Uludağ Üniversitesi Eğitim Fakültesi Dergisi / Journal of Uludag University Faculty of Education
e-ISSN: 2667-6788 https://dergipark.org.tr/tr/pub/uefad
Yıl **2023,** Cilt **36**, Sayı **1**, **23**-**50** | DOI: https://doi.org/10.19171/uefad.1231069 **Araştırma Makalesi | Research Article**

Test Eşitlemede Yerel Bağımsızlık Varsayımının İhlalinin Delta ve Bootstrap Eşitleme Hatalarına Etkisinin Çeşitli Değişkenlere Göre İncelenmesi[1]

Investigation of the Effect of Violation of the Local Independence Assumption in Test Equating on Delta and Bootstrap Equating Errors According to Various Variables

Mehmet Fatih DOĞUYURT
Doktora Öğrencisi ◆ Gazi Üniversitesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı ◆ doguyurtmfatih@gmail.com ◆ ORCID: 0000-0001-9206-3321

Şeref TAN
Prof. Dr. ◆ Gazi Üniversitesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı ◆ seraftan@gazi.edu.tr ◆ ORCID: 0000-0002-9892-3369

Özet

Bu araştırmada doğrusal, eşit yüzdelikli ve polinominal loglinear öndüzgünleşirilmiş eşit yüzdelikli test eşitleme yöntemlerinde hataların belirlenmesinde kullanılan delta ve bootstrap eşitleme hatası kestirme yöntemlerinin örneklem büyüklüğü, madde sayısı ve ikinci boyuta yüklenen madde yüzdesi değişkenleri bakımından incelenmesi amaçlanmıştır. Araştırmada eşitleme yöntemleri farklı koşullarda simülasyon verileri ile kontrollü olarak karşılaştırıldığından araştırma bir simülasyon çalışması niteliğindedir. Simülasyon çalışmalarında elde edilen simülasyon verisinin gerçek yanıtları temsil etmesi için simülasyon verileri PISA 2018 matematik sınavı birinci formundan elde edilen verilere ait dağılımlarından yararlanarak üretilmiştir. Araştırmada, örneklem büyüklüğü (250, 1000, 5000), test uzunluğu (20, 40, 60), ikinci boyuta yüklenen madde oranı (%15-%30-%50) olmak üzere 36 koşul incelenmiştir. Bu koşullar altında 2PLM'ye uyumlu iki kategorili ve 100 replikasyon ile 3600 veri seti elde edilmiştir. Araştırmamızda "random gruplar deseni" kullanılmıştır. Genel olarak örneklem büyüklüğü azaldıkça elde edilen hata miktarlarında artış olduğu en az hata içeren koşulun 5000 örneklem büyüklüğü ve testte yer alan madde sayısının 20 olması durumunda elde edildiği, en iyi performansa sahip eşitleme yönteminin doğrusal eşitleme yöntemi olduğu ve eşitleme hatalarını belirlemede kullanılan yöntem olarak da delta yöntemi olduğu sonucuna ulaşılmıştır. Ayrıca testin tek boyutlu yapısının bozulup çok boyutlu olması durumunda ikinci boyuta yüklenen madde oranı bakımından elde edilen eşitleme hatalarında sistematik bir bulguya rastlanılmadığı, araştırmada ele alınan test eşitleme yöntemleri, eşitleme hatalarını belirlemede kullanılan yöntem, örneklem büyüklüğü ve testte yer alan madde sayısı koşullarına göre değişkenlik gösterdiği tespit edilmiştir.

Anahtar Sözcükler: Test eşitleme, Geleneksel Eşitleme Yöntemleri, Eşitlemenin standart hatası, Delta yöntemi, Bootstrap.

Abstract

This study aimed to examine the delta and bootstrap equating error estimation methods used in determining errors in linear, equipercentile and polynomial loglinear presmoothing equipercentile test equating methods regarding sample size, number of items, and percentage of items loaded in the second dimension. The

[1] Bu çalışma, Mehmet Fatih Doğuyurt'un Prof. Dr. Şeref Tan danışmanlığında hazırladığı doktora tezi için üretilen veriler kullanılarak hazırlanmıştır.

Kabul Tarihi / Received: 08.01.2023 | Onay Tarihi / Accepted: 16.03.2023

study is simulation research since the equating methods in the research are compared using the simulation data under different controlled conditions. The simulation data obtained in the simulation studies were produced using the distributions of the data obtained from the first form of the PISA 2018 mathematics exam to represent the actual responses. In the study, 36 conditions including sample size (250, 1000, 5000), test length (20, 40, 60), and the ratio of items loaded on the second dimension (15%, 30%, 50%) were examined. Under these conditions, 3600 data sets were obtained with two categories and 100 replications compatible with 2PLM. "Random groups design" was used in our study. It has been concluded that, in general, the number of errors increases as the number of samples decreases and that the condition with the minor error is obtained when the sample size of 5000 and the number of items in the test is 20. It has also been concluded that the equating method with the best performance in equating and determining errors is the linear equating method and the delta method, respectively. In addition, there is no systematic finding in the equating errors in terms of the rate of items loaded into the second dimension when a multidimensional structure is obtained as a result of the corruption of the one-dimensional structure of the test, and the equating errors vary according to the test equating methods discussed in the research, the method used to determine the equating errors, the sample size and the number of items in the test.

Keywords: Test equating, Traditional equating Methods, Standard error of equating, Delta method, Bootstrap

1. Giriş

Eğitimle kazandırılmak istenen yeni davranışların beklenen düzeyde kazandırılmış olup olmadığına karar verebilmek için, öncelikle bu davranışların, geçerliği ve güvenirliği yeterli derecede yüksek olan ölçme araçlarıyla ölçülmesine ihtiyaç vardır (Özçelik, 2010). Standartlaştırılmış koşullar altında bir yapının derecesini veya miktarını sayısal olarak tanımlaması amaçlanan ölçme aracı (Kane, 2006) olarak tanımlanan test, eğitim ve psikoloji alanında yaygın olarak kullanılmaktadır. Cook ve Eignor (1991) test uygulamasının birincil amacının sınava giren bir grup bireyin becerilerini mümkün olduğunca adil ve objektif bir şekilde ölçmek veya değerlendirmek için bir araç sağlamak olarak belirtmişlerdir. Yapılan geniş ölçekli testlerden elde edilecek puanların geçerliğinin ve güvenirliğinin yüksek olması için test geliştiricileri ve uygulayıcıları birçok önlem almaktadır. Bu önlemlerden biri de sınava giren öğrencilerin cevapları birbirinden kopya etmesini engellemek için birden çok test formu oluşturmaktır. Bu formlar aynı soruların farklı sırada sunulması şeklinde olabileceği gibi, aynı yapıyı ölçen ve farklı sorulardan oluşan alternatif formlar şeklinde de hazırlanabilmektedir. Ayrıca akademik kabul için önemli olan ALES, YÖKDİL ve YDS gibi sınavlar bir takvim yılı içinde birden fazla uygulanmakta ve bu sınavlardan elde edilen puanlar birkaç yıl kullanılabilmektedir. Yıl içerisinde birden fazla uygulanan bu sınavlarda aynı yapıyı ölçen farklı sorular kullanılır. Çünkü aynı test soruları birden çok kez kullanılırsa, o zaman sınava daha önce girenler, daha sonra sınava girenleri içerik ve maddelerin cevapları hakkında bilgilendirebilir veya istenen puanı elde etmek için sınava birden çok kez giren kişi, yapı(lar) hakkında artan bilgisinden dolayı değil, bunun yerine sorulara önceden maruz kaldığından veya teste girme yeteneğinin artmasından dolayı bir kez sınava girene göre daha yüksek puan alabilir. Bu nedenle, aynı sınavı kullanmak yerine, genellikle bir test planı yardımıyla aynı test özelliklerine dayalı olarak sınavın farklı formları oluşturulur (Desjardins ve Bulut, 2018). Oluşturulan bu formlar alternatif veya paralel form/test olarak adlandırılır. Paralel testler aynı yapıyı veya gizil özelliği ölçen, aynı gerçek puana ve hata varyanslarına sahip testler olarak (De Gruijter ve Leo, 2007) tanımlanmaktadır. Paralel testler öğrencilere aynı zaman diliminde uygulanabileceği gibi farklı zaman dilimlerinde de uygulanabilir. Test uygulayıcıları uygulanan bu test formlarının paralel olduğu sayıltısı ile hareket eder. Peki, gerçekten bir testin paralel formlarını oluşturmak mümkün müdür? Başka bir ifade ile belirli bir yetenek seviyesine sahip öğrencinin oluşturulan bu formlardan herhangi birini alması durumunda

öğrencinin ölçülmek istenen yeteneğinde alabileceği diğer olası formlara nazaran pozitif veya negatif yönde bir değişimin olmaması garanti edilebilir mi?

