Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi
Yıl: 2022, Cilt: 55, Sayı: 1, 1-36
DOI: 10.30964/auebfd.1001128, E-ISSN: 2458-8342, P-ISSN: 1301-3718

Madde Tepki Kuramına Dayalı Gerçek Puan Eşitlemede Ölçek Dönüştürme Yöntemlerinin İncelenmesi[1]

MAKALE TÜRÜ	Başvuru Tarihi	Kabul Tarihi	Yayım Tarihi
Araştırma Makalesi	26.09.2021	22.12.2021	05.01.2022

Ömay Çokluk Bökeoğlu [2]
Ankara Üniversitesi
Arzu Uçar [3]
Hakkari Üniversitesi
Ebru Balta [4]
Ağrı İbrahim Çeçen Üniversitesi

Öz

Bu araştırmada, Madde Tepki Kuramı'na (MTK) dayalı gerçek puan eşitlemede, ölçek dönüştürme yöntemlerinin (ortalama-ortalama (OO), ortalama-standart sapma (OS), Stocking-Lord (SL), Haebara (HB)) farklı koşullar altında eşitleme hatalarının karşılaştırılması amaçlanmıştır. Araştırmanın amacı doğrultusunda, yöntemlerin hatalarını karşılaştırmak için örneklem büyüklüğü (500, 1000, 3000, 10000), test uzunluğu (40, 50, 80), ortak madde oranı (%20-%30-%40), parametre kestirim modeli (iki ve üç parametreli lojistik model (2PLM ve 3PLM)) ve grupların yetenek dağılımı (benzer (N(0-1) - N(0-1)), farklı (N(0-1) - N(0.5,1)) koşulları altında 2PLM ve 3PLM'ye uyumlu iki kategorili 50 yineleme ile 7200 veri seti oluşturulmuştur. Veri toplamı deseni olarak "denk olmayan gruplarda ortak madde/test (NEAT) deseni" kullanılmıştır. Veri üretiminde ve analizinde R yazılımı kullanılmıştır. Araştırmadan elde edilen bulgular, eşitleme hatası (RMSD) ölçütüne göre değerlendirilmiştir. Çalışmanın sonucunda, tüm koşullar göz önünde bulundurulduğunda, SL yönteminin RMSD değerlerinin, diğer yöntemlere göre daha yüksek olduğu görülmekle birlikte, OO ve OS yöntemlerinin birbirine benzer RMSD değerleri ürettiği görülmüştür. Ayrıca, ölçek dönüştürme yöntemlerine ilişkin RMSD değerleri karşılaştırıldığında, 2PLM ve 3PLM'nin kullanıldığı durumlarda benzer sonuçlar elde edildiği, örneklem büyüklüğü ve test uzunluğu arttıkça SL yöntemi dışında diğer yöntemlerin eşitleme hatalarında azalma oluştuğu ve ortak madde oranının %40 ve grupların yetenek dağılımının benzer olduğu durumlarda, yöntemlerin, RMSD değerlerinin daha düşük olduğu gözlenmiştir.

Anahtar sözcükler: Haebara, MTK gerçek puan eşitleme ortalama-ortalama, ortalama-standart sapma, ölçek dönüştürme, Stocking-Lord, test eşitleme.

Etik kurul kararı: Bu araştırma, 01.01.2020 tarihinden önce simülatif verilerle gerçekleştirildiği için etik kurul kararı zorunluluğu taşımamaktadır.

[1]Bu araştırma makalesi 02-05 Mayıs 2018 tarih aralığında Akdeniz Üniversitesi'nde düzenlenen V^{th} International Eurasian Educational Research Congress'de sözlü bildiri olarak sunulmuştur.

[2]Prof. Dr., Eğitim Bilimleri Fakültesi, Eğitim Bilimleri Bölümü, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, e-posta: cokluk@education.ankara.edu.tr, https://orcid.org/ 0000-0002-3879-9204
[3]Dr. Öğr. Üyesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, e-posta: arzukapcik@gmail.com, https://orcid.org/ 0000-0002-0099-1348
[4]*Sorumlu Yazar*: Arş. Gör. Dr., Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, e-posta: ebrubalta2@gmail.com, https://orcid.org/ 0000-0002-2173-7189

Eğitim ve psikoloji alanındaki testler, bireylerin öğrenme eksiklerinin ortaya çıkarılması, okulların ya da bireylerin profilinin oluşturulması, bireylerin işe alınması ya da bir kuruma seçme ve yerleştirme vb. amaçlarla kullanılmaktadır. Kullanılan testler ile elde edilen ölçme puanları doğrultusunda bireyler hakkında yaşamlarının şekillenmesinde etkin rol oynayan kararlar alınmaktadır. Bireyler hakkında uygun kararların verilebilmesi için bu testlerden elde edilen puanların geçerli ve güvenilir olması ve böylelikle de yanlı kararlara yol açmaması beklenmektedir. Bu amaçlarla da, bireyler hakkında yansız kararlar alınmaya çalışılırken farklı bireylerin, farklı testlerle aynı özelliklerinin ölçülmesi gereken durumlar oluşabilmektedir (Crocker ve Algina, 1986; Petersen ve diğ., 1993). Seçme ve yerleştirme, sertifika verme, geçme-kalma vb. durumlarda kararlar vermede kullanılan testler, test geliştiriciler tarafından içerik, bilgi ve beceri açısından benzer şekilde hazırlanmaya çalışılsa da, test formları arasında testin güçlük düzeyi ve başka bazı istatistiksel özellikler açısından farklılıklar oluşabilmektedir (Hambleton ve diğ., 1991; He, 2010; Kolen, 1988; Kolen ve Brennan, 2004). Farklı test formları kullanıldığında bireyler hakkında adil kararların verilebilmesi, bir başka deyişle bu formlardan elde edilen test puanlarının karşılaştırılabilir olması "test eşitleme" çalışmalarının yapılması ile olanaklı olmaktadır (Dorans, 1990; Mohandas, 1996; Skaggs, 1990).

Test eşitleme, ölçülen özelliğin aynı olduğu test uygulamalarından elde edilen puanların eşdeğerliğini sınayan istatistiksel bir süreç olarak tanımlanmaktadır (Kolen ve Brennan, 2004; Kim ve Hanson, 2002). Test eşitlemenin gerçekleştirilebilmesi için; eşitlenecek her iki testin aynı yapıyı ölçmesi, eşitlik özelliğinin (testler eşitlendiğinde bireylerin farklı testleri almasının fark oluşturmaması) bulunması, simetri özelliğinin bulunması (X formu Y formuna eşitlenebiliyorsa, Y formunun da X formuna eşitlenebilmesi), gruplararası değişmezlik özelliğinin olması (eşitleme fonksiyonunun örneklem seçimine bağlı olmaması) ve testlerden elde edilen puanların eşit güvenirlik düzeyine sahip olması koşullarının karşılanması gerekmektedir (Dorans ve Holland, 2000). Test eşitleme çalışması, veri toplama için gerekli olan test eşitleme deseninin seçilmesi, test formlarının eşitleneceği eşitleme yöntemlerinin seçilmesi ve test eşitleme sonuçlarının değerlendirilmesi olmak üzere üç süreçten oluşmaktadır (Harris ve Crouse, 1993). Test formlarının eşitlenmesinde ilk aşamada, test eşitleme deseninin oluşturulması gerekmektedir. Test eşitleme desenleri, testin içeriği ve gruba bağlı olarak değişkenlik gösteren yapılardır (Holland ve diğ., 2007). Desenlerin oluşturulmasında, bireylerin eşit yetenek dağılımları ve ortak madde kullanımı yaklaşımları kullanılmaktadır (Dorans ve diğ., 2010). Alanyazın incelendiğinde, veri toplama desenlerine ilişkin çeşitli sınıflamaların yer aldığı görülmektedir (Hambleton ve diğ., 1991; Kolen ve Brennan, 2004). Kolen ve Brennan (2004), veri toplama desenlerini, tek grup, dengelenmiş tek grup ve denk olmayan gruplarda ortak madde deseni şeklinde sınıflandırmıştır. Tek grup deseninde, eşitlemenin yapılacağı testler aynı bireylere uygulanarak elde edilen verilerin istatistiksel açıdan daha güçlü olması sağlanabilmektedir. Bu desenin sınırlılığı (dezavantajı) ise iki test formunun aynı bireylere uygulanmasından kaynaklı olarak iki test uygulamasının sonuçlarının birbirini etkilemesi olarak görülmektedir (von

Davier, 2010). Tek grup desenindeki öğrenme ve yorgunluk etkisini (sıra etkisi) ortadan kaldırmaya yönelik olarak dengelenmiş tek grup deseni geliştirilmiştir. Dengelenmiş tek grup deseninde, yetenek dağılımları eşdeğer olan bireyler, random olarak iki farklı gruba atanarak, gruplara her iki test formu sırayla uygulanmaktadır (Kolen ve Brennan, 2004; Petersen ve diğ., 1993). Denk olmayan gruplar ortak madde/test deseninde (NEAT), her bir öğrencinin tek bir test formunu alması sağlanarak ortak maddelere verilen tepkiler üzerinden eşitleme algoritması kurulmaktadır (Kolen ve Brennan, 2004). Test eşitleme çalışmalarının en önemli aşaması olan veri toplama işleminin ardından, test formlarını eşitleme yönteminin seçilmesi gerekmektedir (Holland ve diğ., 2007).

Test eşitleme işlemleri, psikometrik kuramlardan Klasik Test Kuramı (KTK) ve Madde Tepki Kuramı'na (MTK) dayalı olarak gerçekleştirilmektedir. KTK'de madde güçlük ve madde ayırtedicilik indekslerinin "gruba bağlı" olmasından dolayı eşitlik ve değişmezlik varsayımları karşılanamamaktadır (Kolen, 1981). MTK'de ise, bireylerin yetenek düzeyleri maddeden, madde parametreleri de gruptan bağımsız olarak belirlenebilmektedir. Böylelikle, KTK ve MTK'nin dayandığı varsayımlar ve matematiksel fonksiyonların birbirinden farklılaşmasından dolayı kullanılan test eşitleme yöntemleri de farklılaşmaktadır. KTK'ye dayalı yöntemler; eşit yüzdelikli (equipercentile), doğrusal (linear) ve ortalama (mean) eşitleme olmak üzere üçe ayrılmaktadır (Barnard, 1996). MTK'ya dayalı yöntemler, madde karakteristik eğrilerinin gösterildiği matematiksel fonksiyonlara göre farklılaşmaktadır. MTK'de değişmezlik özellliği bulunmasından kaynaklı olarak eşitleme işlemi yapılmasına gerek duyulmamakla birlikte veri toplama desenine göre ölçekleme işlemi gerçekleştirilmektedir. Dengelenmiş veya tek grup deseni kullanılarak gerçekleştirilen eşitleme işlemlerinde parametrelerin aynı ölçeğe yerleştirilmesinde ek bir ölçeklemeye gerek duyulmamakla birlikte denk olmayan gruplarda doğrusal bir dönüştürme yapılma gereği bulunmaktadır (Han, 2008; Kolen ve Brennan, 1995). Parametre dönüştürme yolu ile parametrelerin ortak bir ölçeğe yerleştirilme süreci "kalibrasyon" olarak adlandırılmaktadır (Kolen ve Brennan, 2004). Kalibrasyon süreci boyunca, madde ve yetenek parametreleri aynı ölçekte kestirilerek farklı formlardan elde edilen test puanlarının karşılaştırılması sağlanmaktadır (Crocker ve Algina, 1986). MTK'da farklı gruplardan elde edilen madde parametrelerini eşitleme ya da ölçek dönüşümü için "eş zamanlı kalibrasyon (concurrent calibration)" ve "ayrı kalibrasyon (separate calibration)" olmak üzere iki farklı kalibrasyon yöntemi kullanılmaktadır (Kolen ve Brennan, 1995; Nozawa, 2008). Eş zamanlı kalibrasyon yöntemleri, birden fazla testin eşitlendiği ve ortak maddeler için madde parametrelerinin önceden bilinmediği durumda kullanılmaktadır (Bastari, 2000). Eş zamanlı kestirimde, kalibrasyon işlemi tek aşamada gerçekleştirilerek her iki test formundaki tüm maddeler için madde ve yetenek parametreleri aynı anda tek bir analiz ile aynı anda kestirilmektedir (Chu, 2002; Hanson ve Beguin, 2002).

Ayrı kalibrasyon ölçek dönüştürme yöntemlerinde, denk olmayan gruplardan elde edilen madde parametrelerini aynı ölçeğe yerleştirme işleminde, ortak madde deseninde yer alan ortak maddelerin parametreleri kestirilerek gruplarda yer alan

bireylerin yetenek parametreleri ortalaması 0, standart sapması 1 olacak şekilde ölçeklenir (Li, 2009). Böylelikle, ayrı kalibrasyon yöntemlerinde, iki farklı test formu için madde ve yetenek parametreleri kestirildikten sonra kestirilen parametreler için lineer bir dönüşüm işlemi yapılarak kalibrasyon işleminin iki aşamalı gerçekleştirildiği görülmektedir (Cook ve Eignor, 1991; Hanson ve Beguin, 2002; Kim ve Cohen, 1998). Ortak madde deseninde MTK'ye dayalı ayrı kalibrasyon ölçek dönüştürme yöntemleri; moment yöntemleri (ortalama-ortalama yöntemi (Loyd ve Hoover, 1980) ve ortalama-standart sapma yöntemi (Marco, 1977)) ve karakteristik eğri dönüştürme yöntemleri (Haebara, 1980; Stocking ve Lord, 1983) olarak iki kategoride ele alınmaktadır. Bu araştırma kapsamında, moment yöntemlerinden, ortalama-ortalama ve ortalama-standart sapma ve karakteristik eğri dönüştürme yöntemlerinden Haebara ve Stocking-Lord ölçek dönüştürme yöntemleri kullanıldığı için, aşağıda, bu yöntemler hakkında kısa bilgiler sunulmuştur.

