodology. *Educational Measurement: Issues and Practice, 7*(4), 29-36.

Kolen, M. J. (2004). *POLYEQUATE* windows console version [Computer software]. Lowa City IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Lowa.

Kolen, M. J. & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York: Springer.

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

Kolen, M. J. & Brennan, R. L. (2014). *Test equating, scaling and linking: Methods and practices* (3rd ed.). New York: Springer.

Kubiszyn, T. & Borich, G. D. (2013). *Educational testing and measurement: Classroom application and practice* (10th ed.). New Jersey: Wiley.

Lee, W. & Ban, J. (2010). A comparison of IRT linking procedures. *Applied Measurement in Education 23*(1), 23-48.

Li, Y. H., Lissitz R. W. & Yang, Y. N. (1999, April). *Estimating IRT equating coefficients for tests with poltomously and dichotomously scored items*. Paper presented at Annual Meeting of The National Council on Measurement in Education, Montreal, Canada.

Lord F. M. & Wingersky M. S. (1984). Comparison of IRT true-score and equipercentile observed score equatings. *Applied Psychological Measurement, 8*, 452–461.

Lorenzo-Seva, U. & Ferrando, P. J. (2006). *FAKTOR 10.4* [Computer software]. Tarragona: Universitat Rovira i Virgili.

Loyd, B. H. & Hoover, H. D. (1980). Vertical equating using the rasch model. *Journal of Educational Measurement, 17*(3), 179-193.

Marco, G. L. (1977). Item characteristic curve solutions to three intracteble testing problems. *Journal of Educational Measurement, 14*(2), 139-160.

MEB (2013). *PISA 2012 ulusal ön raporu*. Ankara: Sebit.

Muraki, E. (1997). A generalized partial credit model. W.J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153-164) içinde. New York: Springer.

Muraki, E. & Bock, R. D. (2003). *PARSCALE 4.1* [Computer software]. Chicago, IL: Scientific Software International, Inc.

OECD (2009). *PISA Data Analysis Manual: SPSS (Second Edition*). PISA, OECD Publishing, doi: 10.1787/9789264056275-en.

Ogasawara, H. (2000). Asymptotic standard errors of IRT equating coefficients using moments. *Economic Review (Otaru University of Commerce), 51*(1), 1-23.

Ogasawara, H. (2001). Standart errors of item response theory equating / linking by response function methods. *Applied Psychological Measurement, 25* (1), 53- 67.

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

Öztürk-Gübeş, N. & Kelecioğlu, H. (2016). The impact of test dimensionality, common-item set format, and scale linking methods on mixed-format test equating. *Educational Sciences: Theory and Practice*, 16, 715-734.

Petersen, N. S., Kolen, M. J. & Hoover, H. D. (1989). Scaling, norming and equating. R. L. Linn (Ed.), *Educational measurement* (pp. 221-262) içinde. New York: Macmillan.

Sinharay, S. & Hollland, P. W. (2010). A new approach to comparing several equating methods in the context of the NEAT design. *Journal of Educational Measurement, 47*(3), 261-285.

Skaggs, G & Lissitz, R. (1982, March) *Test equating: relevant ıssues and a review of recent research.* Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, California.

Speron, E. (2009). *A comparison of metric linking procedures in item response theory.* Unpublished doctorate dissertation, IIIinois Institute of Technology, Chicago.

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-2010.

Sönmez, V. & Alacapınar, F. G. (2016). *Örneklendirilmiş bilimsel araştırma yöntemleri* (4. Baskı). Ankara: Anı Yayıncılık.

Tanguma, J. (2000, January). *Equating test scores using the linear method: a primer*. Paper presented at the Annual Meeting of the Southwest Educational Research Association, Dallas, Texas.

Tate, R. (2000). Performance of a proposed method for he linking of mixed-format tests with constructed response and multiple choice items. *Journal of Educational Measurement, 37*(4), 329-346.

Tsai, T., Hanson, B. A., Kolen, M. J. & Forsyth, R. A. (2001). A comparison of bootstrap standard errors of IRT equating methods for the common-item nonequivalent groups design. *Applied Measurement in Education*, *14*(1), 17-30, doi: 10.1207/S15324818AME1401_03.

Uysal, İ. (2014). *Madde tepki kuramına dayalı test eşitleme yöntemlerinin karma modeller üzerinde karşılaştırılması.* Yayımlanmamış yüksek lisans tezi, Abant İzzet Baysal Üniversitesi, Bolu.