Hambleton ve diğerleri (1991) ve Aiken (2000) teorik olarak paralel testler oluşturmak mümkün olsa da pratikte bunu gerçekleştirmenin oldukça zor olduğunu belirtmişlerdir. Oluşturulan bu test formlarının psikometrik özellikler bakımından birbirine denk olması için büyük çabalar gösterilse de formlar arasında güçlük bakımından farklılıklar kaçınılmazdır. Farklı zamanlarda bu testleri alan öğrencilerin aldıkları puanlar aynı değerlendirmelerde kullanılmakta ve bu durum sınavlar ile ilişkisi olan tüm paydaşlar açısından eşitlik hissiyatına karşı bir kaygı oluşmaktadır. Bu kaygı altında yatan sorgulama ise farklı test formlarından elde edilen puanların aynı değerlendirme için kullanılmasının ne kadar adil olacağı yönündedir. Bu sorgulamanın cevabı ise şüphesiz ki test eşitlemedir. Test eşitleme, benzer içerik ve güçlük düzeyine sahip test formları arasındaki farklılıkları düzenleyerek, bu formlardan elde edilen puanların birbiri yerine kullanılmasını sağlayan istatistiksel süreçtir (Kolen ve Brennan, 2014). Test eşitleme, genellikle bir testin birden fazla formunun bulunduğu ve farklı formlardaki sınavların birbirleriyle karşılaştırıldığı veya araştırmacıların uygulama etkileri problemlerinin üstesinden gelmek istedikleri durumlarda kullanılır (Felan, 2002).

Eşitleme çalışmasında kullanılacak verinin elde edilmesi durumuna göre farklı eşitleme desenleri kullanılabilir. Desenlerin seçimi hem pratik ve hem de istatistiksel konuları içermektedir (Kolen ve Brennan, 2014). Veri toplamada kullanılan tek grup deseni, dengelenmiş tek grup deseni, random grup deseni ve eşdeğer olmayan gruplar ortak madde deseni eşitleme desenleridir (Kolen ve Brennan, 2014). Eşdeğer gruplar deseni (equivalent-groups design) olarak da bilinen random grup deseninde ortak bir evrenden gelen cevaplayıcılar test formlarına random olarak atanır ve iki grup farklı test formlarını alır (Cook ve Eignor, 1991). Bu desende grup performansları arasındaki fark, formlar arasındaki güçlük farklılıklarının doğrudan göstergesidir. Bu nedenle örneklemden kaynaklı yanlılığı önlemek için büyük ve heterojen örneklemlerle çalışmak gerekir (Livingston, 2004). Bu araştırmada random grup deseni kullanılmıştır.

Literatür incelendiğinde eşitleme yöntemleri genellikle Klasik Test Kuramı'na (KTK) ve Madde Tepki Kuramı'na (MTK) dayalı eşitleme yöntemleri olmak üzere ikiye ayrılmaktadır. Klasik test teorisinde ölçülmek istenen değişkene ait gerçek değer, gerçek puan olarak adlandırılır. Gerçek puan ölçme yoluyla doğrudan elde edilemez, bazı varsayımlarla gözlenen puanlardan kestirilir. Bu sebeple klasik test teorisine gerçek puan teorisi de denir (Baykul, 2015). 1905 yılında Spearman tarafından temelleri atılan KTK, gerçek puanı, ölçme sonuçlarından elde edilen puanla tahmin etmeye çalışan bir kuramdır (Hambleton ve Jones, 1993). Hambleton ve Jones'a (1993) göre klasik test kuramının zayıf teorik varsayımlara sahip olması bu kuramın birçok test durumuna uygulanmasını kolaylaştırmış bunun sonucu olarak da araştırmacılar tarafından yoğun bir şekilde kullanılmıştır.

Klasik Test Kuramı'na dayalı eşitleme yöntemleri; ortalama eşitleme, doğrusal eşitleme ve eşit yüzdelikli eşitleme olmak üzere üçe ayrılır (Kolen ve Brennan, 2014). Ortalama eşitleme yöntemi, geleneksel eşitleme yöntemleri arasında en az katı olanı (Sansivieri vd., 2017) ve belki de iki test formundan elde edilen puanları eşitlemeye yönelik en basit yaklaşımı ifade eder (Finch vd., 2014). Bu eşitleme yöntemi uygulanan iki test formunun sadece ortalamalarına odaklanır. Her iki formun ortalamaları eşit olması halinde iki formdan elde edilen puanların eşit olacağı varsayılır. Bu yaklaşımı kullanmak için, iki test formundaki güçlüğün tüm puan ölçeğinde sabit (veya aynı) olduğunu varsayılır (Finch vd., 2014). Doğrusal eşitleme, puan ölçeği boyunca iki form arasında sabit bir fark olduğunu varsayan ortalama eşitleme

yönteminin aksine (Sunnassee, 2011) iki test formu arasındaki zorluk farklılıklarının puan ölçeği boyunca değişmesine izin veren (Kolen ve Brennan, 2014) bir yöntemdir. Örneğin, doğrusal eşitleme, Form X'in düşük yetenek seviyesindeki öğrenciler için Form Y'den daha zor olmasına ancak yüksek yetenek seviyesindeki öğrenciler için ise daha kolay olmasına izin verir (Kolen ve Brennan, 2014). Doğrusal eşitleme, ortalamalar ve standart sapmalardaki farklılıklar dışında, X ve Y formlarındaki puanların dağılımlarının aynı olduğu varsayımına dayanır (Crocker ve Algina, 2008). Doğrusal eşitlemede, ortalamalara eşit standart sapma uzaklıkta bulunan puanlar eşit olarak ayarlanır (Kolen ve Brennan, 2014). Eğer bu doğruysa, aynı *z* puanına sahip biri X formunda diğeri Y formunda olmak üzere tanımlanan puan çifti eşdeğer puanlar olarak belirlenebilir (Crocker ve Algina, 1986).

$\sigma(X)$ ve $\sigma(Y)$'yi sırasıyla Form X ve Form Y puanlarının standart sapmaları olarak tanımlanırsa doğrusal dönüşüm denklemi aşağıdaki gibi tanımlanır (Kolen ve Brennan, 2014):

$$\frac{x-\mu(X)}{\sigma(X)}=\frac{y-\mu(Y)}{\sigma(Y)}$$

X formu yukarıda verilen denklem kullanılarak Y formu için düzenlendiğinde, eşitlik aşağıda belirtilen şekilde tanımlanabilir:

$$ly(x)=y=\sigma(Y)\left[\frac{x-\mu(X)}{\sigma(X)}\right]+\mu(Y)$$

Burada $ly(x)$, Form X'te gözlemlenen puanları Form Y ölçeğine dönüştürmek için doğrusal dönüşüm denklemidir. Terimleri yeniden düzenleyerek, $ly(x)$için alternatif bir ifade aşağıdaki şekilde belirtilebilir:

$$ly(x)=y=\frac{\sigma(Y)}{\sigma(X)}x+\left[\mu(Y)-\frac{\sigma(Y)}{\sigma(X)}\mu(X)\right]$$

Bu denklem, eğim ve sabit biçiminde doğrusal bir denklem olarak ifade edilir. Eğim, standart sapmaların oranı ve sabit terimi ise köşeli parantez içindedir (Kolen ve Hendrickson, 2013).