Ortalama-ortalama (OO) ölçek dönüştürme yönteminde, eşitleme katsayılarını kestirmede ortak maddelerin ayırt edicilik ve güçlük parametre değerlerinin ortalamaları hesaplanır (Loyd ve Hoover, 1980). A ve B eşitleme katsayılarını hesaplamada kullanılan matematiksel gösterimler Eşitlik 1 ve Eşitlik 2 ile belirtilmektedir.

$$A = \frac{\mu(\alpha_i)}{\mu(\alpha_j)} \tag{1}$$

$$B = \mu(b_j) - A\mu(b_i) \tag{2}$$

Eşitlik 1 ve Eşitlik 2'de, $\mu(\alpha_i)$, i ölçeğine ait ortak maddelerin ayırıcılık parametrelerinin ortalamasını, $\mu(b_i)$, i ölçeğine ait ortak maddelerin güçlük parametrelerinin ortalamasını, $\mu(\alpha_j)$, j ölçeğine ait ortak maddelerin ayırtedicilik parametrelerinin ortalamasını, $\mu(b_j)$, j ölçeğine ait ortak maddelerin güçlük parametrelerinin ortalamasını, A, eşitleme denkleminin eğimini ve B eşitleme denkleminin sabitini göstermektedir. Ortalama-standart sapma (OS) ölçek dönüştürme yönteminde ise, eşitleme katsayılarını elde edebilmek için ortak maddelerin güçlük parametre değerlerinin ortalaması ve standart sapması kullanılır (Marco, 1977). A ve B eşitleme katsayılarını elde etmede kullanılan matematiksel gösterimler Eşitlik 3 ve Eşitlik 4 ile belirtilmektedir.

$$A = \frac{\sigma(b_j)}{\sigma(b_i)} \tag{3}$$

$$B = \mu(b_j) - A\mu(b_i) \tag{4}$$

Eşitlik 3 ve Eşitlik 4'te, $\sigma(b_i)$, i ölçeğine ait ortak maddelerin güçlük parametre değerlerinin standart sapmasını, $\mu(b_i)$, i ölçeğine ait ortak maddelerin güçlük parametre değerlerinin ortalamasını, $\sigma(b_j)$, j ölçeğine ait ortak maddelerin güçlük parametre değerlerinin standart sapmasını, $\mu(b_j)$, j ölçeğine ait ortak maddelerin güçlük parametre değerlerinin ortalamasını, A, eşitleme denklemindeki eğimi ve B eşitleme denklemindeki sabiti göstermektedir. OO ve OS yöntemlerinden A ve B katsayılarının kestirim eşitlikleri incelendiğinde farklı parametreler kullanıldığı görülmektedir. Haebara (1980) ile Stocking ve Lord (1983), ölçek dönüştürmede moment yöntemlerinin tüm madde parametre kestirimlerini eş zamanlı hesaba katmamasından kaynaklı olarak karakteristik eğrileri benzer fakat parametre değerleri farklı olan maddeler için hatalı sonuç üretmesinin önüne geçebilmek için, madde güçlük ve madde ayırt edicilik parametrelerinin her ikisini de dikkate alan yöntemler geliştirmişlerdir. Karakteristik eğri yöntemlerinde dönüşüm sabitleri belirlenerek test ya da madde karakteristik eğrileri arasındaki fark hesaplanarak gerçek puanlar eşitlenir (Baker ve Al-Karni, 1991; Hambleton ve diğ., 1991). NEAT deseniyle toplanan verilerde uygulanan ve ortak maddelerin karakteristik eğrileri arasındaki farkı azaltmak için geliştirilen Stocking ve Lord (1983) yöntemine ilişkin matematiksel gösterim Eşitlik 5 ile belirtilmektedir.

$$SLdiff(\theta_i) = \left[\sum_{j:V} p_{ij}(\theta_{ji}, a_{ji}, b_{ji}, c_{ji}) - p_{ij}\left(\theta_{ji}; \frac{a_{ij}}{A}, Ab_{ij} + B, c_{ij}\right)\right]^2,$$

$$SLcritt = \sum_i SLdiff(\theta_i) \qquad (5)$$

Eşitlik 5'te, $p_{ij}(\theta_{ji}, a_{ji}, b_{ji}, c_{ji})$, madde karakteristik fonksiyonunu ve $p_{ij}\left(\theta_{ji}; \frac{a_{ij}}{A}, Ab_{ij} + B, c_{ij}\right)$ eşitlenmiş madde karakteristik fonksiyonunu göstermektedir. Eşitlik 5 incelendiğinde, Stocking-Lord (SL) yaklaşımının, belli bir yetenek düzeyinde, madde karakteristik eğrileri arasındaki farkı, her bir madde karakteristik eğrisi için madde karakteristik fonksiyonu ile eşitlenmiş madde karakteristik fonksiyonu arasındaki farkın toplamının karesini alarak hesapladığı görülmektedir. Haebara (1980)'nın geliştirdiği madde karakteristik eğri yöntemine ilişkin matematiksel gösterim Eşitlik 6 ile belirtilmiştir.

$$Hdiff(\theta_i) = \sum_{j:V} \left[p_{ij}(\theta_{ji}, a_{ji}, b_{ji}, c_{ji}) - p_{ij}\left(\theta_{ji}; \frac{a_{ij}}{A}, Ab_{ij} + B, c_{ij}\right)\right]^2,$$

$$Hcritt = \sum_i Hdiff(\theta_i) \qquad (6)$$

Eşitlik 6 incelendiğinde, Haebara (HB) yaklaşımının, belli bir yetenek düzeyinde madde karakteristik eğrileri arasındaki farkı, her bir madde karakteristik eğrisi için madde karakteristik fonksiyonu ile eşitlenmiş madde karakteristik fonksiyonu arasındaki farkın karesini toplayarak hesapladığı görülmektedir.

Eşitleme sürecinde, ölçek dönüştürme işlemi gerçekleştirildikten sonra aynı ölçekte yer alan parametreler için MTK gerçek puan ve gözlenen puan eşitleme işlemi uygulanır. MTK gerçek puan ve MTK gözlenen puan eşitleme yöntemleri X ve Y formuna ait puanları ilişkilendirmek için kullanılır (Kolen, 2007). MTK gözlenen puan eşitlemede, madde ve yetenek parametreleri kestirilerek X ve Y formu için kuramsal (teorik) olarak oluşturulan gözlenen puan dağılımlarından elde edilen puanlar doğrusal ve eşit yüzdelikli eşitleme yöntemi kullanılarak eşitlenmektedir (Kolen ve Brennan, 2004; von Davier, 2008). MTK gerçek puan eşitleme yönteminde ise gerçek puanlar farklı yetenek düzeylerine göre kestirilerek X formu üzerinde seçilen gerçek puana karşılık gelen yetenek düzeyi belirlenir. Ardından, belirlenen bu yetenek düzeyine ilişkin Y formunda karşılık gelen gerçek puan hesaplanır (Hambleton ve Swaminathan, 1985; Kolen ve Brennan, 2004). MTK gerçek puan eşitlemede her iki gruba ilişkin yetenek dağılımlarının kestirilmesine gerek kalmamakta ve eşitleme sonuçları kuramsal olarak gruptan bağımsız elde edilmektedir (Han ve diğ., 1997).

Test eşitleme sürecinde, eşitlemenin doğruluğunun belirlenmesinde random ve sistematik eşitleme hataları kullanılmaktadır (Felan, 2002; Kolen, 1988; Ryan ve Brockmann, 2009). Random eşitleme hatası, cevaplayıcı örnekleminden kaynaklanan bir hata türü olarak; sistematik hata ise eşitlemenin varsayımlarının ihlal edilmesi ve yanlılıktan kaynaklanan hata türü olarak tanımlanmaktadır. Random eşitleme hataları, örneklem büyüklüğü arttıkça azalmaktadır. Sistematik hata ise, özellikle eşitlenmemiş gruplarda ortak madde deseninde, grupların birbirinden farklılaştığı, ortak maddelerin içerik ve istatistiksel olarak testi temsil etmediği vb. durumlarda ayrıca ortak maddelerin testteki konumundan kaynaklı olarak ortak madde fonksiyonunun farklılaştığı durumlarda artmaktadır (Kolen, 1988). Test eşitleme sürecinde toplam hata, random ve sistematik hatanın toplamı olarak tanımlanır. Eşitleme sürecinde random ve sistematik hata türü kontrol altına alınmaya çalışılmalıdır (Kolen ve Brennan, 2004; Zeng, 1991). Test eşitleme sürecinin kontrolünde, eşitleme hatalarını en aza indirebilmek için eşitleme deseninin ve yönteminin, eşitlemenin amacına, yanıtlayanların özelliklerine ve test verisine bağlı olarak seçilmesi gerekmektedir (Yang ve Houang, 1996).

Özellikle yılda birden çok kez uygulanan veya sınav güvenliği nedeniyle paralel olduğu varsayılan farklı test formlarının kullanıldığı ulusal düzeyde gerçekleştirilen geniş ölçekli ve yüksek riskli sınav uygulamalarında (ALES, KPSS, LGS, YGS, vb.) ve uluslararası karşılaştırmalar yapmak için kullanılan testlerde de (TIMSS, PISA, PIRLS), test puanlarının eşitliğinin sınanması her geçen yıl daha da önem kazanmaktadır. İki ya da daha fazla testin aynı içerikte ve benzer istatistiksel özelliklere sahip olacak şekilde ortak bir ölçek üzerine yerleştirilmesi ile farklı madde özelliklerine sahip veriler birleştirilerek bireylere ilişkin gelişimler ve eğilimler belirlenmekte ve böylelikle bölgeler ile ülkeler arasındaki farklılıkların ortaya çıkarılmasında bilgi verici sonuçların ortaya çıkarılması sağlanmaktadır. Böylece, test geliştiriciler ve uygulayıcılar açısından bireyler hakkında karar vermede adil davranılmasına katkıda bulunulmaktadır (Angoff, 1971; Eid, 2005; Felan, 2002; Michaelides, 2003; von Davier ve Wilson, 2007). Test geliştirici ve uygulayıcılarına

katkı sağlamada, test puanlarının eşitlenmesinde, olabildiğince hatadan arınık eşitleme çalışmalarının gerçekleştirilmesi önem taşımaktadır. MTK gerçek puan ve MTK gözlenen puan eşitleme yöntemlerinin karşılaştırıldığı çalışmalarda, bazı araştırmacılar (Hagge, 2010; Tsai ve diğ., 2001) gözlenen puan eşitlemenin daha düşük hata oranına sahip olduğunu, bazıları (Han ve diğ., 1997; Cho, 2007) gerçek puan eşitlemenin daha düşük hata oranına sahip olduğunu, diğerleri (Brossman ve Lee, 2013; Lord ve Wingersky, 1984; Wang, 2012) ise gözlenen puan ve gerçek puan eşitleme yöntemlerinin benzer sonuçlar verdiğini belirtmişlerdir. Cook ve Eignor (1991) çalışmasında, yetenek dağılımları farklı olan gruplara farklı güçlük düzeyinde testlerin uygulandığı durumlarda, MTK gerçek puan eşitleme yöntemlerinin KTK'ye dayalı eşitleme yöntemleri ve MTK gözlenen puan eşitleme yöntemlerine kıyasla daha az hata oranına sahip olduğunu belirtmiştir. Bununla birlikte, MTK'ye ve KTK'ye dayalı test eşitleme yöntemlerini karşılaştıran bazı çalışmalar (Caldwell, 1984; Chen, 2001; Han ve diğ., 1997; Petersen ve diğ., 1983; Yang, 1997; Yang ve Houang, 1996), MTK'ye dayalı eşitleme yöntemlerinin daha kararlı sonuçlar verdiğini belirtirken, bazı çalışmaların (Harris ve Kolen, 1986; Hills ve diğ., 1988; von Davier ve Wilson, 2007) MTK'ye ve KTK'ye dayalı test eşitleme yöntemlerinin benzer sonuçlar verdiğini belirtmektedir. Test eşitleme sürecinde, karakteristik eğri ve moment ölçek dönüştürme yöntemlerinin ele alındığı ilgili çalışmalar incelendiğinde (Baker ve Al-Karni, 1991; Gök ve Kelecioğlu, 2014; Gül ve diğ., 2017; Hanson ve Béguin, 2002; Karkee ve Wright, 2004; Kim ve Cohen, 1998; Kolen ve Brennan, 2004; Kilmen, 2010; Kim ve Lee, 2004; Kim ve Kolen, 2006; Meng, 2012; Way ve Tang, 1991) karakteristik eğri yöntemleri ile daha kararlı sonuçlar elde edildiği görülmektedir. Tate (2000), çok kategorili puanlanan maddeler için uyarlanan OS ve SL yöntemlerinden elde edilen eşitleme hatalarının birbirine benzer olduğunu, Speron (2009), ise bazı koşullarda OS yönteminin en fazla hataya sahip olduğunu belirtmiştir. Karakteristik eğrisi yöntemleri karşılaştırıldığında, HB ve SL yöntemlerinin benzer eşitleme hatasına sahip olduğu belirtilmiştir (He, 2011; Lee ve Fitzpatrick, 2008). Moment yöntemleri karşılaştırıldığında, OO ve OS yöntemlerinin, kestirim için farklı parametreler kullanmasından kaynaklı olarak farklı eşitleme hatalarına sahip sonuçlar ürettiği görülmektedir. Aksekioğlu (2017) ile Kolen ve Brennan (2004), OS yönteminin, madde güçlük parametresinin kestiriminin madde ayırtedicilik parametresinin kestirimine göre daha kararlı olmasından dolayı OO yöntemine tercih edilebilir olduğunu, Baker ve Al-Karni (1991), Way ve Tang (1991), Ogasawara (2000) ve Gündüz (2015), OO yönteminin ortalamaların standart sapmalardan daha kararlı sonuçlar verdiği için OS yöntemine göre daha kararlı olduğunu belirtmişlerdir. Ayrıca, moment ve karakteristik eğri dönüştürme yöntemleri ile ilgili araştırmalarda, test uzunluğu, örneklem büyüklüğü, parametre kestiriminde kullanılan model türü, kullanılacak eşitleme deseni, ortak madde türü ve yetenek dağılımı değişkenlerinin farklı düzeylerinin bağımsız değişken olarak ele alındığı görülmektedir. Böylece, ölçek dönüştürme ve test eşitleme yöntemlerinden elde edilen sonuçların, eşitleme koşulları değiştikçe farklılık gösterdiği bazı durumlarda ise farklı yöntemlerin benzer sonuçlar verdiği görülmektedir. Böylelikle, araştırmalarda ele alınan koşullar farklılaştığında, ölçek dönüştürme yöntemlerinden

