GEFAD / GUJGEF42(2): 1053-1081(2022)

Madde Tepki Kuramında Eşitleme Hatalarının Belirlenmesinde Kullanılan Delta ve Bootstrap Yöntemlerinin Çeşitli Değişkenlere Göre İncelenmesi*

Investigation of Delta and Bootstrap Methods for Calculating Error of Test Equation in IRT in Terms of Some Variables

Rana SALMANER DOĞAN[1], Şeref TAN[2]

[1]*Gazi Üniversitesi, Eğitim Bilimleri, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı. rasalmaner@gmail.com*
[2]*Gazi Üniversitesi, Eğitim Bilimleri, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı. seraftan@gazi.edu.tr*

Makale Türü/Article Types: *Araştırma Makalesi/Research Article*

Makalenin Geliş Tarihi: 11.04.2021 ***Yayına Kabul Tarihi: 17.02.2022***

ÖZ

Alan yazın incelendiğinde test eşitleme ve faydaları ile ilgili birçok çalışmanın olduğu görülmektedir. Ancak test eşitlemelerin faydalarının yanı sıra sınırlılıkları da vardır. Bunlardan en bilineninin eşitleme hataları olduğu söylenebilir. Bu araştırmada TIMMS 2015 4. sınıf matematik verisi kullanılarak bootstrap ve delta yöntemleri ile elde edilen Madde Tepki Kuramı gözlenen ve gerçek puan eşitleme hatalarının farklı örneklem büyüklükleri ve ölçek dönüştürme yöntemlerine göre incelenmesi amaçlanmıştır. Çalışma bir yöntemin daha kapsamlı ve farklı değişkenler açısından incelenmesi yönüyle betimsel bir çalışmadır. Çalışmada birey yetenek düzeylerinin yakın olması amaçlanarak, TIMMS başarı ölçeğinin orta noktasının (500) üstünde ve altında yer alan ülkelerin (Avusturalya, Kanada, İtalya, İspanya, Hırvatistan, Slovak Cumhuriyeti, Yeni Zelanda, Türkiye ve Gürcistan) verisi içinden rastgele olarak seçilmiş 500, 1000 ve 3000 kişilik örneklemler kullanılmıştır. TIMMS 2015 sınavında kullanılan 14 kitapçık türünden hangisinin kullanılacağının belirlenmesi amacıyla Madde Tepki Kuramı varsayımları incelenmiş ve kitapçıklar arasından model uyum indekslerinin hepsinin kabul edilebilir düzeyde olduğu kitapçık çifti seçilmiştir. Yapılan analizler sonucunda genel olarak her iki

* **Alıntılama:** Salmaner Doğan, R. ve Tan, Ş. (2022). Madde tepki kuramında eşitleme hatalarının belirlenmesinde kullanılan delta ve bootstrap yöntemlerinin çeşitli değişkenlere göre incelenmesi. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi, 42*(2), 1053-1081.

yöntemde de en düşük hata değerlerinin Stocking Lord yönteminde elde edildiği ve delta yöntemiyle elde edilen hataların tüm örneklem büyüklüğünde bootstrap yönteminde elde edilen hatalardan yüksek olduğu bulunmuştur.

Anahtar Sözcükler: *Eşitlemenin standart hatası, Delta yöntemi, Bootstrap*

ABSTRACT

After investigation of the literature, it can be seen that there are lots of studies about test equating and its benefits. However, besides its good sides, it has some limitations, and it can be said that the most known one of them is test equating error. In this study, it is aimed the examination of Item Response Theory observed and true score equating errors, obtained using TIMMS 2015 4th Grade math data by bootstrap and delta methods, according to different sample sizes and scale transformation methods. The study is a descriptive research in terms of investigating one method in a more detailed way, and according to different variables. It is used randomly chosen 500, 1000, and 3000 sized samples from the countries (Australia, Canada, Italia, Spain, Croatia, Slovak republic, New Zealand, Turkey, and Georgia) above and below center point (500) of TIMMS success scale. To determine which one would be used among 14 types of TIMMS 2015 booklets, it was controlled Item Response Theory assumptions, and it was chosen the pair of booklets, having all acceptable model fit indices. As results of the analysis, it was observed that for both methods, scale transformation method, having smallest equating errors, was Stocking Lord, and that for all sample sizes, errors estimated by bootstrap method were smaller than one by delta.

Keywords: *Standard error of equating, Delta method, Bootstrap*

GİRİŞ

Uluslararası öğrenci başarısını karşılaştırma projeleri (TIMMS, PIRLS, PISA, vb.), katılan ülkelerin kendi eğitim sistemlerini değerlendirmelerine ek olarak öğrencilerin matematik, fen bilgisi ve okuma alanlarındaki bilgi ve becerilerini yıllara göre takip etmelerini sağlayan, geniş ölçekli değerlendirme içeren projelerdir (MEB, 2011). Ülkelerden beklenen, sonuçlardan yola çıkarak ülke genelinde gerekli reformları gerçekleştirmeleri ve söz konusu projelere katılımı sağlayarak bu reformların etkisini takibe almalarıdır. Bu kapsamda bu projeler ve içlerinde yer alan standart testler önemli görülebilir.

Standart testlerin, bireyler hakkında alınacak kararlardaki önemi düşünüldüğünde, bazı özellikler taşıması gerekmektedir. Bunlar adayların birbiriyle kıyaslanmasını sağladığı (Tekin, 2009) için geçerli ve güvenilir ölçmelere olanak tanıması, kullanışlı olması, duyarlı ölçümler yapması ve yanlı kararlara yol açmamasıdır (Turgut ve Baykul, 2012). Standartlaştırılmış testler farklı alanlarda öğrencilerin eğitimsel ve temel yeteneklerindeki gelişimini ölçen genellikle geniş ölçekli testlerdir (Popham,1999). Geniş ölçekli değerlendirmelerde karşılaşılan zorluklar test güvenliğini sağlamak ve potansiyel yanlılık ile haksız test şartlarını oluşmaması için test materyallerinin gizliliğini sağlamaktır (Çağlak, 2015). Bu zorluklardan bir diğeri bireyler hakkında verilecek gelişimi değerlendirme veya seçme kararlarının geçerli ve tutarlı olabilmesi için her yıl veya yılda birden fazla aynı amaçla uygulanan testlerin çoklu formlarından elde edilen puanların birbirleri ile karşılaştırılabilir hâle dönüştürülmesi gerekliliğidir (Kilmen, 2010). Birden fazla test formu kullanıldığı ya da katılımcılar bir testin farklı formlarını aldığı durumlarda uygulamada meydana gelecek etkilerin önüne geçmek için farklı testlerin eşitlenmiş olması beklenmektedir (Felan, 2002). Test güçlük farklarını ayarlamak için kullanılan istatistiksel yönteme "eşitleme" denir (Kolen ve Brennan, 2004: 2).

Eşitlemede veri toplamak için yaygın olarak kullanılan üç desen vardır (Kolen ve Brennan, 2004, s.12). Randum grup deseninde testleri alan grubun aynı olması, grup

değişkenini sabit tutmaktadır. Böylece test formları arasındaki güçlük farkları gruba bağlı olmadan değerlendirilmektedir. Bu güçlük farklarından yola çıkılarak testler arasında eşitleme denklemi elde edilmektedir (Hambleton, Swaminathan, ve Rogers, 1991; Kolen, 1988; Livingston, 2004). Bir diğeri dengelenmiş tek grup desenidir. Bu desende test alacak gruptaki bireyler ikiye ayrılarak, farklı test koşullarına farklı sıralarla atanırlar. Bu desende grupların yetenek bakımından birbirine olabildiğince benzer olması istenir. Ayrıca sıra etkisinin incelenmesinde kullanılabilir (Kolen ve Brennan, 2004, s.14). Son olarak bu çalışmada kullanılan eş değer olmayan gruplar için ortak madde desenidir. Bu desende iki formda ortak maddeler bulunmaktadır. Ve formlar farklı katılımcı gruplarına uygulanır (Kolen ve Brennan, 2004, s.18). Madde Tepki Kuramı'na (MTK) göre eşitlemede, ortak maddeler, madde ve yetenek parametrelerinin aynı ölçek üzerine yerleştirilmesinde kullanılır (Cook ve Eignor, 1991). Farklı test formlarının eşitlenmesinin ek olarak, bu desen farklı yıllarda farklı formların kullanılmasıyla gerçekleşen durumlarda da kullanılır (Michalides, 2003).