Eşit yüzdelikli eşitleme yöntemi, geleneksel eşitleme yöntemleri arasında en az varsayım gerektiren buna rağmen diğer yöntemlere kıyasla daha karmaşık olan bir yöntemdir (Finch vd., 2014). Bu yöntem en genel tanımla, iki formdaki hangi puanların aynı yüzdelik sıralamasına sahip olduğunu belirlemeyi içerir (Crocker ve Algina, 1986). Kavramsal olarak, eşit yüzdelikli eşitleme, belirli bir puanın bir formdaki yüzdelik dilimini bulmayı ve bu puanı diğer formdaki aynı yüzdelik dilimdeki puana eşitlemeyi içerir (Finch vd., 2014). Eşit yüzdelikli eşitlemede, yeni formdaki bir puan ve referans formundaki bir puan grupta aynı yüzdelik sıralamasına sahiplerse, bu puanlar bir grup test katılımcısı için eşdeğerdir.

Test uygulanan bireyler, bir evrenden veya evrenlerden çekilen örneklem olarak kabul edildiği için, rastgele hata nedeniyle grafik çizildiğinde ham puan dağılımları genellikle düzensiz görünür (Cui, 2006). Rastgele hatayı azlatmanın bir yöntemi ise düzgünleştirme yöntemleridir. Düzgünleştirme yöntemleri, evren karakteristiğine düzgünlük özelliği sahip deneysel dağılımlar ve eşit yüzdelikli ilişkilerin kestirimlerini üretmek için geliştirilmişlerdir (Kolen ve Brennan, 2014).

Kolen ve Brennan (2014) eşitlemede, ön-düzgünleştirme (pre-smoothing) ve son-düzgünleştirme (post-smoothing) olmak üzere iki tip düzgünleştirme yöntemini tartışmıştır. Düzgünleştirme işlemi, eğer eşitleme işleminden önce yapılırsa ön-düzgünleştirme (presmoothing), eşitleme işleminden sonra yapıldığında ise son-düzgünleştirme (postsmoothing) şeklinde ifade edilir (Hanson vd., 1994; Kolen ve Brennan, 2014).

Ön düzgünleştirme yöntemleri ise polinominal loglinear ve beta-4 (güçlü gerçek puan yöntemi) yöntemleri olarak ikiye ayrılır (Kolen ve Brennan, 2014). Literatürde bu ön-düzgünleştirme yöntemlerinin dışında önerilen yöntemler de mevcuttur. Cui (2006), Cui ve Kolen (2009) kübik B-spline ön-düzgünleştirme yöntemi (Cubic B-spline Presmoothing Method) ve doğrudan ön-düzgünleştirme yöntemini (Direct Presmoothing Method) önermişlerdir. Bu araştırmada eşit yüzdelikli eşitlemede polinominal loglinear ön-düzgünleştirme yöntemi kullanılmıştır.

Polinomal loglinear yöntemde, gözlemlenen puanların yoğunluğunun logaritması, C + 1 dereceli bir polinomla uyumludur. Derece parametresi C, düzgünleştirilmiş dağılımın orijinal dağılıma uyum derecesini kontrol eder (Cui ve Kolen, 2009). Polinominal log lineer yöntem, aşağıdaki dağılım şeklinin modeline uyum sağlar (Kolen ve Brennan, 2014):

$$\log[N_X f(X)] = \omega_0 + \omega_1 + \omega_2 + \cdots + \omega_C x^C$$

Yukarıda verilen denklemde, yoğunluğun logu, C derecesinin daha alt düzey bir polinomu olarak ifade edilir. Örneğin, C= 2 ise, o zaman $\log[N_X f(X)] = \omega_0 + \omega_1 + \omega_2 x^2$ ve model ikinci dereceden bir polinomdur. Modeldeki ω parametresi maksimum likelihood yöntemi ile kestirilebilir (Kolen ve Brennan, 2014).

Polinominal loglinear ön-düzgünleştirme yöntemi kullanırken önemli bir husus C parametresinin seçimidir (Liu, 2011). C parametresinin seçimi, grafiksel uyum ile birlikte çeşitli ki-kare istatistiklerine dayalı olarak yapılan öznel bir seçimdir (Haberman, 1974; Hanson, 1990).

Bu noktada bahsedilmesi gereken diğer bir husus da yerel bağımsızlık varsayımıdır. Madde Tepki Kuramının bir varsayımı olan yerel bağımsızlık, bireyin maddeleri, birbirinden bağımsız olarak yanıtlamasıdır. Bir diğer deyişle, bireyin bir maddeyi yanıtlama olasılığının, istatistiksel olarak diğer maddeleri yanıtlama olasılıklarını etkilememesidir (Crocker ve Algina, 1986). Eğer bir testte, bireylerin maddelere verdikleri cevaplar yerel bağımsız değil ise diğer bir boyut bağımlılığa neden olmaktadır. Ancak yerel bağımsızlık testleri ile madde çiftleri arasındaki bağımlılığa odaklanılmaktadır. Bu bağımlılık maddelerin büyük bir kısmını etkilemediği sürece ayrı bir boyut olarak ortaya çıkmayabilir (DeMars, 2010). Araştırmamızda yer alan veri setlerininde boyutluluğun nedeni olarak yerel bağımsızlık varsayımının ihlali durumu ele alınmıştır. Bu durum veri üretimi başlığı altında detaylıca anlatılmaktadır.

Eşitleme doğruluğunu tanımlamak için kullanılan eşitlemede hata kavramı Cook ve Eignor'a (1991) göre bireyin yetenek düzeyi ile almadığı test için kestirilen yetenek düzeyi arasındaki fark olarak açıklanmaktadır. Bir eşitleme işlemi eğer hatalardan arınık ise farklı testlerden elde edilen yetenek düzeylerinin eşit olması beklenen durum olacaktır.

Literatür incelendiğinde rastgele ve sistematik olmak üzere iki tür eşitleme hatasından bahsedilmektedir (Felan,2002; Kolen ve Brennan, 2014). Eşitlemenin standart hatası olarak da bilinen random eşitleme hatası, eşitleme yapılacak test formlarının örneklem özellikleri ile ilgili bir hata türüdür. Eşitleme yapılacak test formlarının uygulandığı örneklemin büyüklüğü arttıkça eşitlemenin standart hatası

küçülür ve çok büyük örneklemler için önemsiz hâle gelir. Kavramsal olarak eşitlemenin standart hatası, eşitlenmiş puanların standart sapması olarak tanımlanabilir (Felan, 2002; Kolen, 1988; Kolen ve Brennan, 2014). Diğer hata türü ise eşitleme yanlılığı olarak da tanımlanan sistematik eşitleme hatasıdır. Bu hatanın altında yatan faktörler ise eşitleme yöntemlerini kullanmak için gerekli koşulların ve varsayımların ihmal edilmesinden kaynaklanmaktadır (Kolen ve Brennan, 2014).