hangisinin tercih edilebileceğinin değerlendirilmesi gerektiği düşünülmektedir. Bu noktadan hareketle bu araştırmada, testlerden elde edilen gerçek puanların eşitlenmesinde, MTK'ye dayalı ölçek dönüştürme yöntemlerinin (ortalama-ortalama, ortalama-standart sapma, Stocking-Lord, Haebara) çeşitli faktörlere (örneklem büyüklüğü, test uzunluğu, grupların yetenek dağılımı, ortak madde oranı, parametre kestiriminde kullanılan model türü) göre eşitleme hatalarının karşılaştırılması amaçlanmıştır. Alanyazın incelendiğinde, araştırma ele alınan koşulların düzeylerinin hepsinin ele alındığı çalışmaya rastlanmamıştır. Son yıllarda özellikle MTK'ye dayalı testlerin geliştirilmesi ve kullanımı da göz önünde bulundurulduğunda, araştırmadan elde edilen sonuçların, test puanlarının eşitlenmesine yönelik olarak test geliştiricilere ve uygulayıcılara katkı sağlayacağı düşünülmektedir.

Yöntem

Bu bölümde araştırma modeli, simülasyon koşulları ve gerçekleştirilen analizlerle ilgili bilgiler yar almaktadır.

Araştırma Modeli

Bu araştırmada, simülasyon verileri kullanılarak testlerden elde edilen gerçek puanların eşitlenmesinde, MTK'ye dayalı ölçek dönüştürme (ortalama-ortalama, ortalama-standart sapma, Stocking-Lord, Haebara) yöntemlerinin çeşitli faktörlere (örneklem büyüklüğü, test uzunluğu, grupların yetenek dağılımı, ortak madde oranı, parametre kestiriminde kullanılan model türü) göre eşitleme hataları karşılaştırılmıştır. Araştırmada, ölçek dönüştürme yöntemlerinin, çeşitli koşullara göre eşitleme hatalarını belirlemek için Monte-Carlo simülasyon çalışması gerçekleştirilmiştir. Araştırmada, MTK'ye dayalı ölçek dönüştürme yöntemlerinin gerçek puan eşitlemede eşitleme hatalarının çeşitli koşullara göre ayrıntılı incelenerek ve karşılaştırılarak ortaya çıkarılması yönüyle araştırmanın betimsel araştırma özelliği taşıdığı söylenebilir. Betimsel araştırmalar, verilen bir durumu eksiksiz ve dikkatli bir şekilde tanımlandığı çalışmalardır (Fraenkel ve diğ., 2012). Bu çerçevede bu araştırma, betimsel araştırma niteliğinde bir simülasyon çalışması olarak değerlendirilebilir.

Veri Toplama Deseni

Bu araştırmada, test puanlarının MTK'ye dayalı ölçek dönüştürme işlemlerinde, denk olmayan gruplarda ortak madde/test deseni (NEAT) kullanılmıştır. NEAT deseni, esnek ve karmaşık bir desen olması ile birlikte alanyazında sıklıkla kullanılmaktadır (Bastari, 2000; Felan, 2002; Kolen ve Brennan, 2004; Meng, 2012; Michaelides, 2003; Sinharay ve Holland, 2008). Araştırmada kullanılan veri toplama deseni, Tablo 1'de yer almaktadır.

Tablo 1

Araştırmanın Veri Toplama Deseni

Örneklem	X Formu	Ortak Madde(A)	Y Formu
Grup 1	+	+	
Grup 2		+	+

Tablo 1'de gösterilen bu desende, test formları (X ve Y) ve ortak test formları oluşturulmuştur. Bu araştırmada X formu eski test formunu, Y formu ise eşitlenecek (yeni) test formunu belirtmektedir. X formunu alan grup Grup 1, Y formunu alan grup ise Grup 2 ile temsil edilmektedir. Gruplardaki birey sayısı ve ortak testte yer alan madde sayısı, araştırmanın amacı doğrultusunda farklılaşmaktadır. NEAT deseninde, gruplar, ortak maddelerden oluşan iki test formundan birini yanıtladıktan sonra ortak maddelerden elde edilen parametreler aynı ölçeğe yerleştirilmektedir.

Simülasyon Deseni

Araştırmada, çeşitli koşulların, gerçek puan eşitlemede, MTK'ye dayalı ölçek dönüştürme yöntemleri üzerine etkisini belirlemek için Monte-Carlo simülasyon çalışması gerçekleştirilmiştir. Araştırmada, gerçek veri ile çalışmada ele alınan koşulların (simülasyon faktörleri; örneklem büyüklüğü, test uzunluğu, ortak madde oranı, parametre kestiriminde kullanılan model türü, grupların yetenek dağılımı ve ölçek dönüştürme yöntemleri) tümünün sağlanmasının olanaklı olmamasından dolayı simülasyon verisi kullanılmıştır.

Ölçek dönüştürme yöntemlerinde, her bir grup için örneklem büyüklüğünün, eşitleme sonuçlarının doğruluğunu ve kesinliğini arttırdığı çeşitli araştırmalar (Cui ve Kolen, 2008; Godfrey, 2007; Hanson ve Beguin, 2002; Kim ve Lee, 2004; Meng, 2012; Norman-Dvorak, 2009; Nozawa, 2008; Qu, 2007; Zhao, 2008) ile ortaya konmuştur. Kolen ve Brennan (2004), MTK'ye dayalı ölçek dönüştürme yöntemlerinde, NEAT deseninde üç parametreli lojistik model'de (3PLM) her bir form için örneklem büyüklüğünün 1500 olması gerektiğini, Skaggs ve Lissitz (1986), 3PLM için örneklem büyüklüğünün en az 1000 olması gerektiğini belirtmiştir. Bastari (2000) ve Han (2008), 1000 örneklem büyüklüğünü, orta büyüklükte, Lee (2007) ile Liou, Cheng ve Johnson (1997) ise büyük örneklem olarak değerlendirmişlerdir. Hanson ve Beguin (2002) ile Gök ve Kelecioğlu (2004) çalışmalarında, 1000 ve 3000 örneklem büyüklüğünü ele alarak sırasıyla " küçük" ve " büyük" örneklem büyüklüğü olarak belirtmişlerdir. Spence (1996) çalışmasında, test eşitleme çalışmalarında en az 500 örneklem büyüklüğü ile çalışılmasını önermiştir. Kim ve Cohen (1998), 300 ve 1.000, Kilmen (2010), 500 ve 1.000, Lee ve Ban (2009), 500 ve 3000, Speron (2009), 250 ve 1.000 örneklem büyüklükleri ile eşitleme hatalarını incelediği çalışmalarında, örneklem büyüklüğü arttıkça eşitleme hatalarının azaldığını belirtmişlerdir. Karkee ve Wright (2004) çalışmasında, 31.813 bireye ilişkin test verilerini kullanarak MTK'ye dayalı ölçek dönüştürme yöntemlerinin hatalarını karşılaştırmıştır. Drasgow ve diğerleri (1995) çalışmasında, 3PLM için 3.000'den büyük örneklem büyüklüklerinin kullanılmasını önermiştir. Böylelikle, bu çalışmada, Türkiye'de MEB ve ÖSYM kurumları tarafından gerçekleştirilen geniş ölçekli sınav uygulamalarında minimum örneklem büyüklüğü de göz önünde bulundurularak, 500, 1.000, 3.000 ve 10.000 olarak belirlenmiştir. Ayrıca, her iki test formu için örneklem büyüklüğü eşit olarak ele alınmıştır.

Test uzunluğu, ölçek dönüşümlerini ve testleri eşitlemeyi etkileyen faktörlerden biridir. Testte yer alan madde sayısının değişimlendiği çalışmalarda, Skaggs (2005), 50 maddeden, Livingston ve Kim (2010), 110 maddeden, Babcock ve diğerleri (2012), 100 maddeden oluşan test formlarını kullanmışlardır. Ayrıca, Gök ve Kelecioğlu

(2014), 30, 60 ve 80 maddeden oluşan test formlarını kullanmış ve test uzunluğunun, test eşitlemenin doğruluğuna etkisini incelemiştir. Araştırmada, test uzunluğu, ülkemizde uygulanan geniş ölçekli sınavlarda yer alan ortalama madde sayıları da dikkate alınarak her bir formda 40, 50 ve 80 madde olacak şekilde değişimlenmiştir.

Ortak maddelerin kullanıldığı veri toplama desenlerinde ölçek dönüştürme işlemleri ortak maddeler aracılığıyla gerçekleşmektedir. NEAT deseninde ortak (ankor) test formu iç ve dış ortak test olmak üzere ikiye ayrılır. Bu çalışmada iç ankor test kullanılmıştır. Ortak maddelerden oluşan ankor test, ölçek dönüşümleri yapılacak testlerin içeriğini kapsayacak şekilde olmalı ve test maddeleri ile yüksek korelasyon göstermelidir. Angoff (1984) ve Budescu (1985), test formlarında yer alan ortak madde sayısının 20'den ya da toplam madde sayısının %20'sinden az olmamasını, Hambleton ve diğerleri (1991), testteki madde sayısının %20 - %25'i kadar ortak madde olması gerektiğini, Kolen ve Brennan (2004), 40 veya daha fazla maddeden oluşan karma modeldeki testlerde ortak maddelerin oranının %20 ve üzerinde olmasını, Kang ve Petersen (2009), ortak madde oranının %20'nin altında olduğu durumların incelenmesini, Spence (1996), testteki madde sayısının %20'si ile %30'u arasında değişen oranda ortak madde yer alması gerektiğini belirtmişlerdir. Godfrey (2007), Kim ve Cohen (1998), Meng (2012) ve Norman-Dvorak (2009) çalışmalarında, ortak madde oranının artması ile birlikte eşitleme hatasının azaldığını belirtmişlerdir. Böylece, simülasyon deseninde test uzunluğu faktör düzeyleri ele alınırken ortak madde oranı da göz önünde bulundurulmuştur. Böylece, 20 maddelik testlerin tek boyutluluğu sağlamasına karşın ortak madde sayısı ve oranı da göz önünde bulundurularak belirlenen desen için 20 maddelik testi kullanmanın uygun olmayacağı görülmektedir. Bu nedenle, bu çalışmada, veri toplama deseninde test formlarında yer alan ortak madde oranı %20, %30 ve %40 olarak belirlenmiş ve ortak madde oranının ölçek dönüştürme sürecine etkisi araştırılmıştır.

Veri üretmede kullanılan modelin türünün, bazı araştırmalarda (Chon ve diğ., 2007; Kaskowitz ve De Ayala, 2001; Kim ve Lee, 2006; Kim ve Kolen, 2006; Lord, 1983; Ironson, 1983; Uysal, 2014) ölçek dönüştürme ve test eşitleme süreci üzerinde etkisi olduğu belirtilmiştir. Bu araştırmada ikili şeklinde puanlanan maddeler, iki ve üç parametreli lojistik modele (2PLM ve 3PLM) göre üretilmiştir. Böylelikle, 3PLM ile geniş ölçekli ve yüksek riskli sınav uygulamalarında şans ile doğru yanıtlanma olasılığı göz önünde bulundurularak önerilerin getirilmesi amaçlanmıştır.

Kolen (1985), geniş ölçekli sınavlarda, grupların yeteneklerinin çarpık dağılıma sahip olabileceğini belirtmekle birlikte, Li ve Lissitz (2000) ölçek dönüştürme ve eşitleme sürecinde grupların ortalamaları arasındaki 0.5'lik farkın gruplar arasındaki farklılıkların etkisini ortaya koymada yeterli olduğunu göstermiştir. Bu çalışmada, Bastari (2000) ve Cao (2008)'nun çalışmasında olduğu gibi benzer ve farklı yetenek dağılımları üzerinde çalışılmıştır. Tablo 2'de ilgili alanyazında yer alan çalışmalar göz önünde bulundurularak araştırmada ele alınan değişkenler ve değişimlenen düzeyler sunulmuştur.