Eşitleme yöntemleri genellikle Klasik Test Kuramı'na (KTK) ve Madde Tepki Kuramı'na (MTK) dayalı eşitleme yöntemleri olmak üzere ikiye ayrılmaktadır. Klasik Test Kuramı'na dayalı yöntemler; eşit yüzdelikli eşitleme, doğrusal eşitleme ve ortalama eşitleme olmak üzere üçe ayrılır (Kolen ve Brennan, 2004, s.29). Eşitleme yöntemlerinin avantajlarına bakıldığında zor testi alan bireylere karşı olası haksızlıkları engellemekte ve test formlarından kaynaklanan yanlılık problemlerini ortadan kaldırmaktadır (Angoff, 1971; Cook ve Eignor, 1991; Hambleton, Swaminathan ve Rogers, 1991). Ancak bunun yanı sıra iki formun puan dağılımlarının aynı olduğu durumlarda, doğrusal eşitleme ile eşit yüzdelikli eşitleme aynı sonucu verir (Kolen, 1988; Felan, 2002; Livingston, 2004). Güçlük düzeyleri farklılık gösteren iki formun doğrusal eşitliği, başarılı bir grup ile daha az başarılı bir grupta farklılık gösterebilir. Bu durumun ortaya çıkmasının bir nedeni Klasik Test Kuramı olabilir.

Klasik Test Kuramı'na dayalı yöntemlerin tersine MTK' ye dayalı yöntemler, test formlarını eşitlemede her bir madde için doğrusal olmayan ve maddenin doğru cevaplanma olasılığı ile yetenek düzeyi arasındaki ilişkiyi gösteren bir eğri olan "madde

karakteristik eğrisini" kullanır (Lord, 1980). Madde Tepki Kuramı'na dayalı olarak yapılan test eşitleme çalışmalarında, testlerden alınan puanların aynı ölçekte yer almasını sağlamak için ortak maddeler kullanılır (Cook ve Eignor, 1991). Bunun faydası ise, bir teste ait çok sayıdaki maddenin parametrelerinin tek oturumda hesaplanmasını sağlamasıdır. Bazı oturumlarda testte madde sayısı her bireye uygulanamayacak kadar fazla olabilir. Bu durumda, soruların uygulandığı yanıtlayıcı grubundaki her birey belli bir test deseni kapsamında oluşturulmuş soru kümelerine ve ortak maddelere yanıt verebilirler (Verhelst, 2004).

Madde Tepki Kuramı'na dayalı olarak ortak maddeli test eşitleme yönteminin kullanıldığı durumlarda, ölçek dönüşümü için çeşitli yöntemler geliştirilmiştir. Bu çalışma kapsamında bahsedilen yöntemlerin hepsi kullanılmıştır. Bunlardan ilki moment yöntemleridir. Bu yöntemler ortalama-ortalama ve ortalama- standart sapmadır. Loyd ve Hoover (1980) tarafından tanımlanan ortalama- ortalama yöntemi güçlük parametresinin (b) yanında ayırıcılık (a) parametresinin ölçek puan dönüşümünde kullanıldığı yöntemdir. Ortalama, standart sapma yöntemi ise Marco (1977) tarafından tanımlanmıştır. A ve B'nin hesaplanmasında güçlük parametresinin standart sapması ve ortalaması kullanılır (Kolen ve Brennan, 2004: 183). Bir diğer ölçek dönüştürme yöntemi ise karakteristik eğri yöntemidir. Bu yöntemde iteratif süreçleri içeren Haebara yaklaşımı ile Stocking-Lord yaklaşımları bulunur. Haebara yöntemi Haebara (1980) tarafından geliştirilmiştir. Bu yönteme göre bir yetenek düzeyindeki katılımcılar için, madde karakteristik eğrileri arasındaki fark, her bir maddeye ait madde karakteristik eğrileri arasındaki farkın karelerinin toplamıdır. (Kolen ve Brennan, 2004, s.184). Stocking-Lord'a (1983) göre ise belli bir yetenek düzeyindeki katılımcılar için, madde karakteristik eğrileri arasındaki fark, her bir maddeye ait madde karakteristik eğrileri arasındaki farkın toplamının karesidir (Kolen ve Brennan, 2004, s.185).

Test eşitlemelerinin bahsedilen faydalarının yanı sıra birtakım sınırlılıkları bulunmaktadır. Bunlardan en bilineninin eşitleme hataları olduğu söylenebilir. Test eşitlemede hata kavramı bireyin yetenek düzeyi ile almadığı test için kestirilen yetenek düzeyi arasındaki farkla açıklanmaktadır (Cook ve Eignor, 1991). Eşitleme hatasının

düşüklüğü, eşitlemede seçilen yöntemin doğru bir seçim olduğunu göstermektedir (Kolen ve Brennan, 2004, s.248). Bahsedilen eşitleme hataları seçkisiz ve sistematik olmak üzere iki başlık altında incelenebilmektedir (Felan, 2002). Seçkisiz eşitleme hatası birey puanlarının birey evreni ya da evrenlerinden alınmış örneklemler olduğu varsayıldığında kullanılır (Kolen ve Brennan, 2004, s.247). Eşitlemenin standart hatası eşitleme bağlantılarının kestiriminde kullanılan seçkisiz hata indeksidir ve sistematik hatayı direkt olarak etkilemez (Kolen ve Brennan, 2004, s.248).

Kolen ve Brennan (2004, s.250) eşitlemenin standart hatası ile ilgili bölümünde standart hatanın kestirilmesi için genel olarak iki çeşit yöntem geliştirildiğini belirtmiştir. Bu iki yöntem de çalışmada kullanılan MTK'ye dayalı olarak daha ayrıntılı bahsedilmiştir. Bunlardan ilki olan bootstrap yönteminde birey evreni ya da evrenlerinden çekilen örneklemler kullanılarak yapılan eşitlemenin hipotetik tekrarları kullanılarak eşitlenen puanların standart sapmasıdır (Patton, Cheng, Yuan ve Diano, 2014). Bootstrap yöntemi MTK eşitleme standart hatasını elde etmek kullanılmıştır (Kolen ve Brennan, 2004; Tsai, Hanson, Kolen ve Forsyth, 2010). İkinci yöntem analitik başka bir deyişle delta yönteminde ise bir eşitlik elde edilir. Örneklem istatistikleri kullanılarak standart hatalar kestirilir (Kolen ve Brennan, 2004: 259). Bu yöntem de MTK eşitleme hatalarının kestiriminde Osawaga, (2001) ve Zhang (2020) tarafından kullanılmıştır. Çalışmalarda eşitlenmiş puanların kestirilen madde parametrelerinin ve MTK ölçek dönüştürme katsayılarının fonksiyonları olduğu vurgulanmış ve madde parametrelerinin ve ölçek dönüştürme yöntemlerinin varyans- kovaryans matrisi ile MTK eşitlemelerin standart hatalarını hesaplamak için matematiksel formüller oluşturulmuştur.

Araştırmanın Önemi ve Amacı

Eşitleme hataları ile ilgili olarak alanyazın incelendiğinde bu konuda birçok çalışma yapıldığı belirlenmiştir. Bunlardan bazılarında araştırmacılar Klasik Test Kuramı'nda değişen madde fonksiyonu ve şans başarısı gibi belirli bazı kavramların eşitlemeye olan etkisini incelerken (Bozdağ ve Kan, 2010; Demirus ve Gelbal, 2016), bazılarında ise aynı koşullar altında hangi KTK eşitleme yönteminin daha az hata ürettiğini belirlemeyi amaçlamıştır (Demir ve Güler, 2014; Öztürk ve Anıl, 2012, Zeng, 1991).