Uygulamada standart hatayı tahmin etmek için iki genel yöntem geliştirilmiştir. Bu yöntemler, bootstrap ve delta yöntemleri olarak adlandırılmaktadır (Kolen ve Brennan, 2014). Bu yöntemlerden ilki olan bootstrap yöntemi, çok çeşitli istatistiklerin standart hatalarını kestirmek için kullanılan bir yöntemdir. Bootstrap yönteminde, eldeki verilerden birçok örneklem alınır ve her örneklemede eşitleme fonksiyonları kestirilir. Standart hatalar, bu birçok yeniden örneklemeden elde edilen veriler kullanılarak hesaplanır. Pratikte bootstrap yöntemini uygularken sözde rasgele sayı üreteci kullanarak rastgele örneklemler çizmek için bir bilgisayar kullanılır. İkinci yöntem ise; prosedürlerin örneklem istatistiklerini kullanarak standart hataları tahmin etmek için kullanılan ve bir denklemle sonuçlanması bakımından analitik olan delta yöntemidir. Delta yöntemi, eşitleme fonksiyonlarının standart hatalarını kestirmek için yaygın olarak kullanılan istatistiksel bir yöntemdir. Bu yöntem, hesaplama süresinin en aza indirilmesi gerektiğinde veya bir eşitleme çalışması için istenen örneklem büyüklüklerini tahmin ederken yararlı olabilir. Sonuç olarak istenilen bilgiye ve standart hatalardan yapılacak kullanıma bağlı olarak her iki yöntem türü de kullanılabilir (Kolen ve Brennan, 2014).

1.1. Araştırmanın Amacı ve Önemi

Test eşitleme çalışmalarının altında yatan gerekçenin farklı test formlarından alınan puanların karşılaştırılması ve bu puanların birbirinin yerine kullanılması gayesidir. Bu gaye doğrultusunda araştırmacılar birçok eşitleme yöntemi geliştirmiş ve geliştirilen bu yöntemler çeşitli faktörler bakımından araştırmacılar tarafından karşılaştırılmıştır. Literatür incelendiğinde ise en uygun eşitleme yönteminin hangisi olduğu konusunda araştırmacılar tarafından ortak bir kanı olmadığı, farklı koşullar altında test eşitleme yöntemlerinden elde edilen sonuçlarda değişkenlik gösterdiği görülmektedir.

Standart hataların elde edildiği yöntem bakımından çalışmalar incelendiğinde Parshall ve diğerleri (1995) doğrusal eşitleme yönteminde küçük örneklem büyüklüğünde istatistiksel yanlılığı ve standart hataları bootstrap yöntemi kullanarak karşılaştırmıştır. Ogasawara (2001) çalışmasında MTK eşitleme ve standart hataları delta yöntemi kullanılarak karşılaştırırken; Tsai vd. (2001) çalışmalarında bootstrap hatalar kullanılarak MTK eşitleme yöntemlerini karşılaştırmıştır. Cui ve Kollen (2008) eşit yüzdelikli eşitlemede random hataları parametrik ve nonparametrik yöntemlere göre karşılaştırmıştır. Zhang (2022) ve Zhang ve Zhao (2019) standart hataların hesaplanmasında kullanılan bootstrap ve delta yöntemlerine ek olarak çoklu veri atama yönteminiyle de standart hataları hesaplamış ve bu üç yöntemi karşılaştırmıştır. Salmaner Doğan ve Tan (2022) bootstrap ve delta yöntemleri ile elde edilen standart hataları Madde Tepki Kuramı gözlenen ve gerçek puan eşitleme hatalarının farklı örneklem büyüklükleri ve ölçek dönüştürme yöntemlerine göre incelenmiş ve elde edilen hataları delta ve bootstrap yöntemleri bakımından karşılaştırmıştır. Geleneksel eşitleme yöntemlerinin standart hatalarının örüntüsü ve davranışı geniş çapta araştırılmış olmasına ve sonuçların iyi bilinmesine rağmen (Tsai vd., 2001) geleneksel eşitleme yöntemlerinde hataların elde edildiği yöntemler bakımından karşılaştıran çalışmaya rastlanılmamıştır. Araştırmamız bu açıdan alana katkı sağlayacağı düşünülmektedir.

Alan yazın incelendiğinde çok boyutlu madde tepki kuramı bağlamında 2010 yılından sonra araştırmacıların bu alana yönelik çalışmalar yaptığının ve çeşitli veri yapıları altında farklı eşitleme yöntemleri geliştirdikleri görülmektedir (Brossman ve Lee, 2013; Kim, 2018; Lee ve Brossman, 2012; Lee vd., 2015; Lee ve Lee, 2016; Tao ve Cao, 2016). Bu eşitleme yöntemlerinde ise; geliştirilen eşitleme yöntemleri genellikle eşit yüzdelikli eşitleme yöntemi ile karşılaştırılmıştır. Bu araştırma ile eşitlenecek test formunun çok boyutlu olması durumunda geleneksel eşitleme yöntemlerinin performansları kendi içinde karşılaştırılmıştır. Çalışmamız bu bakımdan da alana katkı sunacağı düşünülmektedir.

Araştırmanın amacı geleneksel test eşitleme yöntemlerinde hataların belirlenmesinde kullanılan delta ve bootstrap yöntemlerinin örneklem büyüklüğü, madde sayısı ve ikinci boyuta yüklenen madde oranı değişkenleri bakımından incelenmesidir. Bu kapsamda aşağıda yer alan araştırma problemlerine cevap aranacaktır:

1. Doğrusal eşitleme yöntemi kullanılarak yapılan eşitlemede örneklem büyüklüğü, madde sayısı ve ikinci boyuta yüklenen madde oranına göre delta ve bootstrap yöntemlerinin ürettiği hatalar nasıl değişmektedir?
2. Eşit yüzdelikli eşitleme yöntemi kullanılarak yapılan eşitlemede örneklem büyüklüğü, madde sayısı ve ikinci boyuta yüklenen madde oranına göre delta ve bootstrap yöntemlerinin ürettiği hatalar nasıl değişmektedir?
3. Polinominal loglinear öndüzgünleşirilmiş eşit yüzdelikli eşitleme yöntemi kullanılarak yapılan eşitlemede örneklem büyüklüğü, madde sayısı ve ikinci boyuta yüklenen madde oranına göre delta ve bootstrap yöntemlerinin ürettiği hatalar nasıl değişmektedir?
4. Doğrusal Eşitleme, Eşit Yüzdelikli Eşitleme ve Polinominal loglinear öndüzgünleşirilmiş eşit yüzdelikli eşitleme yöntemleri ile elde edilen eşitleme hatalarının bootstrap ve delta yöntemleri için karşılaştırılmaları nasıldır?

2. Yöntem

2.1. Araştırmanın Türü

Geleneksel test eşitleme yöntemlerinde hataların belirlenmesinde kullanılan delta ve bootstrap yöntemlerinin örneklem büyüklüğü, madde sayısı ve ikinci boyuta yüklenen madde oranı değişkenleri bakımından karşılaştırmak amaçlanmış ve bu amaç doğrultusunda belirlenen koşullara özgü veriler türetilerek en az eşitleme hatasını veren yöntemin bulunması planlamıştır. Çalışma, bu özelliklerinden dolayı bir simülasyon çalışmasıdır. Araştırmada eşitleme yöntemleri farklı koşullarda simülasyon verileri ile kontrollü olarak karşılaştırılmaktadır.

2.2. Araştırma Deseni

Bu çalışmada random gruplar deseni kullanılmıştır. Random grup deseninde; ortak bir evrenden gelen cevaplayıcılar test formlarına random olarak atanır ve iki grup farklı test formlarını alır (Cook ve Eignor, 1991). Bu desen altında eşdeğer grupları oluşturmak için iki yaygın yaklaşım vardır. Birincisi, evrenden random olarak iki örneklemin seçilmesi (Finch and French, 2019) ve seçilen örneklemlerden birine Form X diğerine ise Form Y uygulamasıdır (Dorans vd., 2010; Finch and French, 2019; von Davier vd., 2004). Diğer yöntem ise spiralleme (sarmal) işlemidir. Bu terim, testin iki formunu değişen sırayla

paketlemek için kullanılan bir test jargonudur (Livingston, 2014). Test kitapçıkları paketlendiğinde Form X ve Form Y olarak paketlenir ve kitapçıklar dağıtıldığında, ilk sınava girene Form X, ikinci sınava giren Form Y, üçüncü sınava giren Form X vb. şekilde dağıtımı yapılır. Bu sarmal süreç tipik olarak, Form X ve Form Y'yi alan karşılaştırılabilir, eşdeğer gruplara yol açar (Kolen ve Brennan, 2014). Bu desende grup performansları arasındaki fark, formlar arasındaki güçlük farklılıklarının doğrudan göstergesidir. Bu nedenle örneklemden kaynaklı yanlılığı önlemek için büyük ve heterojen örneklemlerle çalışmak gerekir (Livingston, 2004).