Tablo 2

Simülasyon Deseni

Faktörler (Değişkenler)	Koşullar (Düzeyler)	Koşul Sayısı
Örneklem Büyüklüğü	500-1,000-3,000-10,000	4
Test Uzunluğu	40-50-80	3
Ortak Madde Oranı	%20-%30-%40	3
Parametre Kestirim Modeli	2PLM-3PLM	2
Grupların Yetenek Dağılımı	Benzer (N(0-1) - N(0-1))	2
	Farklı(N(0-1) - N(0.5,1))	

Tablo 2'de yer alan simülasyon deseni incelendiğinde, bu çalışmada, örneklem büyüklüğü (4 düzey), test uzunluğu (3 düzey), ortak madde oranı (3 düzey), parametre kestirim modeli (2 düzey) ve grupların yetenek dağılımı (2 düzey) olmak üzere toplam 144 (4x3x3x2x2) koşul incelendiği görülmektedir.

Verilerin Üretilmesi

Çalışmanın verileri her bir koşul (düzey) için 50 yineleme yapılarak üretilmiştir. Alanyazında yer alan çalışmalarda (Harwell ve diğ., 1996; Hanson ve Beguin, 2002; Hu ve diğ., 2008) araştırma sonuçlarının tutarlı ve genellenebilir olabilmesi için her veri seti üzerinde en az 50 yineleme yapıldığı gözlendiği için bu araştırmada da 50 yineleme yapılmıştır. Çalışmada kullanılan OO, OS, HB ve SL ölçek dönüştürme yöntemlerinin performansını karşılaştırmak için 7200 (144x50) veri seti üretilmiştir. Veri setleri üretilmeden önce ana testler ve ortak maddeler için madde parametreleri ve bu maddeleri yanıtlayacak bireylere ait yetenek parametreleri için, ilgili araştırmalar incelenerek, madde ve yetenek parametre değerleri üretilmiştir. Veriler R yazılımında *"irtoys"* paketi (Partchev, 2016) kullanılarak 3PLM ve 2PLM'ye uygun tek boyutlu iki kategorili tepkiler üretilmiştir. NEAT deseni kullanıldığından, test ve ortak test formları oluşturulmuştur. Eşitleme için üretilen iki test formunun (X ve Y formu) ve ortak test formlarına ait madde ayırt edicilik parametresi (*a*) tek biçimli (uniform) dağılımdan (0.5–2.0) aralığında, *c* parametresi tek biçimli (uniform) dağılımdan (0.2 - 0.3) aralığında ve X ve Y formlarına ait madde güçlük parametresi olan *b* parametresi ise normal dağılımdan (0,1) üretilmiştir.

Ölçek dönüştürme Süreci ve Değerlendirme Ölçütü

Araştırmada, NEAT desenine göre madde parametrelerini kestirmek için ayrı ve aynı kalibrasyon ölçek dönüştürme yöntemlerinin performansı, örneklem büyüklüğü (4 koşul), test uzunluğu (3 koşul), ortak madde oranı (3 koşul) ve grupların yetenek dağılımı (2 koşul) olmak üzere toplam 144 koşul altında incelenmiştir. OO, OS, HB ve SL ölçek dönüştürme yöntemleri ile test formlarının aynı ölçek üzerine yerleştirilme işlemleri *"SNSequate"* paketi (González, 2014), madde ve yetenek parametre kestirimleri *"ltm"* paketi (Rizopoulos, 2015) kullanılarak elde edilmiştir. Madde parametrelerinin kestiriminde Maksimum En Çok Olabilirlik, yetenek parametrelerinin kestiriminde ise Beklenen Sonsal Dağılım yöntemleri kullanılmıştır. Kestirilen madde ve birey parametrelerine ait kalibrasyonlar ve bu parametrelerin,

ayrı kalibrasyon yöntemlerinden OO ve OS karakteristik eğri dönüştürme yöntemlerinden HB ve SL ölçek dönüştürme yöntemleri ile aynı ölçek üzerine yerleştirilme işlemleri için *"plink"* paketi (Weeks, 2010), araştırmada ele alınan değişkenlerin ölçek dönüştürme yöntemleri üzerindeki ortak etkilerinin belirlenmesinde *"lattice"* paketi (Sarkar, 2017) kullanılmıştır. Araştırmada, örneklem büyüklüğü, test uzunluğu, ortak madde oranı, maddelerin üretildiği model ve grupların yetenek dağılımı koşullarına göre ölçek dönüştürme yöntemlerinden elde edilen sonuçlar, bireylerin gerçek yetenek düzeyi ile kestirilen yetenek düzeyleri arasındaki farka dayandığı için Harris ve Crouse (1993) ile Walker ve Kim (2010) tarafından kullanışlı bir istatistiksel indeks olarak ele alınan Eşitlik 7 ile belirtilen RMSD ölçütüne göre değerlendirilmiştir.

$$RMSD = \sqrt{\frac{\sum_i f_i(\theta^* - \theta)^2}{\sum_i f_i}} \quad (7)$$

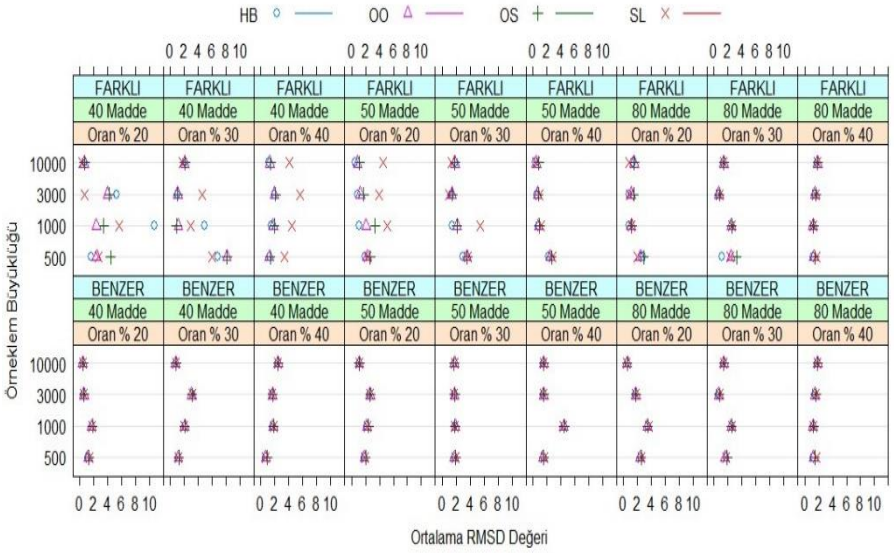
Eşitlik 7'de, θ, gerçek yetenek düzeyini, θ^*, kestirilen yetenek düzeyini, f ise yetenek düzeylerine ait frekansı belirtmektedir. Toplam ölçek dönüştürme hatası (sistematik ve random hata) olan RMSE'nin düzenlenmiş biçimi olarak sonuçların ne kadar yanlı ve doğru olduğunu yansıtan RMSD hesaplanırken, bireylerin kestirilen yetenek düzeyinden gerçek yetenek düzeyleri çıkartılıp kareleri alınarak tüm bireyler için elde edilen değerler toplanmaktadır. Ardından elde edilen bu toplam, yetenek düzeylerine ait frekansa bölünerek karakökü alınır. Araştırmada, koşullara ilişkin madde parametre ve yetenek kestirimlerindeki değişkenliği değerlendirmeye yönelik olarak 50 yineleme ile elde edilen RMSD değerlerinin ortalaması alınarak, bu değerlere ilişkin ortak ve temel etki grafikleri kullanılarak incelenmiştir.

Bulgular

Bu bölümde, gerçek puan eşitlemede, MTK'ye dayalı Ortalama-ortalama (OO), Ortalama-standart sapma (OS), Stocking-Lord (SL) ve Haebara (HB) ölçek dönüştürme yöntemlerinin eşitleme hatalarının çeşitli faktörler altında nasıl değiştiğine yönelik bulgular sunulmuştur. Şekil 1'de, parametre kestirim modeli olarak 2PLM kullanıldığı durumda, yöntemlerin RMSD değerleri ortak etkilerine yönelik elde edilen bulgular gösterilmektedir.

Şekil 1

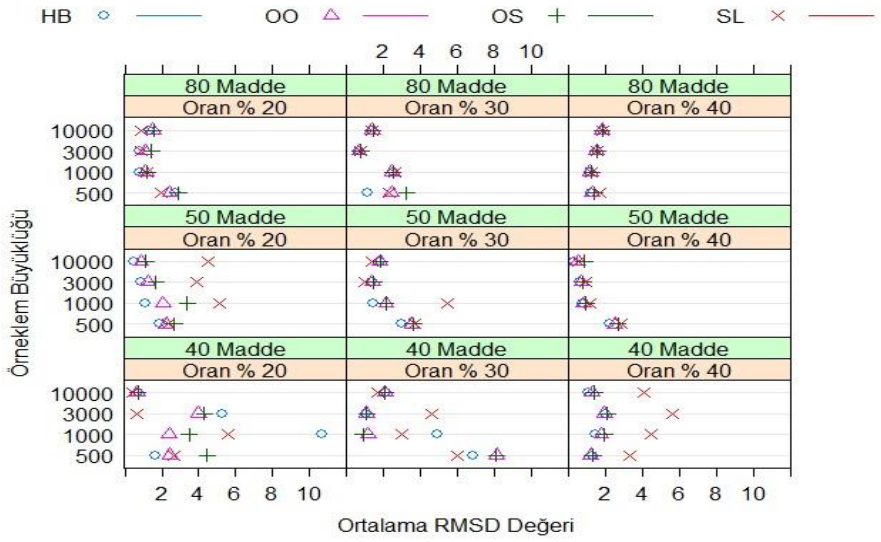
2PLM Kullanıldığında Ölçek Dönüştürme Yöntemlerinin RMSD Değerleri Ortak Etki Grafiği



Şekil 1 incelendiğinde, kullanılan tüm ölçek dönüştürme yöntemlerinin RMSD değerlerinin, farklı dağılımlarda, genel olarak birbirinden farklılaştığı görülmektedir. Bununla birlikte, ölçek dönüşümünün benzer yetenek dağılımlarında gerçekleştiği durumlarda, yöntemlerin RMSD değerlerinin tüm koşullarda birbirine yakın değerlerde olduğu gözlenmiştir. Dolayısıyla, benzer dağılımlarda gerçekleştirilen ölçek dönüştürme işleminde, yöntemlerin, çeşitli koşullar altında birbirine benzer RMSD değerleri ürettiği sonucuna ulaşılmıştır. Ayrıca, en büyük örneklem büyüklüğünde, test uzunluğunun 40 madde, grupların yetenek dağılılmlarının farklı ve ortak madde oranının %20 ve %30 olduğu koşullarda tüm yöntemlerde en düşük RMSD değerinin gözlendiği belirtilebilir. Test uzunluğunun 40 madde, grupların yetenek dağılımlarının farklı ve ortak made oranının %40 olduğu durumda ise SL yönteminin RMSD değerinin, diğer yöntemlerden daha yüksek olduğu görülmüştür. Şekil 2'de, parametre kestirim modeli olarak 2PLM kullanıldığı ve grupların farklı yetenek dağılımlarında yöntemlerin RMSD değerleri ortak etkilerine yönelik elde edilen bulgular gösterilmektedir.

Şekil 2

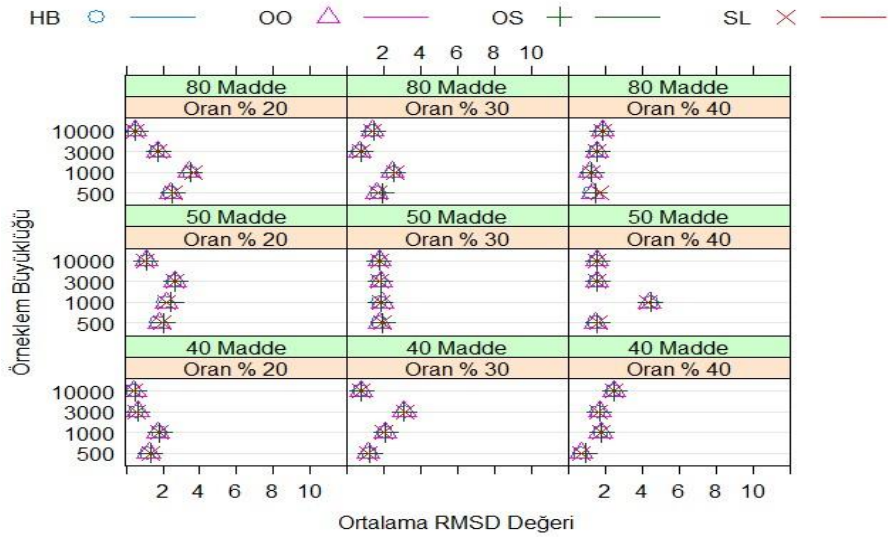
2PLM'nin Kullanıldığı Grupların Farklı Yetenek Dağılımlarında Ölçek Dönüştürme Yöntemlerinin RMSD Değerleri Ortak Etki Grafiği



Şekil 2 incelendiğinde, genel olarak tüm test uzunluklarında, ortak testin uzunluğu arttıkça HB, OO ve OS ölçek dönüştürme yöntemlerinin ortalama RMSD değerlerinde azalma olduğu gözlenmiştir. Bununla birlikte, genel olarak örneklem büyüklüğü arttıkça da RMSD değerlerinin azaldığı görülmektedir. Ortak test madde oranının %20 olduğu durumda tüm ölçek dönüştürme yöntemlerinin en düşük RMSD değerine sahip olduğu görülmüştür. Test uzunluğunun 40 madde ve ortak madde oranın %40 olduğu tüm koşullarda, SL yönteminin diğer yöntemlerden daha yüksek RMSD değerlerine sahip olduğu gözlenmiştir. Şekil 3'te, parametre kestirim modeli olarak 2PLM kullanıldığı ve grupların benzer yetenek dağılımlarında yöntemlerin RMSD değerleri ortak etkilerine yönelik elde edilen bulgular gösterilmektedir.