Ayrıca Madde Tepki Kuramı'na (MTK) dayalı eşitleme yöntemlerini örneklem büyüklüğü, madde sayısı, ortak madde oranı ve yetenek dağılımı gibi değişkenler açısından eşitleme hatalarını kullanarak karşılaştıran çalışmalar (Battauz, 2013; Gök ve Kelecioğlu, 2014; Gül, Doğan Gül, Bökeoğlu ve Özkan, 2017; Kilmen, 2010; Li, Jiang ve Von Davier, 2012) ile MTK gerçek ve gözlenen puan eşitleme hatalarının karşılaştırıldığı bazı çalışmalar vardır (Aksekioğlu, 2017; Arıkan, 2017; Uyar, Aksekioğlu ve Öztürk Gübeş, 2020).

Hata elde etme yöntemlerini ele alan çalışmalar incelendiğinde ise Osawaga (2001) çalışmasında MTK eşitleme ve standart hataları delta yöntemi kullanılarak hesaplanmıştır. Tsai, Hanson, Kolen ve Forsyth (2010) çalışmalarında bootstrap hatalar kullanılarak MTK eşitleme yöntemlerini karşılaştırmıştır. Zhang (2020) ve Zhang ve Zhao (2019) standart hataların hesaplanmasında kullanılan bootstrap ve delta yöntemleriyle çoklu veri atama yöntemini karşılaştırmıştır.

Alanyazında bu çalışma delta ve bootstrap hata kestirim yöntemlerini karşılaştırması ve bunu geniş ölçekli standart bir test verisi üzerinden yapması açısından önem arz etmektedir. Genel olarak bu konu ile ilgili alan yazında yer alan çalışmalarda simülasyon veri üzerinden kurgusal düzeneklerle incelendikleri görülmektedir. Bu çalışmada test uygulayıcılarının ve araştırmacıların da kullanmak durumunda olduğu gerçek veri setleri kullanılması, gerçek ve gözlenen yöntemlerin örneklem büyüklüğü ve ölçek dönüştürme yöntemine göre incelenmesi açısından katkı sağlayabilir.

Araştırmanın amacı Madde Tepki Kuramı (MTK) test eşitlemede hataların belirlenmesinde kullanılan iki yöntem olan delta ve bootstrap yöntemlerinin örneklem büyüklüğü ve ölçek dönüştürme yöntemleri değişkenleri bakımından incelenmesidir. Bu kapsamda aşağıda yer alan araştırma problemlerine cevap aranacaktır:

1. Gerçek puan eşitlemede örneklem büyüklüğüne ve MTK eşitlemede kullanılan ölçek dönüştürme yöntemlerine göre delta ve bootstrap yöntemlerinin ürettiği hatalar nasıl değişmektedir?

2 Gözlenen puan eşitlemede örneklem büyüklüğüne ve MTK eşitlemede kullanılan ölçek dönüştürme yöntemlerine göre delta ve bootstrap yöntemlerinin ürettiği hatalar nasıl değişmektedir?

3. Gerçek ve gözlenen puan eşitleme hatalarının bootstrap ve delta yöntemleri için karşılaştırılmaları nasıldır?

YÖNTEM

Araştırmanın Türü

Araştırma MTK test eşitlemede hataların belirlenmesi için kullanılan delta ve bootstrap yöntemlerini ölçek dönüştürme yöntemleri olan ortalama-ortalama (O-O), ortalama-standart sapma (O-S), Stocking Lord (SL) ve Haebara (HA) ve örneklem sayısı (500, 1000, 3000) gibi değişkenler açısından karşılaştırmayı amaçladığından bir yöntemin daha kapsamlı ve farklı değişkenler altında incelenmesi açısından betimsel bir çalışmadır. Betimsel çalışmalarda araştırmacıların amacı verilen bir durumu mümkün olan en dikkatli ve tam şekilde tanımlamaktır (Büyüköztürk, Çakmak, Akgün, Karadeniz ve Demirel, 2014: 15).

Evren ve Örneklem

Çalışma evrenini 2015 yılında TIMMS 4. sınıf matematik sınavına giren 56 ülkeden 272,907 öğrenci oluşturmaktadır. Bu evrenden TIMMS başarı ölçeğinin orta noktasının (500) üstünde olan 5 ülke (Avusturalya, Kanada, İtalya, İspanya, Hırvatistan) ve ölçeğin orta noktasının (500) altında yer alan 4 ülke (Slovak Cumhuriyeti, Yeni Zelanda, Türkiye ve Gürcistan) verileri birleştirilmiştir. Birleştirilen veri setlerinden 1. Ve 2. kitapçık verilerinden rastgele olarak 500, 1000 ve 3000 kişilik örneklemler seçilerek analizlere dâhil edilmiştir.

Verilerin Toplanması

Çalışmada TIMMS 2015 4. Sınıf matematik sınavı verileri kullanılmıştır. Veriler internette herkesin kullanımına açık olarak yayınlandığından etik kurul izni gerektirmemektedir. Veriler TIMMS ve PIRLS uluslararası çalışma merkezinin resmi internet sitesinden SPSS dosyası olarak indirilmiştir.

Veri Analizi

Analizde MTK 3 parametreli lojistik model kullanıldığından kitapçıklardaki maddelerin hepsi 0-1 ikili olarak yeniden kodlanmıştır. Bu yapılırken kısmı açık uçlu maddelerde tam cevap 1 olarak puanlanırken diğer cevaplar 0 olarak kodlanarak ikili puanlanan veri seti elde edilmiştir. Tüm kitapçıklar için MTK varsayımları olan tek boyutluluk ve yerel bağımsızlık incelenmiştir.

Tek boyutluluk varsayımının test edilmesi için Mplus 6 (Muthén ve Muthén, 2010) paket programı kullanılarak tetrakorik Açımlayıcı Faktör Analizi (AFA) uygulanmıştır ve model-veri uyumunun belirlenmesi için tek faktörlü modelin uyum indeksleri (p değeri, x^2/sd, RMSEA, CFI ve TLI değerleri, faktör yükleri ve faktör özdeğerleri) incelenmiştir (Byrne, 2013). Yapılan analizden elde edilen model uyum indeksleri Tablo 1'de gösterilmiştir.

Tablo 1. Kitapçıkların Tetrakorik Korelasyona Dayalı AFA Uyum İndeksleri

Uyum indeksleri / Kitapçık numarası	Madde sayısı	Örneklem (N)	P değeri	x^2/sd	Faktör yükleri	RMSEA	CFI	TLI (NFI)
1	25	4060	0.00	3.74	0.26-0.77	0.026	0.97	0.97
2	23	4051	0.00	4.50	0.31-0.75	0.029	0.97	0.96
3	23	4048	0.00	9.72	0.24-0.74	0.046	0.93	0.92
4	25	4055	0.00	18.6	0.26-0.72	0.066	0.82	0.80
5	29	4039	0.00	18.6	0.30-0.91	0.066	0.96	0.96
6	28	4072	0.00	15.3	0.27-0.91	0.030	0.99	0.99
7	25	4074	0.00	5.30	0.28-0.69	0.033	0.95	0.95
8	23	4101	0.00	6.47	0.38-0.67	0.037	0.95	0.94
9	27	4064	0.00	22.66	0.34-0.86	0.073	0.98	0.98
10	27	4070	0.00	26.61	0.26-0.83	0.079	0.98	0.98
11	24	4080	0.00	6.96	0.28-0.74	0.038	0.94	0.93
12	24	4048	0.00	12.42	0.30- 095	0.053	0.94	0.94
13	26	4094	0.00	41.84	0.24-0.92	0.100	0.92	0.92
14	28	4075	0.00	32.00	0.29-0.81	0.083	0.92	0.92