2.3. Simülasyon Koşulları

Araştırmada, hataların belirlenmesinde kullanılan delta ve bootstrap yöntemlerinin örneklem büyüklüğü, madde sayısı ve ikinci boyuta yüklenen madde oranı değişkenleri bakımından karşılaştırmak amacıyla bir simülasyon çalışması yürütülmüştür. Bu doğrultuda simülasyon koşulları; madde sayısı (20, 40, 60), örneklem büyüklüğü (250, 1000 ve 5000) ve ikinci boyuta yüklenen madde yüzdesi (%0, %15, %30 ve %50) olarak belirlenmiştir. Simülasyon koşulları Tablo 2.1'de sunulmuştur.

Tablo 1. *Araştırmanın Simülasyon Koşulları*

Faktörler	Koşullar	Koşul Sayısı
Örneklem Büyüklüğü	250-1000-5000	3
Test Uzunluğu	20-40-60	3
İkinci Boyuta Yüklenen Madde Oranı	%0-%15-%30-%50	4

Tablo 1'de görüldüğü gibi örneklem büyüklüğü üç, test uzunluğu üç, ikinci boyuta yüklenen madde oranı dört farklı koşul içermek üzere toplamda 36 (3x3x4) koşul ele alınmıştır. Bu koşullar altında her bir test formu 100 replikasyon yapılarak türetilmiştir.

2.4. Veri Türetimi

Çalışmamızda kullanılacak verileri elde etmek için R istatistik yazılımı 4.1.1 (R Core Team, 2019) sürümü kullanılmıştır. R istatistik yazılımı altında veri türetilirken tek boyutlu verilerin türetilmesinde bireylere ait madde yanıtlarını ve gizil özellikler için tek değişkenli normal dağılımların oluşturulmasında "mirt" (Chalmers, 2012) paketinde "*for*" döngüsüyle "*simdata*" komutu, iki boyutlu verilerin türetilmesi için bireylere ait madde yanıtlarını ve gizil özellikler için iki değişkenli normal dağılımların oluşturulmasında "mirtCAT" (Chalmers, 2016) paketinde "for" döngüsüyle "generate_pattern" komutu ile 100 replikasyon kullanılarak gerçekleştirilmiştir. Elde edilen veri setlerinin ve bu veri setlerinin oluşturulmasında kullanılan dağılımların elde edildiği gerçek veri setlerine ait parametre kestirimleri ve diğer istatistikler ise MPlus 8.3 sürümü ile elde edilmiştir.

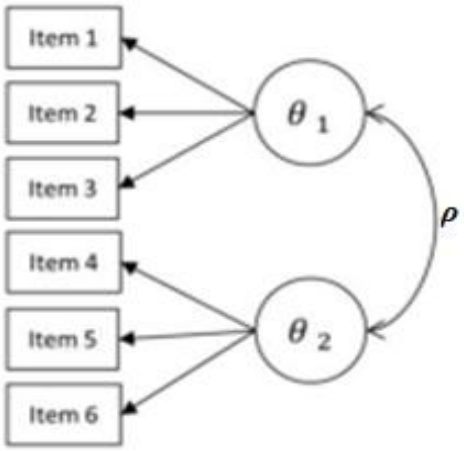
Way ve diğerleri (1988), simülasyon çalışmalarında elde edilen simülasyon verisinin gerçek yanıtları temsil etmesi gerektiği araştırmacılar tarafından dikkat edilmesi gereken en önemli nokta olarak belirtilmişlerdir. Bu doğrultuda çalışmamızda elde edilen simülasyon verileri PISA 2018 matematik sınavı birinci formundan (Form-1) elde edilen verilere ait dağılımlarından yararlanarak türetilmiştir.

Veriler iki adım takip edilerek oluşturulmuştur. İlki, simülasuyon verilerinin türetilmesinde kullanılacak olan gerçek madde parametrelerine ait dağılımlar, PISA 2018 Matematik Testi Türkiye örnekleminden faydalanarak elde edilmiştir. PISA 2018 matematik testi birinci formunda yer alan ve ikili puanlanan maddelerden oluşan veri seti önce tek boyutlu olarak analiz edilmiştir. Yapılan doğrulayıcı faktör

analizi sonucunda uyum indeksleri incelendiğinde RMSEA değerinin 0,033; CFI ve TLI değerinin 0,95'e eşit olduğu bulunmuştur. RMSEA değerlerinin 0,05'den küçük; CFI ve TLI değerinin 0.95'ten eşit veya büyük olması model veri uyumunun mükemmel olduğu (Kline, 2015; Tabachnick ve Fidell, 2013) ve testin tek boyutlu bir yapıya sahip olduğunu doğrulamaktadır. Ancak, verilere ait madde çiftleri arasında artık korelasyonlar incelendiğinde üç maddenin artıkları (residual) arasında 0,20 ve üzeri korelasyon olduğu tespit edilmiştir. Yen (1993), madde çiftleri artıkları arasında 0,2'lik bir kesme değeri önermiş ve bu değerden daha büyük değerlerin yerel bağımsızlık varsayımının ihlal ettiğini savunmuştur. Buradan hareketle yerel madde bağımsızlığını ihlal eden bu üç madde ikinci boyuta yüklenerek analiz tekrar edilmiş ve uyum indeksleri incelenmiştir. Elde edilen RMSEA değerinin 0,023 CFI ve TLI değerlerinin 0,98 olduğu ve küçük olsa da tek boyutlu yapıdan elde edilen uyum indekslerine göre iyileşme olduğu tespit edilmiştir. Bu bulgu üç maddenin (madde 16-17-18) yerel bağımsızlık varsayımını ihlal ettiği ve ikinci bir boyuta yüklendiği şeklinde yorumlanmış ve bu iki boyutlu yapıdan elde edilen dağılımlar veri türetmek için kullanılmıştır.

Mevcut çalışma için basit yapı olarak adlandırılan iki boyutlu bir yapı düşünülmüştür. İlk olarak Thurstone (1947) tarafından ortaya atılan faktör analizinde kullanılan bir terim olan "basit yapı", her bir maddenin yalnızca bir faktöre yüklendiği ve diğer faktörler üzerinde herhangi bir çapraz yükün olmadığı durumları ifade eder (McDonald, 2000; Sass ve Schmitt, 2010). Belirli bir madde için faktör yükleri, gizil yetenek ile madde arasında açık bir ilişki olduğunu düşündürecek şekilde nispeten büyük olmalı ya da gizil yetenek ile madde arasında hiçbir ilişki olmadığını gösterecek şekilde nispeten küçük olmalıdır (Finch, 2006; McLeod vd., 2001). Başka bir ifade ile, gizil yetenekler bazı maddelerde yüksek yüke sahipken, diğer maddelerde düşük (0'a yakın) yüke sahiptir (Finch, 2006). Basit yapıyı daha iyi bir şekilde açıklamak için Şekil 1 sunulmuştur.

Şekil 1. *İki Yetenekli Basit Yapılı Çok Boyutlu Model*



Şekil 1'de ρ korelasyonlu iki gizil yetenek (θ_1, θ_2) ve altı madde vardır. Birinci yetenek olan θ_1, ilk üç maddede yüksek yüke sahiptir ve ikinci yetenek olan θ_2, geri kalan üç maddede yüksek yüke sahiptir. Ek olarak, bu iki yetenek arasında bir ilişki olduğu görülmektedir. Eğer yetenekler arasında korelasyon olmasaydı, bu model tek boyutlu iki ayrı modeli yansıtmış oldurdu.