Şekil 3

2PLM'nin Kullanıldığı Grupların Benzer Yetenek Dağılımlarında Ölçek Dönüştürme Yöntemlerinin RMSD Değerleri Ortak Etki Grafiği



Şekil 3 incelendiğinde, benzer yetenek dağılımında gerçekleştirilen ölçek dönüştürme işlemi sonrasında yöntemlerin, tüm koşullar altında birbirine yakın RMSD değerlerine sahip olduğu görülmüştür. Bu çalışmada kullanılan tüm ölçek dönüştürme yöntemlerinin en düşük RMSD değerleri, ortak madde oranının %20, test uzunluğunun 40 madde ve örneklem büyüklüğünün 3000 ve üzeri olduğu koşullarda gözlenmiştir. Şekil 4'te, parametre kestirim modeli olarak 3PLM kullanıldığı durumda yöntemlerin RMSD değerleri ortak etkilerine yönelik elde edilen bulgular gösterilmektedir.

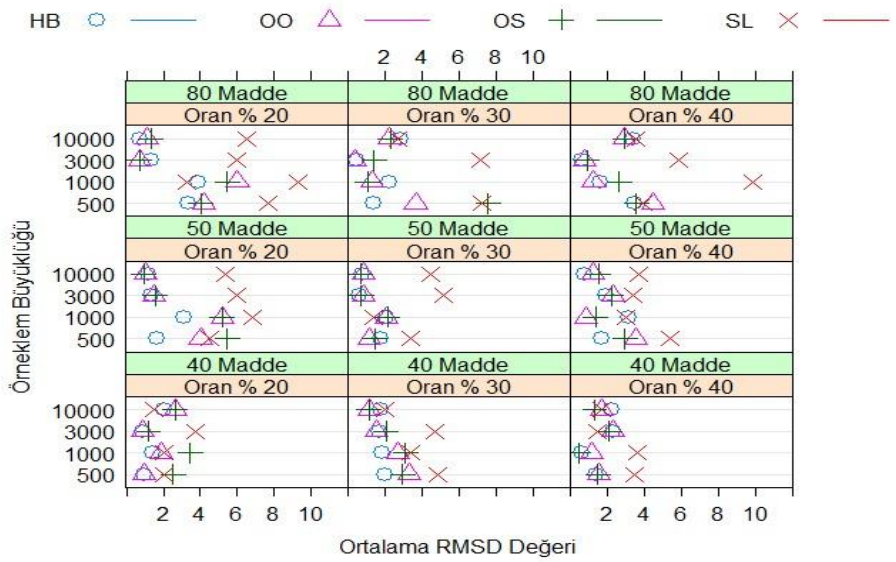
Şekil 4

3PLM Kullanıldığında Ölçek Dönüştürme Yöntemlerinin RMSD Değerleri Ortak Etki Grafiği

Şekil 4 incelendiğinde, 3PLM kullanıldığında benzer yetenek dağılımlarında gerçekleştirilen ölçek dönüştürme işlemleri sonucunda, yöntemlerin RMSD değerlerinin, çeşitli koşullar altında birbirine yakın değerler aldığı görülmektedir. Farklı yetenek dağılımlarında ise OO, OS, HB ve SL ölçek dönüştürme yöntemlerinin RMSD değerlerinin birbirinden farklılaştığı görülmüştür. Ayrıca, farklı yetenek dağılımları kullanıldığında gerçekleştirilen ölçek dönüştürmede genel olarak SL yönteminin diğer yöntemlere göre yüksek RMSD değerine sahip olduğu gözlenmiştir. OO ve HB yöntemlerinin 10,000 örneklem büyüklüğüne sahip olduğu tüm koşullarda RMSD değerlerinin birbirine yakın değerler aldığı görülmüştür. Çalışmada kullanılan ölçek dönüştürme yöntemlerinden OO ve OS yöntemlerinin RMSD değerlerindeki kararlı davranış, grupların yetenek dağılımının benzer olduğu, test uzunluğunun 80 madde ve ortak madde oranının %20 olduğu durumda gözlenmiştir. Böylelikle, örneklem büyüklüğü arttıkça OO ve OS yönteminin tüm koşullardaki RMSD değerinde düzenli bir azalmanın olduğu görülmüştür. Şekil 5'te, parametre kestirim modeli olarak 3PLM kullanıldığı ve grupların farklı yetenek dağılımlarında yöntemlerin RMSD değerleri ortak etkilerine yönelik elde edilen bulgular gösterilmektedir.

Şekil 5

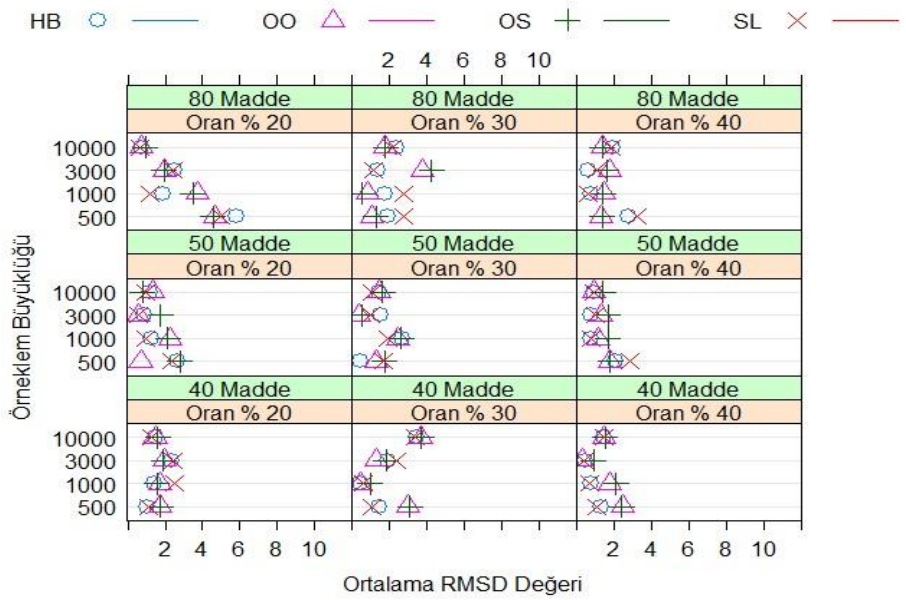
3PLM'nin Kullanıldığı Grupların Farklı Yetenek Dağılımlarında Ölçek Dönüştürme Yöntemlerinin RMSD Değerleri Ortak Etki Grafiği



Şekil 5 incelendiğinde, test uzunluğunun 40 ve 80 madde olduğu tüm koşullarda ortak madde oranı artıkça, ölçek dönüştürme yöntemlerinin RMSD değerlerinin azaldığı görülmektedir. Test uzunluğunun 50 ve 80 madde olduğu tüm koşullarda, genel olarak SL yönteminin yüksek RMSD değerine sahip olduğu gözlenmiştir. Parametre kestirim modeli olarak 3PLM'nin kullanıldığı grupların yetenek dağılımlarının farklı olduğu durumlarda, tüm koşullar altında, genel olarak en düşük RMSD değeri HB yöntemi kullanıldığında elde edildiği görülmektedir. 3000 örneklem büyüklüğünde ve test uzunluğunun 80 madde olduğu tüm koşullarda HB, OO, OS ölçek dönüştürme yöntemlerinin en düşük RMSD değerine sahip olduğu görülmüştür. Şekil 6'da, parametre kestirim modeli olarak 3PLM kullanıldığı ve grupların benzer yetenek dağılımlarında yöntemlerin RMSD değerleri ortak etkilerine yönelik elde edilen bulgular gösterilmektedir.

Şekil 6

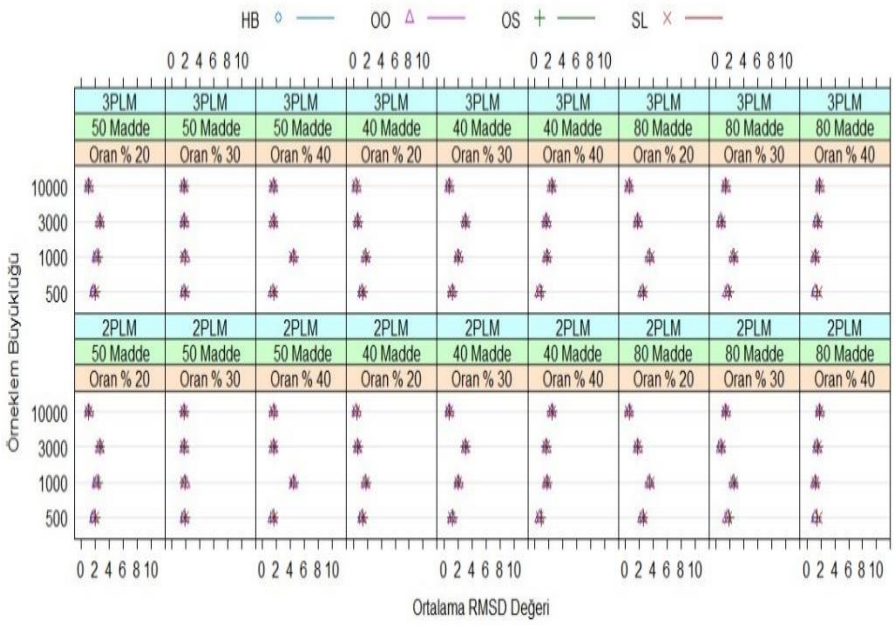
3PLM'nin Kullanıldığı Grupların Benzer Yetenek Dağılımlarında Ölçek Dönüştürme Yöntemlerinin RMSD Değerleri Ortak Etki Grafiği



Şekil 6 incelendiğinde, çalışmada kullanılan tüm ölçek dönüştürme yöntemlerinin, ortak madde oranı %40 olduğu koşulda, düşük RMSD değerine sahip olduğu görülmektedir. Örneklem büyüklüğünün, OO ve OS ölçek dönüştürme yöntemlerinin RMSD değerlerindeki tutarlı etkisi, test uzunluğu 80 madde ve ortak madde oranı %20 olduğu koşullar altında gözlenmiştir. Örneklem büyüklüğü arttıkça RMSD değerinde azalma olduğu görülmüştür. Test uzunluğunun 50 madde ve ortak madde oranının %40 olduğu, ayrıca test uzunluğunun 40 madde ve ortak madde oranının %20 olduğu durumlarda, çalışmada kullanılan tüm yöntemlerin birbirine yakın RMSD değerine sahip olduğu gözlenmiştir. Şekil 7'de, parametre kestirim modeli olarak farklı modellerin kullanıldığı ve grupların benzer yetenek dağılımlarında yöntemlerin RMSD değerleri ortak etkilerine yönelik elde edilen bulgular gösterilmektedir.

Şekil 7

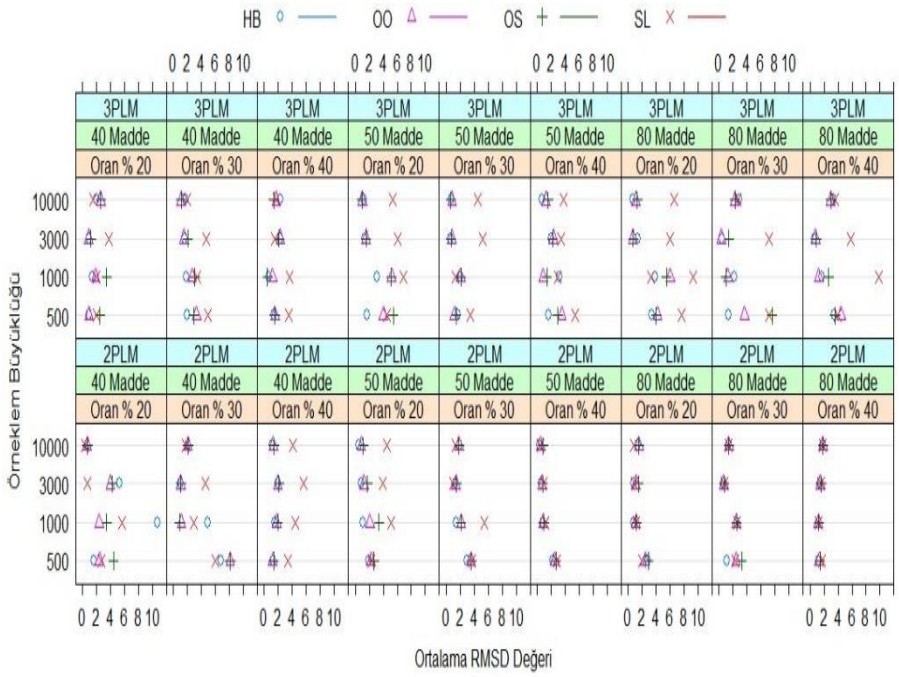
Farklı Modellerin Kullanıldığı Grupların Benzer Yetenek Dağılımlarında Ölçek Dönüştürme Yöntemlerinin RMSD Değerleri Ortak Etki Grafiği



Şekil 7 incelendiğinde, grupların benzer yetenek dağılımlarında kullanılan farklı ölçme modellerindeki çeşitli yöntemlerle gerçekleştirilen ölçek dönüştürme sonucunda elde edilen RMSD değerlerinin benzer olduğu görülmüştür. Bir başka deyişle, parametre kestirim modelinin benzer yetenek dağılımlarında gerçekleştirilen ölçek dönüştürme yöntemlerinin RMSD sonuçlarında farklılaştırma meydana getirmediği görülmüştür. Şekil 8'de, parametre kestirim modeli olarak farklı modellerin kullanıldığı ve grupların farklı yetenek dağılımlarında yöntemlerin RMSD değerleri ortak etkilerine yönelik elde edilen bulgular gösterilmektedir.