Yang, W. L. & Houang, R. T. (1996, April). *The effect of anchor length and equating method on the accuracy of test equating comparisons of linear and IRT-based equating using an anchor-item design.* Paper presented at American Educational Research Association, New York, USA.

Yen, W. & Fitzpatrick, A. R. (2006). Item response theory. R. L. Brennan (Ed.), *Educational measurement* içinde (4. Baskı). Westport, CT: Praeger Publishers.

Zhu, W. (1998). Test equating: What, why and how? *Research Quarterly for Exercises and Sport, 69*(1), 11–23.

Extended Abstract

Purpose

One of the problems discussed in the field of measurement is whether scores obtained from two different forms designed to measure the same characteristics are equal or not. While developing a test, it is quite difficult to create a form in parallel to each other in terms of all characteristics. Therefore, forms developed may show some difficulty differences. In a test application carried out in such a case, if the first form is easier than the second form, students receiving the first one may hold a more advantageous position. This situation prevents a fair assessment and may lead to incorrect decisions. Therefore, it is necessary to place scores obtained from different forms into a common scale. Through this process, called as test equating, it becomes possible to directly compare scores taken from different forms.

One of the test applications in which different forms are used at the same time is Program for International Student Assessment (PISA). With this application implemented Organization for Economic Cooperation and Development (OECD) once every three years, it is measured at what level students in the 15-year-old group are able to use the information they learn in mathematics, reading and science fields in problem situations in daily life. With this application, it is also possible to make international comparisons and to assess the effect of educational decisions taken.

Presence of different items in these booklets used in the internationally important application makes it difficult to compare obtained scores. From the point of justness of this comparison, it is necessary to equate students' scores. However, selection of method is critical for test equating. Because selecting wrong method may result in equating errors. Therefore, it is necessary to compare results obtained with different methods. Analysing the field literature, it is seen that scale linking methods are generally compared over simulative data. Simulation studies are valuable in terms of giving an opportunity for researchers to review many factors for test equating studies. However, obtained

results should also be supported with real data. In the literature, there are only a limited number of studies in which scale linking methods are compared over actual data. For this reason, it is considered that this study will contribute to equating studies to be carried out with tests prepared in a mixed format.

In this study, two booklets consisting of mixed items were equated with scale linking methods, and the method with least errors was analysed over real data. Within the scope of this study, answers to the following problems were searched:

In equating scores obtained from two booklets (Booklet-4, Booklet-11) selected from PISA 2012 Mathematics literacy test;

1. How are the equated scores that are obtained from mean-mean method?
2. How are the equated scores that are obtained from mean-sigma method?
3. How are the equated scores that are obtained from Stocking-Lord method?
4. How are the equated scores that are obtained from Haebara method?
5. Which scale are linking and test equating method has the least equating errors?

Method

In this study, two booklets selected from the PISA 2012 Mathematics literacy test were equated with scale linking methods based on item response theory, and error rates obtained from various methods were analysed. Since methods and techniques included in the study were tests over real data, the study is a descriptive research. The research study group consisted of a total of 716 students including 348 students answering Booklet-4 and 368 students answering Booklet-11 in the Turkey sample. Research data consists of answers that students participating in the PISA 2012 Turkey application gave to mathematics literacy test in Booklet-4 and Booklet-11. The reason for selecting these booklets is that they provided unidimensionality assumption required for conducting test equating with item response theory. For test equating, "common-item nonequivalent groups design" was used. This design is one of the most widely used patterns in the literature. According to this design, the participants answer one of the test forms containing common items, and equating relationship was provided over common items. A total of 35 items, 12 of which are common, were included in booklets selected for the study. While three of common items were scored partially (0-1-2), nine items were scored in binary (0-1). Outside common items, there were 23 separate items scored in binary in these booklets.

In analysis of data, primarily descriptive statistics were calculated. Considering distribution of scores taken from both booklets, it was observed that scores showed a distribution close to normal since measures of central tendency were close to each other, and skewness and kurtosis coefficients were within the range of [+1, -1]. After analysing descriptive statistics, unidimensionality assumption of item response theory was tested. In the literature, a dominant factor effecting test performance seems adequate to test unidimensionality assumption. Therefore, exploratory factor analysis was conducted with FAKTOR 10.4 software for both booklets. Since partially scored items were included in research data, factor analysis was carried out by creating polychoric correlation matrix. As a result of exploratory factor analysis, eigen value of the first factor for the booklet-4 was equal to four times of the second factor value. Analysing the booklet-11, it was observed that eigen value of the first factor is greater than four times of that of the second factor.