Yukarıda verilen açıklamalar doğrultusunda araştırmada basit yapılı olarak isimlendirilen iki boyutlu bir yapı ele alınmıştır, ve bu yapı iki set maddeden oluşturulmuştur. Birinci madde grubu öncelikli olarak bir boyuta (Boyut 1) yüklenir ve diğer madde grubu ise yerel madde bağımsızlığı varsayımını ihlal etmesinden dolayı başka bir boyuta (Boyut 2) yüklenir.

Literatür incelendiğinde boyutlar arası korelasyonun .70 ve üzeri olması durumunun tek boyutluluk ve çok boyutluluk arasındaki eşiği bulanıklaştırdığı (Kahraman, 2013), tek boyutlu MTK model uygulamalarının, test özellikleri yüksek oranda ilişkili olduğunda yani boyutlar arasındaki korelasyonun .80 ve üzerinde olduğunda ise tek boyutluluk varsayımın ihlaline karşı dayanıklı olduğunu (Kahraman ve Kamata, 2004; Kahraman ve Thompson, 2011) göstermektedir. Lee ve Brossman (2012), geleneksel eşitleme yöntemlerinden elde edilen eşitleme sonuçları, random gruplar tasarımı altında .8 veya üzeri bir gizli özellik korelasyonu ile makul ölçüde kabul edilebilir olacağını ileri sürmüşlerdir. Ayrıca Lee ve Brossman (2012) basit yapılı çok boyutlu eşitleme çalışmalarında gizil özellikler arasındaki korelasyon azaldıkça elde edilen hataların büyüklüğünün arttığını tespit etmişlerdir. Uğurlu (2020) ise basit yapılı çok boyutlu testlerle yürütmüş olduğu eşitleme çalışmasında gizil yetenekler arasındaki korelasyonun .5 olması koşulunun eşitleme ilişkisini en doğru yansıtan koşul olduğu bulgusuna ulaşmıştır. Tüm bu bulgular ışığında araştırmamızda basit yapılı iki boyutlu olarak türetilen veri setlerinde iki boyut arası korelasyon .5 olarak sabitlenmiştir.

Tek boyutlu yapıya sahip veri setinde madde ayırt edicilik parametreleri ortalaması 0 ve standart sapması 0,2 olan log normal dağılımdan elde edilirken madde güçlük parametresi ise ortalaması 0 ve standart sapması 1 olan normal dağılımdan elde edilmiştir. Bireylere ait θ parametreleri ise ortalaması 0 ve standart sapması 1 olan normal dağılımdan türetilmiştir. Yerel bağımsızlık varsayımının ihlal edildiği veri setleri ise PISA 2018 Matematik testi birinci formu Türkiye örnekleminden elde edilen dağılıma göre türetilmiştir. Birinci boyuta yüklenen maddelere ait madde ayırt edicilik parametreleri ortalaması 0,53 ve standart sapması 0,178, ikinci boyuta yüklenen maddelere ait ayırt edicilik parametreleri ise ortalaması 0,70 ve standart sapması 0,21 olan log-normal dağılımdan elde edilirken; madde güçlük parametresi ise ortalaması 0,327 ve standart sapması 0,5 olan normal dağılımdan elde edilmiştir. Madde parametreleri oluşturulduktan sonra, iki değişkenli normal dağılımlar BN(0, 0, 1, 1, 0.5) kullanılarak yetenek parametreleri oluşturulmuştur. Elde edilen tüm veriler tek boyutlu 2 parametreli lojistik model altında türetilmiştir.

Veri üretiminde dikkat edilen diğer hususta şu şekildedir: Yerel bağımsızlık varsayımını ihlal eden madde yüzdesi %15 olması koşulunda yukarıda belirtilen dağılıma doğrultusunda 2PLM'e göre veri üretilmiştir. Bu koşulun teste yer alan maddelerin %30'unun yerel bağımsızlık varsayımını ihlal edip ikinci boyuta yüklenmesi durumunda ise ilk koşula göre kestirilen madde parametreleri sabitlenmiş sadece ikinci boyuta yüklenecek yeni madde parametreleri yukarıda bahsedilen dağılımdan elde edilerek yeni veri seti oluşturulmuştur.

2.5. Verilerin Analizi

Bu çalışmada doğrusal, eşit yüzdelikli ve polinominal loglinear öndüzgünleştirilmiş eşit yüzdelikli eşitleme yöntemleri kullanılarak test eşitleme yapılmıştır. Araştırmamızda kullanılan verilerin üretimi R istatistik yazılımı 4.1.1 sürümünde Chalmers (2012) tarafından yazılan mirt paketi ve Chalmers (2022) tarafından yazılan mirtCAT paketi kullanılarak yapılmıştır. Test eşitleme işlemi ise R yazılımda "equate"

paketi (Albano, 2016) kullanılarak yapılmıştır. Araştırmamız "equate" paketi kullanılarak delta yöntemi ve bootstrap yöntemine göre elde edilen eşitleme hataları üzerinden yürütülmüştür.

Araştırmamızda delta ve bootstrap eşitleme hataları kestirim yöntemlerinden elde edilen eşitleme hatalarının karşılaştırılması 36 farklı koşul altında yapıldığı yukarıda detaylı olarak belirtilmiştir. Her bir form türetilirken 100 replikasyon kullanılmış ve böylece sadece bir koşul için 100 Form Y ve 100 Form X test verisi elde edilmiştir. Elde edilen bu test verileri doğrusal, eşit yüzdelikli ve polinominal loglinear öndüzgünleştirilmiş eşit yüzdelikli eşitleme yöntemleri kullanılarak eşitlenmiştir. Bu da her bir eşitleme yöntemi altında sadece bir koşul için 100 eşitleme işlemi yapıldığı anlamına gelir. Daha sonra yapılan bu eşitleme işlemi sonucunda elde edilen bootstrap ve delta eşitleme hatalarının ortalaması alınmış ve elde edilen değer testte yer alan madde sayılarına bölünerek 0 ile 1 arasında ölçeklendirilmiştir. Bu ölçeklendirme ise madde sayısısı koşuluna göre hataları karşılaştırabilmek amacıyla gerçekleştirilmiştir. Burada belirtilmesi gereken bir nokta da her bir eşitleme işleminde bootstrap eşitleme hatalarını elde etmek içinde ayrıca 100 replikasyon kullanıldığıdır.

3. Bulgular

3.1. Birinci Araştırma Problemine Ait Bulgular

Doğrusal eşitleme sonucunda araştırmaya dahil edilen tüm simülasyon koşulları için elde edilen delta ve bootstrap hata ortalamaları Tablo 2'de verilmiştir.

Tablo 2. *Doğrusal Eşitleme Yönteminde Delta ve Bootstap Yöntemleriyle Elde Edilen Eşitleme Hataları*

		Madde Sayısı					
Örneklem	**İkinci Boyuta**	**20**		**40**		**60**	
Büyüklüğü	**Yüklenen Madde Oranı**	**Delta**	**Bootstrap**	**Delta**	**Bootstrap**	**Delta**	**Bootstrap**
250	0%	0,0175	0,0264	0,0308	0,0254	0,0510	0,0274
	15%	0,0183	0,0290	0,0346	0,0262	0,0577	0,0300
	30%	0,0172	0,0312	0,0341	0,0345	0,0542	0,0286
	50%	0,0178	0,0290	0,0332	0,0260	0,0564	0,0290
1000	0%	0,0043	0,0146	0,0089	0,0144	0,0127	0,0136
	15%	0,0044	0,0136	0,0092	0,0145	0,0132	0,0143
	30%	0,0042	0,0144	0,0094	0,0145	0,0136	0,0135
	50%	0,0043	0,0149	0,0093	0,0140	0,0138	0,0160
5000	0%	0,0009	0,0070	0,0017	0,0064	0,0024	0,0062
	15%	0,0009	0,0064	0,0017	0,0058	0,0025	0,0060
	30%	0,0009	0,0069	0,0018	0,0065	0,0026	0,0058
	50%	0,0009	0,0069	0,0017	0,0064	0,0026	0,0064