Tablo 1 incelendiğinde, p değerlerinin hepsinin anlamlı olduğu, x^2/sd değeri içinse 1.ve 2. kitapçıkların dışındaki tüm kitapçıkların x^2/sd değeri 5'ten büyük olduğu görülmektedir. x^2 değerinin örneklem büyüklüğüne duyarlılığından dolayı model uyumunu değerlendirmede alternatif olarak geliştirilen CFI ve TLI değerleri incelendiğinde ise 4. Kitapçık dışında tüm kitapçıkların TLI değerlerinin 0,90'dan yüksek olduğu ve aynı şekilde CFI değerlerinin de 4. Kitapçık değeri dışındakilerin 0,92 yüksek olduğu görülmektedir. RMSEA değerleri ile ilgili olarak 13. ve 14. kitapçık hariç tüm kitapçık değerlerinin 0,08'in altında olduğu gözlenmiştir. Schermelleh- Engel, Moosbrugger ve Müller (2003) çalışmasında örneklem büyüklüğü 250'den fazla ve madde sayısı 12 ve 30 arasında olduğu durumlar için kabul edilebilir uyumu kriterleri olarak χ^2/sd değerinin 5'den küçük, TLI değerinin 0,90'dan büyük ve RMSEA değerinin 0,10'den küçük ve 0,05'den büyük olması gerektiğini belirtmiştir. 4. Kitapçık dışındaki kitapçıklar için model – veri uyumunun iyi olduğu yorumu yapılabilir. Ancak tüm uyum indekslerinin iyi aralıkta olduğu belirlenen 1. ve 2. kitapçık analize dâhil edilmiştir. Bu bölümden sonra bu kitapçıklar kullanılarak analize devam edilmiştir.

Ek olarak Tablo 2' de 1. ve 2. kitapçıkların öz değerleri verilmiştir.

Tablo 2. 1. ve 2. Kitapçık Özdeğerleri

Özdeğerler / Kitapçıklar	1	2	3	4	5	6	7	8	9	10
1	8.24	1.22	1.13	1.03	0.98	0.95	0.88	0.85	0.80	0.78
2	7.51	1.28	1.18	1.09	0.95	0.86	0.82	0.77	0.75	0.73

Tablo 2'de yer alan faktör özdeğerleri incelendiğinde her kitapçık için de birinci ve ikinci faktör arasındaki oranın 3'den büyük olduğu görülmüştür. Lord (1980:21) tek boyutluluğun belirlenmesinde, faktör analizinde birinci faktör yüküne ait özdeğerinin (eigen value), ikinci faktör yüküne ait özdeğerden çok farklı olmasının ve ikinci faktör yüküne ait özdeğerin diğerlerinden çok farklı olmamasının tek boyutluluğa ölçü sayılabileceğini savunmuştur. Bu bilgiler ışığında iki kitapçığın da baskın bir boyuta sahip olduğu ya da tek boyutlu olduğu şekilde yorum yapılabilir.

Yerel bağımsızlık için alanyazında (Embretson ve Reise, 2000, s.48; Hambleton vd., 1991, s.11) tek boyutluluk varsayımının karşılanmış olmasının, kanıt olarak sunulabildiği belirtilmektedir. Ayrıca Lord (1980, s.21) tek boyutluluk varsayımının karşılanması durumunda belli bir yetenek düzeyindeki bireylerin maddelere verdiği tepkiler arsındaki korelasyonun sıfır olduğunu ve bu nedenle tek boyutlu olan ölçeğin aynı zamanda yerel bağımsızlık varsayımını da karşıladığını belirtmiştir. Bu açıklamalar doğrultusunda kitapçıklar baskın bir tek boyuta sahip olduğundan yerel bağımsızlık varsayımını sağladığı yorumu yapılabilir.

Varsayım analizi sonucunda seçilen 1. ve 2. kitapçıkların eşitlemesi için MTK eş değer olmayan gruplarda ortak madde deseni kullanılmıştır. Çalışma deseni Tablo 3'te gösterilmiştir.

Tablo 3. Çalışmanın Eşitleme Deseni

Grup	1.Kitapçık	Ortak	2.Kitapçık	Toplam
Y	14	11		25
X		11	12	23

Tablo 3 incelendiğinde iki kitapçık arasında 11 ortak maddenin bulunduğu ve madde sayılarının sırasıyla 25 ve 23 olduğu görülmektedir. Bu desen kullanılarak MTK gözlenen ve gerçek eşitlemelerin yapılması için öncelikle ST programı kullanılarak ölçek dönüştürme yöntemlerinin katsayıları hesaplanmış ve daha sonra gözlenen ve gerçek puan eşitlemeler PIE paket programı kullanılarak yapılmıştır.

Madde –Test Korelasyonu

Madde-toplam korelasyonu ölçek maddelerin geçerliliğinin belirlenmesi için kullanılmıştır. Korelasyonun yüksek düzeyde ve pozitif yönlü olması ölçme aracındaki maddelerin benzer davranışları örneklediğini ve ölçeğin iç tutarlılığının yüksek olduğunu ifade etmektedir (Büyüköztürk, 2017, s.183). Çalışmada kullanılacak iki kitapçığa ait düzeltilmiş madde-toplam korelasyonu Tablo 4'te sunulmuştur.

Tablo 4. Kitapçıkların Düzeltilmiş Madde-Toplam Korelasyonu

Kitapçık	Madde	Düzeltilmiş Madde-Toplam Korelasyonu	Madde	Düzeltilmiş Madde-Toplam Korelasyonu	Madde	Düzeltilmiş Madde-Toplam Korelasyonu	Madde	Düzeltilmiş Madde-Toplam Korelasyonu
1	**1**	0,32	**8**	0,32	**15**	0,36	**22**	0,45
	2	0,46	**9**	0,23	**16**	0,42	**23**	0,20
	3	0,38	**10**	0,40	**17**	0,48	**24**	0,36
	4	0,41	**11**	0,20	**18**	0,47	**25**	0,54
	5	0,44	**12**	0,36	**19**	0,29		
	6	0,42	**13**	0,44	**20**	0,42		
	7	0,28	**14**	0,29	**21**	0,29		
2	**1**	0,33	**7**	0,28	**13**	0,49	**19**	0,33
	2	0,39	**8**	0,45	**14**	0,37	**20**	0,36
	3	0,46	**9**	0,21	**15**	0,35	**21**	0,32
	4	0,47	**10**	0,35	**16**	0,29	**22**	0,50
	5	0,30	**11**	0,48	**17**	0,33	**23**	0,41
	6	0,32	**12**	0,38	**18**	0,36		

Madde toplam korelasyonunda korelasyon değerlerinin 0,30 ve üzerine olması o maddenin bireyi iyi derecede ayırt ettiği, 0,20-0,30 arasında olması ise o maddenin zorunlu olduğunda teste alınabileceği anlamına gelir. 0,20'den düşük maddelerin testten çıkarılması gerekir (Büyüköztürk, 2017, s.183). Tablo 4 incelendiğinde, 1. kitapçıktaki madde- toplam korelasyonlarının *r=0*,20 ile *r=0*,54 arasında ve son olarak 2. kitapçıktaki maddelerin madde –toplam korelasyonlarının *r=0*,21 ile *r= 0*,50 arasında değiştiği görülmektedir. Sonuç olarak 0,20' den küçük değeri olan madde bulunmadığından analize madde çıkartılmadan devam edilmiştir.

İki kitapçıktan rastgele seçilen 500, 1000 ve 3000 kişilik veri setleri MTK varsayımları açısından incelenmiştir. Tek boyutluluk için R 4.1.1 programı "psych" (Revelle, 2012) paketi kullanılarak tetrakorik AFA yapılmıştır. Yerel bağımsızlık açısından ise tek boyutluluk sonucu kanıt olarak kullanılmıştır. Analizlere ait sonuçlar Tablo 5'te sunulmuştur.