To test accuracy of unidimensional structure obtained through exploratory factor analysis, confirmatory factor analysis was conducted with LISREL 8.7 program. As a result of the analysis, it was observed that Chi-Square/Degrees of Freedom (x^2/sd) rate was less 3 and Root Mean Square Error of Approximation (RMSEA) was smaller than 0.05. While Normed Fit Index (NFI) value showed good fit, Goodness of Fit (GFI) and Adjusted Goodness of Fit Index (AGFI) values corresponded to weak fit. Besides, Non-normed Fit Index (NNFI) and Comparative Fit Index (CFI) showed perfect fit. Now, analysing Root Mean Square Residual (RMR) and Standardized Root Mean Square Residual (SRMR) value, for both booklets, values closer to 0 were obtained. As a result, analysing fit indices obtained for the unidimensional model, it was observed that the model showed good fit. Therefore, it was determined that unidimensionality assumption was met. Since unidimensionality assumption was met in the study, it was also accepted that local independence assumption was also met.

After analyzing unidimensionality assumption, item and ability parameters were estimated with PARSCALE 4.1 software. In parameter estimation, two-parameter logistics model (2PLM) was used for binary scored items, and generalized partial credit model (GPCM) was used for multiple scored item. Then, scale linking process was carried out for four different methods (mean-mean, mean-sigma, Stocking-Lord, Haebara) with STUIRT software. In the last stage, scores obtained from two booklets were equated with POLYEQUATE software by using IRT true-score equating and IRT observed-score equating methods. To determine the amount of error obtained from different methods, WMSE (Weighted Mean Square Error) coefficient was calculated.

Results

As a result of the study, WMSE coefficients obtained for the true-score equating method were found as 0.0508 in mean-mean method, 0.0717 in mean-sigma method, 0.0298 in Stocking-Lord method and 0.0428 in Haebera methods. When observed-score equating method was used, error rate was obtained as 0.0625 in mean-mean method, 0.0634 in mean-sigma, 0.0453 in Stocking-Lord and 0.0432 in Haebara method. According to the results obtained, it was observed that the least error rate was obtained by Stocking-Lord method when true-score equating method was used and obtained by Haebara method when observed-score equating method was used. The finding was obtained that the maximum error rate was given by mean-sigma in both equating methods. In addition, analysing the results obtained, it was observed that generally IRT true-score equating method gave better results than IRT observed-score equating method.

In the study, finally, results obtained from different methods were compared with DTM (Difference That Matters) coefficient. In this study, "0.5" DTM criterion was used. Standardized DTM value (SDTM) by dividing DTM value to standard deviation of the old form (Booklet-4) was obtained as (0.5/8.24) "0.06". In IRT true-score equating method, it was observed that WMSE values obtained outside mean-sigma method were lower than the SDTM value. In IRT observed-score equating method, on the other hand, WMSE value belonging to equated scored obtained from Stocking-Lord and Haebara method was lower than SDTM value. In light of these findings, it can be suggested to use scale linking methods other than mean-sigma when IRT true-score equating method is selected and to equate scores using Stocking-Lord and Haebera methods when IRT observed-score equating method is used.

Discussion, Conclusions and Suggestions

In this study, scores obtained from two booklets were equated with IRT true-score and observed-score equating methods using different scale linking methods, and results obtained from different methods were compared over real data. As a result of the study, it was found that characteristic curve methods (Stocking Lord, Haebara) gave less amount of error than moment methods (mean-mean, mean-sigma). This finding shows consistency with similar studies in which scale linking methods are compared using various assessment criteria. Analysing results obtained from characteristic curve methods, it was found that Stocking-Lord method gave the least error rate in true-score equating and Haebara method gave the least error in observed-score equating. In some studies conducted in the literature, it has been determined that Stocking-Lord method gave better results, in some studies, Haebara method gave less amount of errors, and in others, similar results

were obtained. According to the results obtained from moment methods, it was observed that the highest error rate was obtained with mean-sigma method.

Besides in the study, results from IRT true-score and IRT observed-score equating methods obtained when using different scale linking methods were analysed. Accordingly, it was found that true-score equating method had a lower level of error rate than observed-score equating method. In studies conducted in the literature, results obtained from both methods vary. The reason for these differences is considered to stem from various factors such as sample size, item type, number of common items... etc. Accordingly, it is possible to say that method with the least error rate may vary based on equating conditions.

In accordance with the results obtained from this study, it can be suggested that characteristic curve methods are used in equating studies to be conducted based on IRT by using common-item nonequivalent groups design in mixed format tests. In future studies, results obtained from concurrent estimation and separate estimation can be compared over real and simulation data.