Şekil 8

Farklı Modellerin Kullanıldığı Farklı Yetenek Dağılımlarında Ölçek Dönüştürme Yöntemlerinin RMSD Değerleri Ortak Etki Grafiği.



Şekil 8 incelendiğinde, farklı yetenek dağılımlarında kullanılan farklı ölçme modellerindeki çeşitli yöntemlerle gerçekleştirilen ölçek dönüştürme sonucunda elde edilen RMSD değerlerinin farklılaştığı görülmüştür. 2PLM'nin kullanıldığı, test uzunluğunun 50 madde ve ortak madde oranının %30 ve %40 olduğu ve test uzunluğunun 80 madde olduğu tüm ortak madde oranlarında yöntemlerin düşük RMSD değerine sahip olduğu gözlenmiştir.

Tartışma, Sonuç ve Öneriler

Bu çalışmada, gerçek puan eşitlemede, MTK'ye dayalı Ortalama-ortalama (OO), Ortalama-standart sapma (OS), Stocking-Lord (SL) ve Haebara (HB) ölçek dönüştürme yöntemlerinin, çeşitli faktörler altındaki (test uzunluğu, örneklem büyüklüğü, grupların yetenek dağılımı, ortak madde oranı ve parametre kestirim modeli) eşitleme hatalarının karşılaştırılması amaçlanmıştır. Bu amaçla, alanyazın dikkate alınarak, araştırmadan elde edilen sonuçların geniş ölçekli ve yüksek riskli sınavlar için önemli ve katkı sağlayıcı nitelikte olacağı düşünülerek, çeşitli koşullara uygun veriler üretilmiş ve bu koşulların hangilerinde en az hatanın elde edildiği araştırılmaya çalışılmıştır.

Bastari (2000), Gök ve Kelecioğlu (2014), Kim ve Lee (2006), Ngudgratoke (2009) ile Wu ve diğerleri (2009) çalışmalarında, ölçek dönüştürmesinin yapıldığı grupların yetenek dağılımlarında farklılaşma olduğu durumlarda, yöntemlerin eşitleme hatalarının daha yüksek olduğunu belirtmişlerdir. Bu araştırmada da yöntemlerin RMSD değerleri karşılaştırıldığında, 2PLM ve 3PLM'nin kullanıldığı durumlarda, benzer sonuçların elde edildiği görülmektedir. Böylelikle, grupların yetenek dağılımının farklılaştığı durumlarda, yöntemlerin RMSD değerlerinin daha yüksek olduğu sonucuna ulaşılmıştır.

Parametre kestirim modeli olarak 2PLM'nin kullanıldığı durumlarda, örneklem büyüklüğü ve test uzunluğunun yüksek değerlerinde daha kararlı madde parametre kestirimleri elde edildiği için, örneklem büyüklüğünün 1000'in üzerinde, test uzunluğunun 80 madde ve ortak madde oranın ise %40 olduğu koşullarda, yöntemlerin RMSD değerlerinin, diğer koşullara göre daha düşük olduğu bulunmuştur. Böylelikle, parametre kestirim modeli olarak 2PLM'nin kullanıldığı durumlarda, ortak madde oranının orta düzeyinde, test uzunluğu ve örneklem büyüklüğü arttıkça yöntemlerin hata değerlerinin azaldığı görülmektedir. Araştırmanın bu bulgusu, parametre kestirim modeli olarak 2PLM'nin kullanıldığı, Gök ve Kelecioğlu (2014), Arai ve Mayekawa (2011) ile Speron (2009) tarafından yapılan araştırmaların bulguları ile benzerlik göstermektedir. Parametre kestirim modeli olarak 3PLM'nin kullanıldığı durumda ise, test uzunluğunun 50 ve 80 madde olduğu koşullarda artan ortak madde oranı ile birlikte SL yönteminin RMSD değerlerinde de artma oluştuğu sonucuna ulaşılmıştır.

Araştırmada, tüm koşullar göz önünde bulundurulduğunda, SL yönteminin RMSD değerlerinin, diğer yöntemlere göre daha yüksek olduğu görülmekle birlikte, OO ve OS yöntemlerinin birbirine benzer RMSD değerleri ürettiği görülmüştür. Bu durumun, parametre kestiriminin SL yönteminde daha fazla hata vermesinden kaynaklı olabileceği düşünülmektedir. Araştırmanın bu bulguları, karakteristik eğri yöntemlerinin moment yöntemlerine göre daha kararlı sonuçlar elde ettiğini belirten çalışmaların (Baker ve Al-Karni, 1991; Gök ve Kelecioğlu, 2014; Gül ve diğ., 2017; Hanson ve Béguin, 2002; Karkee ve Wright, 2004; Kim ve Cohen, 1998; Kolen ve Brennan, 2004; Kilmen, 2010; Kim ve Lee, 2004; Kim ve Kolen, 2006; Meng, 2012; Way ve Tang, 1991) bulguları ile çelişmektedir. Ayrıca, araştırma bulgusunun, karakteristik eğri yöntemlerinden SL yönteminin HB yöntemine göre daha az hata ürettiğini belirten, Kilmen (2010), Kim ve Kolen (2006), Speron (2009), Gök ve Kelecioğlu (2014), Uysal'ın (2014) yaptıkları çalışmaların sonuçları ile farklılık gösterdiği de görülmektedir. Ölçek dönüştürme yöntemlerinden moment yöntemlerinin benzer eşitleme hatası ürettiği bulgusu, OS yönteminin OO yöntemine göre daha kararlı olduğunu belirten Aksekioğlu (2017) ile Kolen ve Brennan (2004)'ın ve OO yönteminin OS yöntemine göre daha az hata ürettiğini belirten, Baker ve Al-Karni (1991), Way ve Tang (1991), Ogasawara (2000) ve Gündüz'ün (2015) çalışmalarının buguları ile farklılık göstermektedir.

Araştırmada ele alınan koşullar ile sınırlandırılan durumlar için sonuçların genellenmesi olanaklı olmaktadır. Dolayısıyla, araştırmada yer alan değişkenler ve değişkenlerin birbiriyle etkileşimi göz önünde bulundurulmalıdır. Araştırmadan elde edilen sonuçlar ve değişkenler göz önünde bulundurulduğunda, grupların yetenek dağılımı, örneklem büyüklüğü ve test uzunluğu ile birlikte ortak madde oranı değişkenlerinin üzerinde durulması gerekmektedir. Genel sonuç olarak, özellikle, grupların yetenek dağılımının benzer olduğu durumlarda, ölçek dönüştürme yöntemlerinin RMSD değerlerinin daha düşük, birbirine yakın ve kararlı olduğu görülmüştür. Bununla birlikte, büyük örneklemlerde parametre kestirim hatasının azalmasından ve daha kararlı olmasından dolayı SL yöntemi dışında diğer yöntemlerin eşitleme hatalarında azalma meydana gelmiştir. Bu sonuçlar dikkate alındığında kullanılan yöntemlerden herhangi birinin en etkili ya da en başarılı yöntem olduğunu söyleyebilmenin olanaklı olmadığı görülmektedir. Bu araştırma, örneklem büyüklüğü, test uzunluğu, grupların yetenek dağılımları, parametre kestirim yöntemi, ortak madde oranı değişkenleri ve bu değişkenlerin değişimlendiği belirli düzeyler ile sınırlandırılmıştır. Benzer çalışma, grupların yetenek dağılım düzeylerinin, testte yer alan maddelerin ortalama güçlük düzeylerinin farklılaştırılmasıyla da gerçekleştirilebilir. Bununla birlikte, kullanılan ortak testin mini ya da midi test şeklinde düzenlenmesi de gerçek puan eşitleme çalışmalarına katkıda bulunacaktır. Araştırma simülatif veri ile gerçekleştirilmiştir. Benzer çalışmalar gerçek veriler ile de gerçekleştirilebilir. Ayrıca, gerçek veriler baz alınarak bu verilere benzer daha çok veri simülasyonu ile yöntemlerin, eşitlemenin standart hatası, random eşitleme hatası gibi farklı değerlendirme ölçütleri kullanılarak karşılaştırılması da başka araştırma konuları olarak ele alınabilir.

Etik Kurul Kararı

Bu araştırma, 01.01.2020 tarihinden önce simülatif verilerle gerçekleştirildiği için etik kurul kararı zorunluluğu taşımamaktadır.

Kaynakça

Aksekioğlu, B. (2017). *Madde tepki kuramına dayalı test eşitleme yöntemlerinin karşılaştırılması: PISA 2012 Fen testi örneği* (Tez Numarası: 454879) [Yüksek lisans tezi, Akdeniz Üniversitesi]. Yükseköğretim Kurulu Ulusal Tez Merkezi. https://tez.yok.gov.tr/UlusalTezMerkezi/

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

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

Arai, S., and Mayekawa, S. (2011). A comparison of equating methods and linking designs for developing an item pool under item response theory. *Behaviormetrica, 38*(1), 1-16. https://link.springer.com/article/10.2333/bhmk.38.1

Babcock, B., Albano, A., and Raymond, M. (2012). Nominal weights mean equating: A method for very small samples. *Educational and Psychological Measurement, 72*(4), 608–628. https://doi.org/10.1177/0013164411428609

Baker, F. B., and Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*(2), 147-162. http://www.jstor.org/stable/1434796

Barnard, J. J. (1996). *In search of equity in educational measurement: traditional versus modern equating methods.* [Paper presentation]. ASEESA's National Conference at the HSRC Conference Center 1996 Annual Meeting, Pretoria, South Africa.

Bastari, B. (2000). *Linking multiple-choice and constructed-response items to a common proficiency scale* [Doctoral dissertation, Massachusetts Institute of Technology]. https://scholarworks.umass.edu/dissertations_1/5557

Brossman, B. G., and Lee. W-C. (2013). Observed score and true score equating procedures for multidimensional item response theory. *Applied Psychological Measurement, 37*(6), *460–481.* https://doi.org/10.1177/0146621613484083

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. *Journal of Educational Measurement, 22*(1), 13-20. https://www.jstor.org/stable/1434562

Caldwell, L. J. (1984). *A comparison of equating error in lineer and rasch model test equating method* [Unpublished doctoral dissertation]. Florida State University.

Cao, Y. (2008). *Mixed-format test equating: Effects of test dimensionality and common-item sets* (Publication No. 3341415) [Doctoral dissertation, University of Maryland-Maryland]. ProQuest Dissertations and Theses Global.

Chen, H. W. (2001). *Calibration of the ITBS test battery to the complete test battery: A comparison five linking methods* (Publication No. 3009576) [Doctoral dissertation, University of Iowa-Iowa]. ProQuest Dissertations and Theses Global.

Cho, Y. (2007). *Comparison of bookstrap standard errors of equating using IRT and equipercentile methods with polytomously-scored items under the common-item-nonequivalent-group design* (Publication No. 3301690) [Doctoral dissertation, University of Iowa-Iowa]. ProQuest Dissertations and Theses Global.

Chon, K. H., Lee, W. C., and Ansley, T. N. (2007). *Assessing IRT model-data fit for mixed format tests.* (No.26). Center for Advanced Studies in Measurement and Assessment. https://www.semanticscholar.org/paper/Number-26-Assessing IRT-Model-Data-Fit-for-Mixed-%E2%88%97-Chon Lee/49c57e474a54beed3010ab0f2af64985ce6ddb50

Chu, K-L. (2002). *Equivalent group test equating with the presence of differantial item functioning* (Publication No. 3065477) [Doctoral dissertation, The Florida State University-Florida]. ProQuest Dissertations and Theses Global.

Cook, L. L., and Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. *Educational measurement: Issues and Practice. 10* (3), 37-45. https://eric.ed.gov/?id=EJ436860

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

Cui, Z., and Kolen, M. J. (2008). Comparison of parametric and nonparametric bootstrap methods for estimating random error in equipercentile equating. *Applied Psychological Measurement, 32*(4), 334-347. https://doi.org/10.1177/0146621607300854

Dorans, N. J. (1990). Equating methods and sampling designs. *Applied Measurement in Education, 3*(1), 3-17. https://doi.org/10.1207/s15324818ame0301_2

Dorans, N. J., and Holland P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. Journal of Educational Measurement, 37(4), 281-306. https://doi.org/10.1111/j.1745-3984.2000.tb01088.x

Dorans, N. J., Moses, T. P., and Eignor, D. R. (2010). *Principles and practices of test score equating.* (No.41). Educational Testing Service. https://www.ets.org/research/policy_research_reports/publications/report/2010/ilrs

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., and Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19*(2), 143–165. https://doi.org/10.1177/014662169501900203

Eid, G. K. (2005). The effects of sample size on the equating of test items. *Education, 126*(1), 165-180. https://www.thefreelibrary.com/The+effects+of+sample+size+on+the+equating+of+test+items.-a0136846803

Felan, G. D. (2002, February 14-16). *Test equating: Mean, linear, equipercentile and item response theory*. [Paper presentation]. Southwest Educational Research Association 2002 Annual Meeting, Austin, TX, United States.

Fraenkel, J. R., Wallen, N. E., and Hyun, H. H. (2012). *How to design and evaluate research in education*. McGraw-Hill.