Tablo 5. Tetrakorik Korelasyona Dayalı AFA Uyum İndeksleri

Kitapçık	Uyum İndeksleri					
	Örneklem (N)	P değeri	χ^2/sd	Faktör yükleri	RMSEA	TLI
1	500	0.00	1.52	0.29-0.77	0.03	0.95
	1000	0.00	2.07	0.21-0.78	0.05	0.97
	3000	0.00	4.03	0.26-0.76	0.04	0.98
2	500	0.00	1.58	0.27-0.78	0.03	0.95
	1000	0.00	2.31	0.26-0.71	0.06	0.96
	3000	0.00	4.99	0.29-0.73	0.04	0.98

Tablo 5 incelendiğinde, p değerlerinin hepsinin 0,05'den küçük olduğu ve χ^2/sd değerlerinin hepsinin 5'den küçük olduğu belirlenmiştir. Ek olarak CFI ve TLI değerlerinin 0,93'den büyük olduğu ve RMSEA değerlerinin 0,07'den küçük olduğu görülmektedir. Madde faktör yüklerine bakılacak olursa minimum değerin 0,21 ve maksimum değerin 0,78 olduğu tespit edilmiştir. Maddelerin faktör yük değerlerinin 0,32 den büyük olması gerekir (Büyüköztürk,) ancak madde-toplam korelasyonu için 0,20-0,30 arasının da kabul edilebilir olduğu ve uyum indekslerinin iyi çıkması nedeniyle maddeler ölçekten çıkartılmamıştır. Bu konu ile ilgili olarak Schermelleh-Engel, Moosbrugger ve Müller (2003) de çalışmasında örneklem büyüklüğü 250'den fazla ve madde sayısı 12 ve 30 arasında olduğu durumlar için kabul edilebilir uyumu kriterleri olarak χ^2/sd değerinin 5'den küçük, TLI değerinin 0,90'dan büyük ve RMSEA değerinin 0,10'den küçük ve 0,05'den büyük olması gerektiğini belirtmiştir. Bu bilgi ışığında model uyum indekslerin de genel olarak modelin veriyle uyumlu olduğu yorumu yapılabilir. Diğer bir deyişle kitapçıklara ait veri setleri ile tek boyutlu model arasındaki uyumun iyi olduğu yorumu yapılabilir.

Ayrıca yerel bağımsızlık için ise alanyazında (Embretson ve Reise, 2000, s.48; Hambleton vd., 1991, s.11) tek boyutluluk varsayımının karşılanmış olmasının, kanıt olarak sunulabildiği belirtilmektedir.

Standart Hataların Kestirilmesinde Kullanılan Yöntemler

Bootstrap yönteminde standart hataların hesaplanması aşağıda açıklanan 5 basamakta gerçekleşmiştir (Tsai, vd., 2001):

1. Orijinal iki formdan randum olarak bootstrap veri setleri çekilmiştir.
2. İki formun madde parametre kestirimleri için MTK kalibrasyonları yapılmıştır.
3. Elde edilen ortak madde parametreleri kullanılarak iki form için MTK ölçek dönüştürme katsayıları (A ve B) hesaplanmıştır.
4. Birinciden üçüncüye kadar olan basamaklar birçok kez tekrar edilmiştir (replikasyon sayısı kadar)
5. Dördüncü basamakta elde edilen eşitlenmiş puanların standart sapması hesaplanarak MTK bootstrap eşitleme hataları elde edilmiştir.

Bu çalışmada MTK gerçek ve gözlenen eşitlemeleri ölçek dönüştürme yöntemlerine göre analiz etmek yukarıdaki basamaklar R4.1.1 programının "EquateIRT" (Battauz, 2015), paketi kullanılarak yapılmıştır. Replikasyon sayısı ile ilgili olarak Efron ve Tibshirani (1994, s.50) standart hata kestirimini replikasyon sayısına göre incelemiştir. Bu çalışmada 25 replikasyonun bilgilendirici olduğu, 50 replikasyon sayısının genel olarak yeterli bir standart hata kestirimi verdiğini ve 200 ve üstü replikasyonun çok nadir ihtiyaç duyulduğunu ancak replikasyon sayısı arttıkça kestrimlerin doğruluğunun artabileceğini belirtilmiştir. Bu çalışmada 500 replikasyon kullanılmıştır.

Delta yöntemi için ise standart hataların hesaplanması R 4.1.1 programı "EquateIRT" paketi aracılığıyla yapılmıştır. MTK eş değer olmayan gruplar için ortak madde deseni için standart hata formülü aşağıdaki gibidir:

$$SEE(\hat{\eta}) = \sqrt{\frac{\partial \eta}{\partial \hat{\beta}}\, acov(\hat{\beta}) \frac{\partial \eta}{\partial \beta}} \qquad (1)$$

Formül 1'de $\frac{\partial \eta}{\partial \beta}$ = elementleri ölçek dönüştürme katsayıları (A ve B) ve madde parametreleri kullanılarak elde edilen puanların kısmi türevleri olan bir vektördür.

$acov(\hat{\beta})$= ölçek dönüştürme katsayıları ve madde parametre kestirimlerinin varyans-kovaryans matrisidir (Osawaga, 2001; Battauz, 2015).

BULGULAR

"Gerçek puan eşitlemede örneklem büyüklüğüne ve MTK eşitlemede kullanılan ölçek dönüştürme yöntemlerine göre delta ve bootstrap yöntemlerinin ürettiği hatalar nasıl değişmektedir?" Araştırma Sorusuna İlişkin Bulgular

Gerçek puan eşitleme sonrasında elde edilen hata ortalamaları Tablo 7'de sunulmuştur.

Tablo 7. Değişkenlere Göre Gerçek Puan Eşitleme Hataları

Örneklem büyüklüğü	Ölçek dönüştürme	Bootstrap	Delta
500	Stocking Lord	0.03	0.34
	Ort.-ort.	0.25	0.71
	Ort.-s.sapma	0.20	0.52
	Haebara	0.05	0.34
1000	Stocking Lord	0.01	0.20
	Ort.-ort.	0.04	0.40
	Ort.-s.sapma	0.06	0.32
	Haebara	0.01	0.20
3000	Stocking Lord	0.01	0.11
	Ort.-ort.	0.01	0.22
	Ort.-s.sapma	0.02	0.23
	Haebara	0.01	0.11

Tablo 7'de görüldüğü üzere gerçek puan eşitleme hatalarında genel olarak bootstrap yöntemi ile elde edilen hataların daha düşük değerler olduğu belirlenmiştir. Ayrıca beklendiği gibi örneklem büyüklüğü arttıkça hatalar küçülmektedir.

Ölçek dönüştürme yöntemlerine göre ise bootstrap hataları için 500, 100 ve 3000 kişilik örneklemde en düşük hatanın Stocking Lord yönteminde olduğu görülebilir. Delta yönteminde ise 500, 100 ve 3000 kişilik örneklemde de Stocking Lord, yöntemin en düşük hataları üretmiştir. En yüksek hatalar ise 500 kişilik örneklem için bootstrap ve delta yönteminde ortalama –ortalama yöntemi kullanılarak elde edilmiştir. 1000 kişilik örneklem için bootstrap yönteminde ortalama- standart sapma ve delta yönteminde ortalama- ortalama yöntemlerinde en yüksek hatalar görülmektedir. Son olarak 3000

kişilik örneklem için her iki yöntemde de ortalama- standart sapma yöntemi en yüksek hata değerini vermiştir.

"Gözlenen puan eşitlemede örneklem büyüklüğüne ve MTK eşitlemede kullanılan ölçek dönüştürme yöntemlerine göre delta ve bootstrap yöntemlerinin ürettiği hatalar nasıl değişmektedir?" Araştırma Sorusuna İlişkin Bulgular

Gözlenen puan eşitlemeden elde edilen eşitleme hata ortalamaları Tablo 8'de sunulmuştur.

Tablo 8. Değişkenlere Göre Gözlenen Puan Eşitleme Hataları

Örneklem büyüklüğü	Ölçek dönüştürme	Bootstrap	Delta
500	Stocking Lord	0.01	0.23
	Ort.-ort.	0.24	0.63
	Ort.-s.sapma	0.20	0.40
	Haebara	0.04	0.24
1000	Stocking Lord	0.01	0.16
	Ort.-ort.	0.04	0.38
	Ort.-s.sapma	0.06	0.29
	Haebara	0.01	0.16
3000	Stocking Lord	0.00	0.10
	Ort.-ort.	0.01	0.21
	Ort.-s.sapma	0.02	0.21
	Haebara	0.00	0.10

Tablo 8 incelendiğinde genel olarak bootstrap yöntemi ile elde edilen hataların daha düşük olduğu görülmektedir. Beklendiği gibi hataların örneklem büyüklüğü arttıkça azaldığı gözlemlenmiştir.