Tablo 2, örneklem büyüklüğü dikkate alınarak incelendiğinde 1000 ve 5000 örneklem büyüklüğü altında yapılan doğrusal eşitleme sonucunda tüm simülasyon koşullarında delta yöntemi ile elde edilen hataların bootstrap yöntemi ile elde edilen hatalardan daha düşük olduğu, 250 örneklem büyüklüğünde ise 20 madde koşulunda delta yöntemi ile hesaplanan hataların bootstrap yöntemi ile hesaplanan hatalardan düşük olduğu ancak 40 ve 60 madde koşulunda ise bootstrap yöntemi ile hesaplanan hataların

delta yönteminden daha düşük olduğu görülmektedir. Örneklem büyüklüğü ile ilgili başka bir bulgu ise beklenilen üzere örneklem büyüklüğü arttıkça hataların azaldığı şeklindedir. Testte yer alan madde miktarı arttıkça delta yöntemi ile elde edilen hatalarda bir artış olduğu dikkat çeken başka bir bulgu olarak karşımıza çıkmaktadır. Bootstrap yönteminde ise testte yer alan madde sayısı arttıkça test formlarının tek boyutlu olması durumunda 1000 ve 5000 örneklem büyüklüklerinde elde edilen hatalarda bir azalma olduğu, 250 örneklem büyüklüğünde ise sistematik bir azalmadan bahsetmenin mümkün olmadığı görülmektedir. Sadece testte yer alan madde miktarı dikkate alındığında delta yöntemi için en az hatanın 20 madde olması durumunda, bootstrap yönteminde ise 1000 ve 5000 örneklem büyüklüğü altında 60 madde olması durumunda hesaplandığı görülmektedir. Bu bulgu her iki yöntemin testte yer alan madde sayısı açısından elde edilen hataları hesaplamada farklılaştığını göstermektedir.

Testte yer alan maddelerden bazılarının ikinci boyuta yüklenmesi sonucunda testin yapısının çok boyutlu olması durumunun yapılan eşitleme sonucuna etkisinin en az olduğu yöntemin delta yöntemi olduğu görülmektedir. Ayrıca delta yöntemi, 250 örneklem büyüklüğünde testte yer alan madde sayısı farketmeksizin ikinci boyuta yüklenen madde oranının %15 olması koşulunda diğer ikinci boyuta yüklenen madde oranlarına göre daha fazla hata ürettiği tespit edilmiştir. Bu sistematik bulgu sadece 250 örneklem büyüklüğünde delta yöntemi için geçerlidir. Testin çok boyutlu bir yapıya sahip olması durumunun test eşitleme sonuçlarına etkisinin en fazla olduğu koşulun ise delta yöntemi için 250 örneklem büyüklüğü altında, testte yer alan madde sayısının 60 olması ve bu maddelerin %15'inin ikinci boyuta yüklenmesi durumda, bootstrap yöntemi için de 250 örneklem büyüklüğü altında, testte yer alan madde sayısının 40 olması ve bu maddelerin %30'unun ikinci boyuta yüklenmesi durumda olduğu tespit edilmiştir. Her iki yöntemin testin yapısının çok boyutlu olması durumu açısından karşılaştırılmasına bakıldığında delta yönteminin bootstrap yöntemine göre hata miktarlarında daha az dalgalanma olduğu görülmektedir. Genel olarak testte yer alan maddelerden bazılarının ikinci boyuta yüklenmesi durumunda ikinci boyuta yüklenen madde oranı açısından sistematik olarak bir artış veya azalış olmadığı, farklı koşullarda tek boyutlu yapıya nazaran farklı hatalar hesaplandığı tespit edilmiştir.

3.2. İkinci Araştırma Problemine Ait Bulgular

Eşit yüzdelikli eşitleme sonucunda araştırmaya dahil edilen tüm simülasyon koşulları için elde edilen delta ve bootstrap hata ortalamaları Tablo 3'te verilmiştir.

Tablo 3. *Eşit Yüzdelikli Eşitleme Yönteminde Delta ve Bootstap Yöntemleriyle Elde Edilen Eşitleme Hataları*

		Madde Sayısı					
Örneklem	**İkinci Boyuta**	**20**		**40**		**60**	
Büyüklüğü	**Yüklenen Madde Oranı**	**Delta**	**Bootstrap**	**Delta**	**Bootstrap**	**Delta**	**Bootstrap**
250	0%	0,0360	0,0363	0,0301	0,0326	0,0319	0,0357
	15%	0,0330	0,0328	0,0318	0,0326	0,0332	0,0340
	30%	0,0318	0,0351	0,0343	0,0329	0,0364	0,0321
	50%	0,0322	0,0342	0,0324	0,0328	0,0315	0,0362
1000	0%	0,0175	0,0174	0,0180	0,0171	0,0172	0,0179
	15%	0,0170	0,0177	0,0167	0,0169	0,0166	0,0157
	30%	0,0169	0,0160	0,0164	0,0154	0,0163	0,0168
	50%	0,0168	0,0161	0,0171	0,0167	0,0164	0,0164
5000	0%	0,0080	0,0078	0,0078	0,0075	0,0078	0,0083
	15%	0,0078	0,0071	0,0075	0,0069	0,0075	0,0076
	30%	0,0076	0,0076	0,0075	0,0072	0,0073	0,0075
	50%	0,0076	0,0074	0,0075	0,0069	0,0072	0,0068

Tablo 3'e bakıldığında örneklem büyüklüğü arttıkça her iki yöntemle hesaplanan hata miktarlarının tüm koşullar altında azaldığı görülmektedir. Delta ve bootstrap yöntemlerini kullanarak hesaplanan hatalar karşılaştırıldığında bu yöntemlerle elde edilen hata miktarlarının birbirine benzer olduğu koşulun 5000 örneklem olduğu, birbirinden en farklılaştığı koşulun ise 250 örneklem büyüklüğü altında olduğu görülmektedir. Delta yönteminde testin tek boyutlu olması durumunda 20 madde koşulunda elde edilen hatalar tüm örneklem büyüklükleri için 60 madde koşulunda elde edilen hatalardan fazla olduğu görülmektedir. Bootstrap yönteminde ise bu durum sadece 250 örneklem büyüklüğü altında görülürken, 1000 ve 5000 örneklem büyüklüklerinde 60 madde koşulunda elde edilen hataların 20 madde koşulu altında elde edilen hatalardan az da olsa yüksek olduğu bulunmuştur. Her iki yöntem için de 1000 ve 5000 örneklem büyüklüğünde testin tek boyutlu yapısının bozulması durumunda ikinci boyuta yüklenen madde oranı farketmeksizin tek boyutlu yapıdan daha az hata ürettiği görülmektedir.

Bazı maddelerin ikinci boyuta yüklenmesi sonucu yapılan eşitleme işleminde testin tek boyutlu olması durumuna göre en fazla farklılaşan koşulun her iki yöntem içinde 250 örneklem büyüklüğünde 20 madde koşulu altında gerçekleştiği görülmektedir. Bu koşul altında bootstrap yönteminin delta yöntemine nazaran daha az farklılaştığı bulunmuştur. Burada yapılan yorumun üretilen hata miktarlarından ziyade testtin tek boyutlu olması durumu ile çok boyutlu olması sonucunda üretilen hata miktarları farkı olduğu dikkat edilmesi gereken husustur. Ayrıca 250 örneklem büyüklüğünde testte yer alan madde sayısı koşullarının hepsi için bootstap yönteminin delta yöntemine göre daha iyi performans gösterdiği, 5000 örneklem büyüklüğünde ise delta yönteminin bootstap yöntemine göre daha iyi performans gösterdiği dikkat çeken başka bir bulgudur. 1000 örneklem büyüklüğünde ise 20 ve 60 madde koşulunda delta yönteminden 40 madde koşulunda ise bootstrap yönteminden elde edilen hataların testin tek boyutlu olması durumunda elde edilen eşitleme hatalarına daha yakın hata üreten yöntem olduğu söylebilir. Örneklem büyüklüğü arttıkça testin basit yapılı iki boyutlu bir yapıya sahip olmasının eşitleme sonuçlarını daha az etkilediği, tek boyutlu olması durumunda yapılan eşitleme sonucu elde edilen hatalara yakın hatalar ürettiği tespit edilmiştir.