Godfrey, K. E. (2007). *A comparison of Kernel equating and IRT true score equating methods* (Publication No. 3273329) [Doctoral dissertation, The University of North Carolina-Chapel Hill]. ProQuest Dissertations and Theses Global.

González, J. (2014). SNSequate: Standard and nonstandard statistical models and methods for test equating. *Journal of Statistical Software*, *59*(7), 1-30. https://www.jstatsoft.org/index

Gök, B., ve Kelecioğlu, H. (2014). Denk olmayan gruplarda ortak madde deseni kullanılarak madde tepki kuramına dayalı eşitleme yöntemlerinin karşılaştırılması. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 10*(1), 120-136. https://dergipark.org.tr/tr/pub/mersinefd/issue/17393/181786

Gül, E., Doğan-Gül, Ç., Çokluk-Bökeoğlu, Ö. ve Özkan, M. (2017). Temel eğitimden ortaöğretime geçiş matematik alt testi asıl sınav ve mazeret sınavlarının madde tepki kuramına göre eşitlenmesi. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 17(*4), 1900-1915. https://dergipark.org.tr/tr/pub/aibuefd/issue/32772/363973

Gündüz, T. (2015). *Test eşitlemede Madde Tepki Kuramına dayalı yetenek parametresine yönelik ölçek dönüştürme yöntemlerinin karşılaştırılması]* (Tez Numarası: 429524) [Yüksek lisans tezi, Gazi Üniversitesi]. Yükseköğretim Kurulu Ulusal Tez Merkezi. https://tez.yok.gov.tr/UlusalTezMerkezi/

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*(3), 144-149. https://doi.org/10.4992/psycholres1954.22.144

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* (Publication No. 3422144) [Doctoral dissertation, University of Iowa, Iowa]. ProQuest Dissertations and Theses Global.

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

Han, K. T. (2008). *Impact of item parameter drift on test equating and proficiency estimates.* (Publication No. 3325324) [Doctoral dissertation, University of Massachusetts, Amherst]. ProQuest Dissertations and Theses Global.

Han, T., Kolen, M. J., and Pohlmann, J. (1997). A comparison among IRT true- and observed score equating and traditional equipercentile equating. *Applied Measurement in Education, 10*(2), 105-121. https://doi.org/10.1207/s15324818ame1002_1

Hanson, B. A., and Beguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*(1), 3–24. https://doi.org/10.1177/0146621602026001001

Harris, D. J., and Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6*(3), 195–240. https://doi.org/10.1207/s15324818ame0603_3

Harris, D. J., and Kolen, M. J. (1986). Effect of examinee group on equating relationships. *Applied Psychological Measurement, 10*(1), 35-43. https://doi.org/10.1177/014662168601000103

Harwell, M., Stone, C. A., Hsu, T. C., and Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20* (2), 101-125. https://doi.org/10.1177/014662169602000201

He, Q. (2010). *Maintaining standards in on-demand testing using item response theory* (No.10/4724). Office of Qualifications and Examinations Regulation. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/605861/0210_QingpingHe_Maintaining-standards.pdf

He, Y. (2011). *Evaluating equating properties for mixed-format tests* (Publication No. 3461151) [Doctoral dissertation, University of Iowa, Iowa City]. ProQuest Dissertations and Theses Global.

Holland, P. W., and Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Eds.), *Educational Measurement* (4th ed., pp. 187-220). Praeger.

Holland, P. W., Dorans, N. J., and Petersen, N. S. (2007). Equating test scores. In C. R. Rao and S. Sinharay (Eds.), *Handbook of statistics* (pp. 169-203). Elsevier. https://doi.org/10.1016/S0169-7161(06)26006-1

Hills, J. R., Subhiyah, R. G., and Hirsch, T. M. (1988). Equating minimum-competency tests: comparisons of methods. *Journal of Educational Measurement, 25*(3), 221- 231. https://www.jstor.org/stable/1434501

Hu, H., Rogers, T. W., and Vukmirovic, Z. (2008). Investigation of IRT-based equating methods in the presenceof outlier common items. *Applied Psychological Measurement*, *32*(4), 311-333. https://doi.org/10.1177/0146621606292215

Ironson, G. H. (1983). Using item response theory to measure bias. In R.K. Hambleton (Eds.), *Applications of item response theory* (2nd ed., pp. 155–174). Educational Research Institute of British Columbia.

Kang, T., and Petersen, N. S. (2009, April, 14-16). *Linking item parameters to a base scale.* [Paper presentation]. National Council on Measurement in Education 2009 Annual Meeting, San Diego, CA, United States.

Karkee, T. B., and Wright, K. R. (2004, April, 12-16). *Evaluation of linking methods for placing three parameter logistic item parameter estimates onto a one-parameter scale*. [Paper presentation]. American Educational Research Association 2004 Annual Meeting, San Diego, CA, United States.

Kaskowitz, G. S., and De Ayala, R. J. (2001). The effect of error in item parameter estimates on the test response function method of linking. *Applied Psychological Measurement, 25*(1), 39-52. https://doi.org/10.1177/01466216010251003

Kolen, M. J. (1981). Comparison of traditional and Item Response Theory methods for Equating Tests. *Journal of Educational Measurement*, *18*(1), 1-11. https://www.jstor.org/stable/1434813

Kolen, M. J. (1985). Standard errors of tucker equating. *Applied Psychological Measurement, 9*(2), 209-223. https://doi.org/10.1177/014662168500900209

Kolen, M. J. (1988). An NCME instructional module on traditional equating methodology. *Educational Measurement: Issues and Practice, 7*, 29-36. https://eric.ed.gov/?id=EJ388096

Kolen, M. J. (2007). Data collection designs and linking procedures. In N. J. Dorans, M. Pommerich, and Holland, P. W. (Eds.), *Linking and aligning scores and scales* (2nd ed. pp. 31-55). Springer. https://doi.org/10.1007/978-0-387-49771-6_3

Kolen, M. J., and Brennan, R. L. (1995). *Test Equating: methods and practices*. Springer Verlag.

Kolen, M. J., and Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices.* Springer.

Kilmen, S. (2010). *Madde tepki kuramına dayalı test eşitleme yöntemlerinden kestirilen eşitleme hatalarının örneklem büyüklüğü ve yetenek dağılımına göre karşılaştırılması* (Tez Numarası: 279926) [Yüksek lisans tezi, Ankara Üniversitesi]. Yükseköğretim Kurulu Ulusal Tez Merkezi. https://tez.yok.gov.tr/UlusalTezMerkezi/

Kim, S., and Cohen, A. S. (1998). A comprasion of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22*(2), 131-143. https://doi.org/10.1177/01466216980222003

Kim, S., and Hanson, B. A. (2002). Test equating under the multiple-choice model. *Applied Psychological Measurement*, *26*(3), 255-270. https://doi.org/10.1177/0146621602026003002

Kim, S., and Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education, 19*(4), 357-381. https://doi.org/10.1207/s15324818ame1904_7

Kim, S., and Lee, W. C. (2004). *IRT scale linking methods for mixed-format tests.* (No.5). American College Testing. https://www.act.org/content/dam/act/unsecured/documents/ACT_RR2004-5.pdf

Kim, S., and Lee, W. (2006). An extension of four IRT linking methods for mixed-format tests. *Journal of Educational Measurement*, *43*(1), 53-76. https://www.jstor.org/stable/20461809

Lee, Y. S. (2007). A comparison of methods for nonparametric estimation of item characteristic curves for binary items. *Applied Psychological Measurement, 31*(2), 121-134. https://doi.org/10.1177/0146621606290248

Lee, W. C., and Ban, J. C. (2009). Comparison of IRT linking procedures. *Applied Measurement in Education, 23*(1), 23-48. https://doi.org/10.1080/08957340903423537

Lee, G., and Fitzpatrick, A. R. (2008). A new approach to test score equating using item response theory with fixed c-parameters. *Asia Pacific Education Review, 9*(3), 248–261. https://www.springer.com/journal/12564

Li, D. (2009). *Developing a common scale for testlet model parameter estimates under the common-item nonequivalent groups design* (Publication No. 3359398) [Doctoral dissertation, University of Maryland, Maryland]. ProQuest Dissertations and Theses Global.

Li, Y. H., and Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement, 24* (2), 115-138. https://doi.org/10.1177/01466216000242002

Liou, M., Cheng, P. E., and Johnson, E. G. (1997). Standard errors of the Kernel equating methods under the common-item design. *Applied Psychological Measurement, 21* (4), 349-369. https://doi.org/10.1177/01466216970214005

Livingston, S. A., and Kim, S. (2010). Random-Groups equating with samples of 50 to 400 test takers. *Journal of Educational Measurement, 47*(2), 175–185. https://www.jstor.org/stable/20778946

Lord, F. M. (1983). Statistical bias in maximum likelihood estimators of item parameters. *Psychometrika*, *48*(3), 477-482. https://doi.org/10.1007/BF02293684

Lord, F. M., and Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score equatings. *Applied Psychological Measurement, 8*(4),453-461. https://doi.org/10.1177/014662168400800409

Loyd, B. H., and Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*(3), 179-193. https://www.jstor.org/stable/1434833

Marco, G. L. (1977). Item Characteristic Curve Solutions to Three Intractable Testing Problems. *Journal of Educational Measurement, 14*(2), 139-160. http://www.jstor.org/stable/1434012

Meng, Y. (2012). *Comparison of Kernel equating and item response theory equating methods* (Publication No. 3518262) [Doctoral dissertation, University of Massachusetts, Amherst]. ProQuest Dissertations and Theses Global.

Michaelides, M. P. (2003, April, 21-25). *Sensitivity of IRT equating to the behavior of test equating items.* [Paper presentation]. American Educational Research Association 2003 Annual Meeting, Chicago, Illinois, United States.

Mohandas, R. (1996). *Test equating, problems and solutions: Equating English test forms for the Indonesian junior secondary schoool final examination administered in 1994* [Doctoral dissertation, Flinders Institute of Technology]. https://flinders-primo.hosted.exlibrisgroup.com/primo-explore/search?vid=FUL&lang=en_US

Ngudgratoke, S. (2009). *An investigation of using collateral information to reduce equating biases of the post-stratification equating method* (Publication No. 3381312) [Doctoral dissertation, Michigan State University-Michigan]. ProQuest Dissertations and Theses Global.

Norman-Dvorak, R. K. (2009). *A comparison of Kernel equating to the test characteristic curve methods* (Publication No. 3350452) [Doctoral dissertation, University of Nebraska-Lincoln]. ProQuest Dissertations and Theses Global.

Nozawa, Y. (2008). *Comparison of parametric and nonparametric IRT equating methods under the common-item nonequivalent groups design* (Publication No. 3347237) [Doctoral dissertation, The University of Iowa-Iowa City]. ProQuest Dissertations and Theses Global.

Ogasawara, H. (2000). Asymptotic standard errors of IRT equating coefficients using moments. *Economic Review (Otaru University of Commerce), 51*(1), 1-23. https://www.researchgate.net/publication/241025868

Partchev, I. (2016). Package "irtoys". (Version 0.2.0). https://cran.r-project.org/web/packages/irtoys/irtoys.pdf

Petersen, N. S., Cook, L. L., and Stocking, M. L. (1983). IRT versus conventional equating methods: a comparative study of scale stability. *Journal of Educational Statistics, 8*(2), 137-156. https://www.jstor.org/stable/1164922

Petersen, N. S., Kolen, M. J., and Hoover, H. D. (1993). Scaling, norming and equating. In Linn, R. L. (Eds.) *Educational measurement* (2nd. pp. 221-262). The Oryx.

Rizopoulos, D. (2015). Package "ltm". https://cran.r-project.org/web/packages/ltm/ltm.pdf

Ryan, J., and Brockmann, F. (2009). A prictitioner's introduction to equating. https://files.eric.ed.gov/fulltext/ED544690.pdf

Sarkar, D. (2017). Package "lattice". https://cran.r-project.org/web/packages/lattice/lattice.pdf

Skaggs, G. (1990). To match or not to match samples on ability for equating: A discussion of five articles. *Applied Measurement in Education, 3* (1), 105-113. https://doi.org/10.1207/s15324818ame0301_8

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42*(2),309–330. https://doi.org/10.1111/j.1745-3984.2005.00018.x

Skaggs, G., and Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56*(4), 495-529. https://doi.org/10.3102/00346543056004495

Speron, E. (2009). *A comparison of metric linking procedures in item response theory* (Publication No. 3370885) [Doctoral dissertation, University of Illinois-Illinois]. ProQuest Dissertations and Theses Global.

Spence, P. D. (1996). *The effect of multidimensionality on unidimensional equating with item response theory* (Publication No. 9703612) [Doctoral dissertation, University of Florida-Florida]. ProQuest Dissertations and Theses Global.

Stocking, M. L., and Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210. https://doi.org/10.1177/014662168300700208

Sinharay, S., and Holland, P. W. (2008). *The missing data assumptions of the nonequivalent groups with anchor test (neat) design and their implications for test equating* (No.09-16). Educational Testing Service. https://files.eric.ed.gov/fulltext/ED507841.pdf

Qu, Y. (2007). *The effect of weighting in Kernel equating using counter-balanced designs* (Publication No. 3282191) [Doctoral dissertation, Michigan State University-East Lansing]. ProQuest Dissertations and Theses Global.

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. http://www.jstor.org/stable/1435244

Tsai, T. H. (1997,March, 24-27). *Estimating minumum sample sizes in random groups equating.* [Paper presentation]. National Council on Measurement in Education Association 1997 Annual Meeting,Chicago, Illinois, United States.