Ölçek dönüştürme yöntemleri göz önüne alındığında bootstrap yöntemi için 500 ve 1000 kişilik örneklemlerde Stocking Lord ve 3000 kişilikte ise Haebera yöntemlerinin en düşük hataları ürettiği belirlenmiştir. Delta yöntemi için ise 500 ve 3000 kişilik örneklemde bootstrap yöntemindekiyle aynı şekilde Stocking Lord, 1000 kişilik örneklemde ise Haebera yöntemleri en düşük hata değerlerini ürettiği görülmektedir. En yüksek hata üreten yöntemler ise bootstrap yöntemi için 500 kişilik örneklem için

ortalama- ortalama yöntemi iken 1000 ve 3000 kişilik örneklemler için otalama-standart sapmadır. Delta yönteminde 500 ve 1000 kişilik örneklemlerde ortalama-ortalama ve 3000 kişilik örneklemde ise ortalama- standart sapma yöntemi en yüksek hata değerini üretmektedir.

Bootstrap Yönteminde Gözlenen ve Gerçek Puan Eşitleme Hatalarının Karşılaştırılması

Yapılan analizin sonuçları Tablo 9'da raporlanmıştır.

Tablo 9. Gerçek ve Gözlenen Puan Eşitlemeye göre Bootstrap Hatalar

Örneklem büyüklüğü	Ölçek dönüştürme	Gerçek puan eşitleme	Gözlenen puan eşitleme
500	Stocking Lord	0.03	0.01
	Ort.-ort.	0.25	0.24
	Ort.-s.sapma	0.20	0.20
	Haebara	0.05	0.04
1000	Stocking Lord	0.01	0.00
	Ort.-ort.	0.04	0.04
	Ort.-s.sapma	0.06	0.06
	Haebara	0.01	0.01
3000	Stocking Lord	0.01	0.00
	Ort.-ort.	0.01	0.01
	Ort.-s.sapma	0.02	0.02
	Haebara	0.01	0.00

Tablo 9 incelendiğinde gerçek ve gözlenen puan eşitleme hatalarının birbirine yakın olduğu, ancak genel olarak gözlenen puan eşitlemedeki hataların gerçek puan eşitlemedekilere göre daha düşük olduğu belirlenmiştir. Sadece 1000 kişilik örneklem için ortalama –ortalama yönteminde gerçek puan eşitleme hatasının gözlenenden düşük olduğu tespit edilmiştir.

Delta yönteminde gözlenen ve gerçek puan eşitleme hatalarının karşılaştırılması

Analiz sonuçları Tablo 10'da sunulmuştur.

Tablo 10. Gerçek ve Gözlenen Puan Eşitlemeye göre Delta Hatalar

Örneklem büyüklüğü	Ölçek dönüştürme	Gerçek puan eşitleme	Gözlenen puan eşitleme
500	Stocking Lord	0.33	0.23
	Ort.-ort.	0.71	0.63
	Ort.-s.sapma	0.52	0.40
	Haebara	0.34	0.24
1000	Stocking Lord	0.20	0.16
	Ort.-ort.	0.40	0.38
	Ort.-s.sapma	0.32	0.29
	Haebara	0.20	0.16
3000	Stocking Lord	0.11	0.10
	Ort.-ort.	0.22	0.21
	Ort.-s.sapma	0.23	0.21
	Haebara	0.11	0.10

Tablo 10 göz önüne alındığında genel olarak bootstrap yönteminde olduğu gibi gözlenen puan eşitleme hatalarının gerçek puan eşitlemedekilerden daha düşük olduğu söylenebilir.

TARTIŞMA ve SONUÇ

Çalışmada TIMMS 2015 4. sınıf matematik verisi kullanılarak bootstrap ve delta yöntemleri tarafından elde edilen MTK gözlenen ve gerçek puan eşitleme hataları farklı örneklem büyüklüğü ve ölçek dönüştürme yöntemleri gibi değişkenlere göre incelenmiştir. Ek olarak da kullanılan ölçek dönüştürme yöntemleri gerçek ve gözlenen puan eşitlemelerden elde edilen hatalar bakımından karşılaştırılmıştır. Çalışma sonucunda gerçek ve gözlenen puan eşitlemenin her ikisinde de bootstrap yönteminde delta yöntemine göre daha düşük hata değerleri elde edildiği belirlenmiştir. Ölçek dönüştürme yöntemlerine göre gerçek puan eşitlemede bootstrap ve delta yönteminde her örneklem büyüklüğünde en düşük hatanın Stocking Lord yönteminde olduğu görülmektedir. Aynı şekilde gözlenen puan eşitlemede de genel olarak en düşük hata Stocking Lord yöntemindeyken, bootstrap yönteminde 3000 kişilik ve delta yönteminde 1000 kişilik örneklemlerde Haebara yöntemindedir. Bootstrap ve delta yöntemlerinde genel olarak gözlenen puan eşitlemede daha düşük hatalar üretilirken, sadece bootstrap yönteminde 1000 kişilik örneklemde ortalama- ortalama yönteminde gerçek puan eşitleme daha düşük hata üretmektedir. Genel olarak gözlenen puan eşitleme hataları hem delta hem de bootstrap yöntemlerinde her dört ölçek dönüştürme yönteminde de gerçek puan eşitleme hatalarından düşüktür.

Tüm örneklem büyüklüklerinde, bootstrap yöntemiyle elde edilen hataların delta yönteminde elde edilen hatalardan düşük olduğu sonucuna ulaşılmıştır. Bu sonuç Zeng (1991) tarafından yapılan çalışmada bulunan delta ve bootstrap yöntemlerinden elde edilen hataların yakın olduğu sonucu ile çelişmektedir.

Ölçek dönüştürme yöntemleri ile ilgili olarak bu çalışmada her örneklem büyüklüğünde bootstrap ve delta yöntemleri kullanıldığında karakteristik eğrisi yöntemlerinin (Stocking Lord ve Haebara) moment yöntemlerinden daha düşük hatalar ürettikleri belirlenmiştir. Baker ve Al- Karni (1991) ve Hanson ve Beguin (2002) çalışmalarında da bu bulguyla benzer sonuçlar elde etmiştir. Ek olarak Ogasawara (2001) de 2 ve 3 parametreli modelleri inceleyerek aynı doğrultuda sonuçlara ulaşmıştır. Yine benzer

olarak Zhang (2020) de çalışmasında her simülasyon durumunda ve gerçek veri setinde karakteristik eğri yöntemleri kullanılarak elde edilen hataların moment yöntemlerine göre daha düşük olduğu sonucuna ulaşmıştır. Sonuç olarak bu konuda daha doğru ve kararlı eşitleme sonuçları verebileceğinden karakteristik eğrisi yöntemlerinin moment yöntemlere göre daha tercih edilebilir olduğu sonucuna ulaşılmıştır (Baker ve Al-Karni, 1991; Hanson ve Beguin, 2002; Kim ve Kolen, 2007; Ogasawara, 2001; Zhang, 2020).

Way ve Tang (1991) tarafından yapılan çalışmada ortalama-standart sapma yönteminin ortalama-ortalama ve Stocking Lord yöntemlerinden daha az kararlı kestirimler sağladığı bulunmuştur. Bu çalışmada ise aksine gerçek ve gözlenen puan eşitlemede 500 kişilik örneklemde bootstrap ve delta yöntemleri için ve 1000 kişilik örneklemde delta yönteminde ortalama – ortalama yönteminin ortalama – standart sapmaya göre daha büyük hata değerleri ürettiği görülmüştür. Diğer yandan bu bulguyla benzer olarak bu çalışmada 1000 kişilik örneklemde bootstrap ve 3000 kişilik örneklemede delta ve bootstrap hataları incelendiğinde ortalama – standart sapma yönteminin ortalama-ortalama yöntemine göre daha yüksek hatalara sahip olduğu sonucuna ulaşılmıştır. Bu durumu açıklamak için Baker ve Al-Karni (1991) ortalamaların standart sapmalara göre daha kararlı parametreler olması dayanak gösterilerek ortalama – ortalama yönteminin ortalama- standart sapmaya göre daha iyi kestirimler verdiğini raporlamıştır.

Ayrıca Ogasawara (2001) tarafından yapılan çalışmada da en iyi yöntemin Stocking Lord olduğu bulgusu, bu çalışmada gerçek puan eşitlemede delta ve bootstrap yöntemlerinde her örneklem büyüklüğünde en düşük hata değeri veren yöntemin Stocking Lord olmasıyla tutarlıdır. Ancak bu durumun aksine bu çalışmada gözlenen puan eşitlemede 1000 kişilik örneklem delta yönteminde ve 3000 kişilik örneklem bootstrap yönteminde en düşük hata değeri üreten çalışmanın Haebara olduğu belirlenmiştir.

Sonuç olarak Gök (2012) çalışmasında bulunan sonuca benzer olarak bu çalışmada da yöntemlerin farklı koşullar altında farklı sonuçlar verdiği bulunmuştur. Gök (2012) çalışmasında bu çalışmada da olduğu gibi farklılaşmanın miktarının değişkenlere göre değişmekte olduğunu ve değişkenlere bağlı olduğunu belirtmiştir.

Elde edilen veriler ışığında delta ve bootstrap yöntemleriyle elde edilen hataların araştırma koşullarına ve kullanılan verilere göre değişmekte olduğu görülmektedir. Bu doğrultuda Gök ve Kelecioğlu (2014) da çalışmalarında çalışma sonuçları farklı koşul ve durumlarda diğer yöntemlerden üstün olan tek bir yöntem olmadığı gibi, hangi yöntemi seçmenin en iyi sonuçları vereceği konusunda da açık bir kanıtın olmadığını; fakat eşitleme çalışmalarında elde edilen sonuçların daha önceki çalışmaların sonuçlarıyla tutarlılığı bir yöntemi seçmek için en bilgilendirici yol olduğu belirtmiştir.

Bu çalışma uygulayıcı ve araştırmacıların ilgilendikleri ve karşılarına çıkan gerçek veri setlerinden biri olan TIMMS verisi ile yapılmıştır. Gerçek veri doğası gereği karmaşık olabilir ve kontrol edilemeyen bazı değişkenler içerebilir. Bu çalışmada değişkenleri kontrol edebilmek amacıyla başarı ortalaması etrafında olan benzer yetenek düzeyine sahip veri setleri analize dâhil edilmiş ve 14 kitapçık arasından model uyum indekslerinin hepsinin kabul edilebilir düzeyde olduğu kitapçık çifti seçilmiştir. Bu doğrultuda gelecek çalışmalarda belirlenen şartlara göre üretilen simülasyon veri ile de analizlerin sınanması önerilmektedir.

Çalışmada kullanılan replikasyon sayısı, formlar arasındaki güçlük farkı, birey yetenek düzeyleri ve zaman değişkenleri de araştırmada kontrol edilen değişkenlere eklenerek çalışma geliştirilebilir. Çalışma dikey eşitleme yapılarak yani yıllar arasındaki veriler kullanılarak da yapılabilir. İncelenen hata kestirim yöntemlerine çoklu veri atama (multiple imputation) da dâhil edilerek alanyazına katkı sağlayan araştırmalar yapılabilir.

KAYNAKLAR

Aksekioğlu, B. (2017). Madde tepki kuramına dayalı test eşitleme yöntemlerinin karşılaştırılması: PISA 2012 fen testi örneği (Yayınlanmamış yüksek lisans tezi). Akdeniz Üniversitesi, Eğitim Bilimleri Enstitüsü. Mersin. Retrieved from http://acikerisim.akdeniz.edu.tr/xmlui/handle/123456789/3143.

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

Akın-Arıkan, Ç. (2017). Kernel eşitleme ve madde tepki kuramına dayalı eşitleme yöntemlerinin karşılaştırılması, (Yayınlanmamış doktora tezi). Hacettepe Üni&rsitesi, Eğitim Bilimleri Enstitüsü. Ankara. Retrieved from http://www.openaccess.hacettepe.edu.tr:8080/xmlui/handle/11655/3902

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, DOI: 10.1111/j.1745-3984.1991.tb00350.x

Battauz, M. (2013). IRT test equating in complex linkage plans. Psychometrika, 78(3), 464–480. DOI:10.1007/s11336- 012-9316-y

Battauz, M. (2015). equateIRT: An R package for IRT test equating. *Journal of Statistical Software*, *68*(1), 1-22. Retrieved from https://www.jstatsoft.org/article/view/v068i07

Bozdağ, S., & Kan, A. (2010). Şans başarısının test eşitlemeye etkisi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *39*(39), 91-108. Retrieved from https://dergipark.org.tr/en/pub/hunefd/issue/7799/102166

Büyüköztürk, Ş. (2017). Sosyal bilimler için veri analizi el kitabı (23. Baskı). Ankara: Pegem Akademik Yayıncılık.

Büyüköztürk, Ş., Çakmak, K. E., Akgün, E. Ö., Karadeniz, Ş., & Demirel, F. (2014). *Bilimsel araştırma yöntemleri.* Ankara: Pegem Akademi.

Byrne, B. M. (2013). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. routledge. DOI: 10.4324/9780203807644

Cook, L. L., & Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practice, 10* (3), 37–45, DOI: 10.1111/j.17453992.1991.tb00207.x

Çağlak,S. (2015). The use of a Meta-Analysıs Technıque in Equatıng and Its Comparıson wıth Several Small Sample Equatıng Methods. Doktora Tezi, Florida Universitesi,USA.
Retrieved from https://www.proquest.com/docview/1759157142?pq-origsite=gscholar&fromopenview=true

Demir, S., & Güler, N. (2014). Study of test equating on the common item non-equivalent group design Ortak maddeli denk olmayan gruplar desenine ilişkin test eşitleme çalışması. *Journal of Human Sciences*, *11*(2), 190-208. Retrieved from https://www.j-humansciences.com/ojs/index.php/IJHS/article/view/2870

Demirus, K. B., & Gelbal, S. (2016). Ortak maddelerin değişen madde fonksiyonu gösterip göstermemesi durumunda test eşitlemeye etkisinin farklı yöntemlerle incelenmesi. *Journal of Measurement and Evaluation in Education and Psychology*, *7*(1), 182-201. DOI: 10.21031/epod.56218

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Embretson, S. E. V., & Reise, S. P. (2000). *Item response theory for psychologists*. New Jersey: Lawrence Erlbaum Associates, Inc.

Felan, G. D. (2002). *Test equating: Mean, linear, equipercentile and item response theory*. Paper presented at the annual meeting of the Southwest Educational Research Association (Austin, TX, February 14-16, 2002).

Gök,B. (2012). *Denk olmayan gruplarda ortak madde deseni kullanılarak madde tepki kuramına dayalı eşitleme yöntemlerinin karşılaştırılması* (Yayımlanmamış Doktora tezi). Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara. Retrieved from http://www.openaccess.hacettepe.edu.tr:8080/xmlui/bitstream/handle/11655/1767/c8ab915b-7862-4913-a87f-4195e382ef66.pdf?sequence=1

Gök, B. , & 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. Retrieved from https://dergipark.org.tr/en/pub/mersinefd/issue/17393/181786?publisher=mersin?publisher=mersin?publisher=mersin?publisher=mersin

Gül, E., Gül, Ç. D., Bökeoğlu, Ö. Ç., & 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. DOI: 10.17240/aibuefd.2017.17.32772-363973

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. USA: Sage.