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

Uysal, İ. (2014). *Madde Tepki Kuramı'na dayalı test eşitleme yöntemlerinin karma modeller üzerinde* karşılaştırılması (Tez Numarası: 370226) [Yüksek lisans tezi, Abant İzzet Baysal Üniversitesi]. Yükseköğretim Kurulu Ulusal Tez Merkezi. https://tez.yok.gov.tr/UlusalTezMerkezi/

von Davier, A. A. (2008). New results on the linear equating methods for the non-equivalent groups design. *Journal of Educational and Behavioral Statistics, 33*(2), 186-203. https://www.jstor.org/stable/20172112

von Davier, A. A. (2010). *Statistical Models For Test Equating, Scaling and Linking*. Springer.

von Davier, A. A., and Wilson, C. (2007). IRT true-score test equating: A guide through assumptions and applications. *Educational and Psychological Measurement, 67*(6), 940-957. https://doi.org/10.1177/0013164407301543

Walker, M. E., and Kim, S. (2010, April). *Linking mixed-format tests using multiple choice anchors*. [Paper presentation]. National Council on Measurement in Education Association 2010 Annual Meeting, San Diego, CA, United States.

Wang, X. (2012). *Effect of simple size on IRT equating of uni-dimentional tests in common item non-equivalent group design: a monte carlo simulation study* [Doctoral dissertation, Virginia Institute of Technology]. https://vtechworks.lib.vt.edu/handle/10919/37555

Way, W. D., and Tang, K. L. (1991, April, 3-7). *A comparison of four logistic model equating methods.* [Paper presentation]. American Educational Research Association 1991 Annual Meeting, Chicago, Illinois, United States.

Weeks, J. P. (2010). Plink: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software, 35*(12), 1-33. https://www.jstatsoft.org/article/view/v035i12

Wu, N., Huang, C-Y., Huh, N., and Harris, D. (2009, April, 12-13). *Robustness in using multiple choice iteOS as an external anchor for constructed-response test equating.* [Paper presentation]. National Council on Measurement in Education Association 2009 Annual Meeting, San Diego, CA, United States.

Yang, W. L. (1997). *The effects of content homogeneity and equating method on the accuracy of common item test equating* (Publication No. 9839718) [Doctoral dissertation, Michigan State University-Michigan]. ProQuest Dissertations and Theses Global.

Yang, W. L., and Houang, R. T. (1996, April, 11-13). *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 presentation]. American Educational Research Association 1996 Annual Meeting, New York City, New York, United States.

Zeng, L. (1991). *Standard errors of linear equating for the single-group design*. (No.91-4). American College Testing. https://www.act.org/content/dam/act/unsecured/documents/ACT_RR91-04.pdf

Zhao, Y. (2008). *Approaches for addressing the fit of item response theory models to educational test data*. (Publication No.3337019) [Doctoral dissertation, University of Massachusett-Amherst]. ProQuest Dissertations and Theses Global.

Ankara University Journal of Faculty of Educational Sciences
Year: 2022, Volume: 55, Issue: 1, 1-36
DOI: 10.30964/auebfd.1001128, E-ISSN: 2458-8342, P-ISSN: 1301-3718

Investigation of Scale Transformation Methods in True Score Equating Based on Item Response Theory[1]

ARTICLE TYPE	Received Date	Accepted Date	Published Date
Research Article	09.26.2021	12.22.2021	01.05.2022

Ömay Çokluk Bökeoğlu [2]
Ankara University
Arzu Uçar [3]
Hakkari University
Ebru Balta [4]
Ağrı İbrahim Çeçen University

Abstract

In this research, it was aimed to compare equating errors of scale transformation methods (mean-mean (MM), mean-sigma (MS), Heabera (HB) and Stocking-Lord (SL)) in true score equating based on item response theory (IRT) under different conditions. In line with the purpose of the study, 7200 dichotomous data sets which were consistent with two and three- parameter logistic model were generated with 50 replication under the conditions of sample size (500, 1,000, 3,000, 10,000), test length (40, 50, 80), rate of the common item (20%, 30%, 40%), type of model used in parameter estimation (two and three-parameter logistic models (2PLM and 3PLM)), and ability distribution of groups (similar (N(0-1) - N(0-1)), different (N(0-1) - N(0.5,1)) for the obtained performance of methods. Common item nonequivalent groups equating design was used. R software was used for data generation and analyses. The results obtained from the study were evaluated by using equating error (RMSD) criterion. As a result of the study, considering all the conditions, it was seen that the RMSD values of the SL method were higher than the other methods, but it was seen that the MM and MS methods produced similar RMSD values. In addition, when the RMSD values of the scale transformation methods are compared, similar results are obtained in cases where 2PLM and 3PLM are used, as the sample size and test length increase, equating errors of other methods except the SL method decrease, and It was observed that the methods had lower RMSD values in cases where the common item rate is 40% and the ability distribution of the groups is similar.

Keywords: Haebara, IRT true score equating, mean-mean, mean-sigma, scale transformation, Stocking-Lord, test equating

[1]This study was presented on May 2-5, 2018 in V^{th} International Eurasian Educational Research Congress and published only as an abstract paper.

[2]Prof. Dr., Faculty of Educational Sciences, Department of Measurement and Evaluation, e-mail: cokluk@education.ankara.edu.tr, https://orcid.org/ 0000-0002-3879-9204

[3]Assist. Prof. Dr., Faculty of Education, Department of Measurement and Evaluation, e-mail: arzukapcik@gmail.com, https://orcid.org/ 0000-0002-0099-1348

[4]*Corresponding Author*: Res. Asst. Dr., Faculty of Education, Department of Measurement and Evaluation, e-mail: ebrubalta2@gmail.com, https://orcid.org/ 0000-0002-2173-7189

Ethical comittee approval: Since this research was conducted before 01.01.2020, it does not require an ethics comittee decision.

Purpose and Significance

Educational and psychological tests are used to reveal the learning deficiencies of individuals, to create the profile of schools or individuals, to recruit individuals or to select and place them in an institution. In line with the test scores, decisions are made about individuals that play an active role in shaping their lives. In order to make accurate decisions about individuals, the scores obtained from these tests are expected to be valid and reliable and thus not lead to biased decisions. Thus, while trying to make unbiased decisions about individuals, situations may arise where the same characteristics of different individuals should be measured with different tests (Crocker and Algina, 1986; Petersen et.al., 1993). Although the test developers try to prepare similar tests in terms of content, knowledge and skills for situations of selection and placement, certificate programs and pass-fail etc, there may be differences between test forms in terms of difficulty level and statistical properties (Hambleton et.al., 1991; He, 2010; Kolen, 1988; Kolen and Brennan, 2004). When different test forms are used in order to make fair decisions about individuals, test scores obtained from these forms should be comparable, that is, test equating studies should be performed (Dorans, 1990; Mohandas, 1996; Skaggs, 1990; Woldbeck, 1998).

Testing the equality of test scores in large-scale and high-risk exam applications conducted at the national level where different test forms are applied more than once a year or where different test forms are assumed to be parallel due to exam security is gaining more and more importance today. By placing two or more tests on a common scale with the same content and statistical properties, data with different item properties are combined to determine developments and trends, and provides informative results in revealing differences between regions and countries. Thus, it contributes to the fair treatment of individuals in terms of test developers and practitioners (Angoff, 1971; Eid, 2005; Felan, 2002; Michaelides, 2003; von Davier and Wilson, 2007). It is important to carry out error-free equating studies as much as possible in contributing to test developers and practitioners and in equalizing test scores. In this study, it was aimed to compare equating errors of scale transformation methods (mean-mean (MM), mean-sigma (MS), Heabera (HB) and Stocking-Lord (SL)) in true score equating based on item response theory (IRT) under different conditions.

Method

The purpose of this study was to compare scale transformation methods in true score equating based on IRT under various factors such as sample size, test length, rate of common item, type of model used in parameter estimation and ability distribution of groups. The simulation study was conducted to compare equating errors of scale transformation methods. Hence, it is expected to contribute to theoretical and descriptive studies related to true score equating based on IRT by

estimating the equating error of various scale transformation methods under these factors using the common item nonequivalent groups equating design. In this framework, this research can be evaluated as a simulation study in the nature of descriptive research. In line with the purpose of the study, 7200 dichotomous data sets which were consistent with two and three-parameter logistic models were generated with 50 replication under the conditions of sample size (500, 1000, 3000, 10000), test length (40, 50, 80), rate of the common item (20%, 30%, 40%), type of model used in parameter estimation (two and three parameter logistic models (2PLM and 3PLM)), and ability distribution of groups (similar (N(0-1) - N(0-1)), different (N(0-1) - N(0.5,1)) for the obtained performance of methods. Thus, in this study, a total of 144 (4x3x3x2x2) conditions were examined, including sample size (4 conditions), test length (3 conditions), rate of the common item (3 conditions), type of model used in parameter estimation (2 conditions), and ability distribution of groups (2 conditions). In the study, data were generated with 50 replication for each condition level. It has been observed in studies (Harwell et.al., 1996; Hanson and Beguin, 2002; Hu et.al., 2008) that at least 50 replication were made on each data set in order for the research results to be consistent and generalizable. R software was used for data generation. Common item nonequivalent groups equating design (NEAT) was used. NEAT design is a widely used equating design in test equating. Under the NEAT design, two groups of examinees are administered two test forms that have some items in common. Due to different test forms used, in addition to the group difference in ability levels, test form difference is also introduced into the two groups' test scores. Test scores on the two test forms should then be equated through the common test scores (Kolen and Brennan, 2014). For the traditional NEAT design, the data are collected as two samples from nonequivalent populations that take different tests (X or Y) and the same anchor. Given values for discrimination parameters (a) and guessing parameters (c) were generated from uniform distributions whereas item parameters (b) were generated from a normal distribution. Specifically, distribution of discrimination parameter (a) and guessing parameters (c) were defined as [0.5, 2.0] and [0.2, 0.3], respectively for both test forms X and Y, in addition to anchor test form. In this study as the calibration method mean-mean (MM), mean-sigma (MS), Heabera (HB) and Stocking-Lord (SL) were used. Test forms were placed on the same scale using the *SNSequate* package (Gonzalez Burgos, 2014), item and ability parameter estimations were carried out using the *ltm* package (Rizopoulos, 2015). Maximum Likelihood Estimation Method was used for the estimation of the item parameters, and the Expected Posterior Distribution methods were used for the estimation of the ability parameters. The *plink* package (Weeks, 2010) for the calibrations of the estimation item and person parameters and the placement of these parameters on the same scale with the MM and MS from the separate calibration methods and the HB and SL scale transformation methods from the characteristic curve transformation methods. The *lattice* package (Sarkar, 2017) was used to determine the common effects on methods. The results obtained from the study were evaluated by using equating error (RMSD) criterion.

Results

As a result of the study, considering all the conditions, it was seen that the RMSD values of the SL method were higher than the other methods, but it was seen that the MM and MS methods produced similar RMSD values. In addition, when the RMSD values of the scale transformation methods are compared, similar results are obtained in cases where 2PLM and 3PLM are used, as the sample size and test length increase, equating errors of other methods except the SL method decrease, and it was observed that the methods had lower RMSD values in cases where the common item rate is 40% and the ability distribution of the groups is similar.

Discussion and Conclusions

In this study, it was aimed to compare the equating errors of MM, MS, HB and SL scale transformation methods in true score equating based on IRT under various factors (test length, sample size, ability distribution of groups, rate of common item and parameter estimation model). For this purpose, taking into account the literature, data were produced under various conditions in order to benefit from results obtained from the study in large-scale and high-risk exams. It was investigated which of conditions under consideration had the least error. Kim and Lee (2006) and Gök and Kelecioğlu (2014) interpreted the differentiation of the ability distributions of the groups according to the RMSE values of the parameters estimated by the scale transformation methods in true score equating based on IRT. Researchers stated that the equating errors of the methods are higher when there is a differentiation in the ability distributions in which the scale transformation is made. When the RMSD values were compared in this study, similar results were obtained when 2PLM and 3PLM were used. Thus, it was concluded that there was an increase in the RMSD values of the methods when the ability distribution of the groups differed.

In cases where 2PLM is used as the parameter estimation model since more stable item parameter estimations are obtained as the sample size and test length increase, the RMSD values of the methods were found to be lower than the other conditions in the conditions where the sample size is over 1000, the test length is 80 and the common item ratio is 40%. However, it was concluded that the error values of the methods were lower in the groups with similar ability distribution compared to the groups with different ability distribution. It was concluded that the RMSD values of the SL scale transformation method increased with the increase in the common item ratio in the conditions where 3PLM was used, and the test length was 50 and 80. In the study, the RMSD value of SL scale transformation method is higher than the other methods, considering all conditions. This can be explained by the fact that parameter estimation gives more errors in the SL method. In addition, it is seen that the MM and MS true score equating methods produce similar RMSD values under all conditions. It was observed that the RMSD values of scale transformation methods were lower, close to each other and stable, especially when the ability distribution of the groups was similar. However, due to the decrease in parameter estimation error in large samples and being more stable, equating errors of other methods other than the SL

method decreased. Considering these results, there is no method to suggest the most effective and superior of all methods used.

This study was limited to the sample size, test length, ability distributions of the groups, parameter estimation method, rate of common item variables and certain levels at which these variables changed. A similar study can be carried out by differentiating the skill distribution levels of the groups and the average difficulty levels of the items in the test. However, arranging the common test used as a mini or midi test will also contribute to test equating studies. This study was carried out with simulation data. It can also be performed with real data using similar scale transformation methods in true score equating based on IRT.

Ethical Comittee Approval

Since this research was conducted before 01.01.2020, it does not require an ethics comittee decision.