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. DOI: 10.1177%2F0146621602026001001

Kilmen, S. (2010). *Madde tepki kuramına dayalı test eşitleme yöntemlerinden kestirilen eşitleme hatalarının örneklem büyüklüğü ve yetenek dağılımına göre*

karşılaştırılması. Yayımlanmamış Doktora Tezi, Ankara: Ankara Üniversitesi Eğitim Bilimleri Enstitüsü. Retrieved from https://dspace.ankara.edu.tr/xmlui/handle/20.500.12575/34147

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. DOI: 10.3102/1076998607302632

Kolen, M. J. (1988). Traditional Equating Methodology. *Educational Measurement: Issues and Practice, 7* (4), 29-36. DOI: 10.1111/j.1745-3992.1988.tb00843.x

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scalling and linking* (Second Edition). USA: Springer.

Li, D., Jiang, Y., & von Davier, A. A. (2012). The Accuracy and Consistency of a Series of IRT True Score Equatings. *Journal of Educational Measurement*, *49*(2), 167–189. Retrieved from http://www.jstor.org/stable/41653582

Livingstone, S. A. (2004). Equating test scores (Without IRT). Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/LIVINGSTON.pdf

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Erlbaum.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the rasch model. Journal of Educational Measurement, 17(3), 179–193. DOI:10.1111/j.1745-3984.1980.tb00825.x

MEB (2011). *TIMSS 2007 ulusal matematik ve fen raporu 8. Sınıflar*. Ankara: Hermes. Retrieved from https://timss.meb.gov.tr/www/raporlar/icerik/3

Michaelides, M. P. (2003). *Effects of common-item selection on the accuracy of item response theory test equating with nonequivalent groups*. Stanford University. Retrieved from https://www.proquest.com/docview/305304806?pq - origsite=gscholar&fromopenview=true

Muthén, L. K., & Muthén, B. O. (2010). 1998–2010 Mplus user‘s guide. *Muthén and Muthén*, 39-49. Retrieved from https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf

Ogasawara, H. (2001). Standard Errors of Item Response Theory Equating/Linking by Response Function Methods. *Applied Psychological Measurement, 25* (1), 53–67. DOI: 10.1177%2F01466216010251004

Öztürk, N., & Anıl, D. (2012). Akademik personel ve lisansüstü eğitimi giriş sınavı puanlarının eşitlenmesi üzerine bir çalışma. *Eğitim ve Bilim*, *37*(165). Retrieved from http://eb.ted.org.tr/index.php/EB/article/view/1141

Popham, W.J. (1999). "Why standardized tests don't measure educational quality". *Educational Leadership.* 56 (6): 8–15.

Patton, J. M., Cheng, Y., Yuan, K. H., & Diao, Q. (2014). Bootstrap standard errors for maximum likelihood ability estimates when item parameters are unknown. Educational and Psychological Measurement, 74(4), 697–712. DOI: 10.1177/0013164413511083

Revelle, W. (2012). Procedures for psychological, psychometric, and personality research. *Acesso em*, *9*. Retrieved from https://cran.r-project.org/web/packages/psych/psych.pdf

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of psychological research online*, *8*(2), 23-74. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.509.4258&rep=rep1&type=pdf

Tekin, H. (2009). *Eğitimde ölçme ve değerlendirme*. Ankara: Yargı.

Tsai, T. H., Hanson, B. A., Kolen, M. J., & Forsyth, A. 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

Turgut, M. F. , & Baykul, Y. (2012). *Eğitimde ölçme ve değerlendirme* (Dördüncü Baskı). Ankara: Pegem Akademi.

Uyar, Ş., Aksekioğlu, B., & Öztürk Gübeş, N. (2020). Comparison of Different Scale Linking Methods in PISA 2012 Mathematics Literacy Test. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi, 46,* 121-148. Retrieved from https://acikerisim.mehmetakif.edu.tr/xmlui/handle/11672/3258.

Velhelst, N. D. (2004). Reference Supplemented to The Preliminary Pilot Version of The Manuel for Relating Language Examinations to the Common European Framework Of Reference for Languages: Learning, Teaching, Assessment, Section G: Item Response Theory. DGIV/EDU/LANG (2004) 13, Council of Europe Language Policy Division, Strasbourg.

Way, W. D., & Tang, K. L. (1991). *A Comparison of Four Logistic Model Equating Methods.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Zeng, L. (1991). Standard Errors of Linear Equating for the Single-Group Design. Retrieved from https://www.act.org/content/dam/act/unsecured/documents/ACT_RR91-04.pdf

Zhang, Z. (2020). Estimating standard errors of IRT true score equating coefficients using imputed item parameters. *The Journal of Experimental Education*, 1-23. DOI: 10.1080/00220973.2020.1751579

Zhang, Z., & Zhao, M. (2019). Standard errors of IRT parameter scale transformation coefficients: Comparison of bootstrap method, delta method, and multiple imputation method. *Journal of Educational Measurement*, *56*(2), 302-330. DOI: 10.1111/jedm.12210

ORCID

Rana SALMANER DOĞAN https//orcid.org/0000-0003-1307-3826
Şeref TAN https//orcid.org/0000-0002-9892-3369

SUMMARY

Standardized tests are generally large-scale tests that measure students' educational and basic abilities development in different fields (Turgut & Baykul, 2012). One of the difficulties encountered in the application of these tests is the necessity of converting the scores, obtained from multiple forms of tests applied for the same purpose every year or more than once a year, into a comparable situation in order to make valid and consistent decisions about evaluating students'development or selecting them (Kilmen, 2010). The statistical method used to overcome this difficulty and adjust tests' difficulty differences is called "test equating" (Kolen & Brennan, 2004: 2).

Considering the advantages of equating methods, it prevents possible injustices against individuals who take the difficult test and eliminates bias problems resulting from test forms (Angoff, 1971; Cook & Eignor, 1991; Hambleton, Swaminathan, & Rogers, 1991). Another benefit of the method is that it allows parameters of a lot of items in a test to be calculated in a single session. In some sessions, the number of items in the test may be too many to be applied to every individual. In this case, each individual in the respondent group, where the questions are applied, can respond to question sets and common items according to a certain test pattern (Verhelst, 2004). Apart from these benefits, test equation still has some limitations. It can be said that the most common of these limitations are equation errors.

In this study, it was aimed to compare the delta and bootstrap methods used to determine standard errors in IRT test equation according to different sample sizes and scale conversion methods. As the study aims to examine a method in a more comprehensive way and according to different variables, it is a descriptive study. 500, 1000 and 3000 people samples chosen randomly from the data of the countries above and below the midpoint of the TIMSS achievement scale (500) (Australia, Canada, Italy, Spain, Croatia, the Slovak Republic, New Zealand, Turkey and Georgia) were used in the study. In order to determine which, one would be used among the 14 booklet types used in the TIMMS exam, the IRT assumptions were examined, and it was chosen the pair of booklets (booklets 1 and 2). In addition, in the literature, it was stated that the result of testing unidimensionality hypothesis can be used as evidence for local independence (Embretson & Reise, 2000, p.11.). The item-test correlation was calculated so as to be evidence of internal reliability. It was found that the lowest correlation was 0.20 and the highest correlation was 0.50. b parameters of the items in the booklets and their averages were calculated. The analyses were continued with the 1st and 2nd booklets, which have all acceptable model fit indices. Three samples obtained from Booklets 1 and 2 were examined in terms of IRT assumptions. Bootstrap error estimation method was applied to randomly selected data by using R 4.1.1 software. Then, using suitable formula for the pattern in the delta method with the same data, the errors were estimated by using the same program.

As a result of the analysis, it was determined that the errors obtained by the bootstrap method were generally lower in both score equation methods. In both score equation methods, it was found that lower error values were generally obtained by Stocking Lord method. Comparing true and observed errors in bootstrap method, the error values were generally close to each other, but

the observed score equalization produced lower errors than the true score. In line with what Kolen (1988) and Kilmen (2010) found in their studies, the standard error values decrease as the sample size increases in this study. It contradicts with the result that the errors obtained from the delta and bootstrap methods found in the study conducted by Zeng (1991) are close.