3.3. Üçüncü Araştırma Problemine Ait Bulgular

Polinominal loglinear öndüzgünleştirilmiş eşit yüzdelikli eşitleme yöntemi kullanılarak yapılan eşitleme sonucunda araştırmaya dahil edilen tüm simülasyon koşulları için elde edilen delta ve bootstrap hata ortalamaları Tablo 4'te verilmiştir.

Tablo 4. *Polinominal Loglinear Öndüzgünleştirilmiş Eşit Yüzdelikli Eşitleme Yönteminde Delta ve Bootstap Yöntemleriyle Elde Edilen Eşitleme Hataları*

		Madde Sayısı					
Örneklem	**İkinci Boyuta**	**20**		**40**		**60**	
Büyüklüğü	**Yüklenen Madde Oranı**	**Delta**	**Bootstrap**	**Delta**	**Bootstrap**	**Delta**	**Bootstrap**
250	0%	0,0347	0,0315	0,0342	0,0302	0,0358	0,0368
	15%	0,0356	0,0328	0,0335	0,0274	0,0345	0,0299
	30%	0,0335	0,0300	0,0337	0,0307	0,0342	0,0305
	50%	0,0336	0,0268	0,0335	0,0313	0,0341	0,0293
1000	0%	0,0180	0,0155	0,0178	0,0163	0,0176	0,0166
	15%	0,0171	0,0150	0,0170	0,0151	0,0169	0,0157
	30%	0,0169	0,0153	0,0169	0,0150	0,0165	0,0140
	50%	0,0170	0,0165	0,0169	0,0155	0,0166	0,0149
5000	0%	0,0080	0,0075	0,0078	0,0063	0,0078	0,0072
	15%	0,0077	0,0075	0,0075	0,0069	0,0075	0,0073
	30%	0,0075	0,0065	0,0074	0,0068	0,0073	0,0068
	50%	0,0075	0,0069	0,0074	0,0071	0,0073	0,0063

Tablo 4'e göre 250, 1000 ve 5000 örneklem büyüklüklerinde genel olarak bootstrap yöntemi ile elde edilen hataların delta yöntemi ile elde edilen hatalardan daha az olduğu görülmektedir. Sadece 250 örneklem büyüklüğü 60 madde koşulunda testtin tek boyutlu olması durumunda bootstrap yöntemi ile elde edilen hatanın delta yöntemi ile elde edilen hatadan daha büyük hesaplanmıştır. Yine örneklem büyüklükleri dikkate alındığında örneklem büyüklüğü arttıkça her iki yöntemle hesaplanan hata miktarlarının da azalma olduğu açıkça görülmektedir. Testte yer alan madde sayısı dikkate alındığında testte yer alan madde sayısı arttıkça 1000 ve 5000 örneklem altında delta yönteminde hata miktarlarında sistematik olarak bir azalma olduğu görülmektedir. Testte yer alan maddelerden bazılarının ikinci boyuta yüklenmesi sonucunda testin yapısının çok boyutlu olması durumunun yapılan eşitleme sonucunda delta yöntemi ile elde edilen hataların referans değerimiz olan tek boyutlu eşitleme hatalarından daha az olduğu tespit edilmiştir. Testin tek boyutlu olması durumunda yapılan eşitleme sonucu elde edilen hatalar ile testin çok boyutlu olması durumunda elde edilen eşitleme hataların birbirine en yakın olduğu yöntem, 5000 örneklem büyüklüğünde delta yöntemi olduğu görülmektedir. Testin çok boyutlu bir yapıya sahip olması durumunun test eşitleme sonuçlarına etkisinin en fazla olduğu koşulun ise her iki yöntem için de 250 örneklem büyüklüğü ve testte yer alan madde sayısının 60 olduğu durumda tespit edilmiştir. Her iki yöntemin belirtilen koşul altında karşılaştırılmasına bakıldığında bootstrap yönteminin delta yöntemine göre hata miktarlarında daha fazla değişim olduğu görülmektedir. Buradaki değişimden kastın referans değerimiz olan tek boyutlu eşitleme hatalarından farklılaşması olduğu dikkat edilmesi gereken önemli noktadır.

3.4. Dördüncü Araştırma Problemine Ait Bulgular

Araştırmaya dahil edilen geleneksel eşitleme yöntemleri ile yapılan eşitleme sonuçlarının tüm simülasyon koşulları için elde edilen delta ve bootstrap hata ortalamaları Tablo 5'te verilmiştir.

Tablo 5. *Geleneksel Eşitleme Yöntemlerinde Delta ve Bootstap Yöntemleriyle Elde Edilen Eşitleme Hataları*

			İkinci Boyuta Yüklenen Madde Oranı (%)											
N	**Eşitleme Yöntemi**	**Hata**	**0%**	**15%**	**30%**	**50%**	**0%**	**15%**	**30%**	**50%**	**0%**	**15%**	**30%**	**50%**
250	**Doğrusal**	**Delta**	0,017	0,018	0,017	0,018	0,031	0,035	0,034	0,033	0,051	0,058	0,054	0,056
		Boots	0,026	0,029	0,031	0,029	0,025	0,026	0,034	0,026	0,027	0,030	0,029	0,029
	EYE	**Delta**	0,036	0,033	0,032	0,032	0,030	0,032	0,034	0,032	0,032	0,033	0,036	0,031
		Boots	0,036	0,033	0,035	0,034	0,033	0,033	0,033	0,033	0,036	0,034	0,032	0,036
	Ön Düz.	**Delta**	0,035	0,036	0,033	0,034	0,034	0,033	0,034	0,034	0,036	0,035	0,034	0,034
		Boots	0,031	0,033	0,030	0,027	0,030	0,027	0,031	0,031	0,037	0,030	0,030	0,029
1000	**Doğrusal**	**Delta**	0,004	0,004	0,004	0,004	0,009	0,009	0,009	0,009	0,013	0,013	0,014	0,014
		Boots	0,015	0,014	0,014	0,015	0,014	0,014	0,015	0,014	0,014	0,014	0,013	0,016
	EYE	**Delta**	0,018	0,017	0,017	0,017	0,018	0,017	0,016	0,017	0,017	0,017	0,016	0,016
		Boots	0,017	0,018	0,016	0,016	0,017	0,017	0,015	0,017	0,018	0,016	0,017	0,016
	Ön Düz.	**Delta**	0,018	0,017	0,017	0,017	0,018	0,017	0,017	0,017	0,018	0,017	0,016	0,017
		Boots	0,016	0,015	0,015	0,017	0,016	0,015	0,015	0,015	0,017	0,016	0,014	0,015
5000	**Doğrusal**	**Delta**	0,001	0,001	0,001	0,001	0,002	0,002	0,002	0,002	0,002	0,002	0,003	0,003
		Boots	0,007	0,006	0,007	0,007	0,006	0,006	0,007	0,006	0,006	0,006	0,006	0,006
	EYE	**Delta**	0,008	0,008	0,008	0,008	0,008	0,007	0,007	0,007	0,008	0,008	0,007	0,007
		Boots	0,008	0,007	0,008	0,007	0,008	0,007	0,007	0,007	0,008	0,008	0,007	0,007
	Ön Düz.	**Delta**	0,008	0,008	0,008	0,008	0,008	0,008	0,007	0,007	0,008	0,007	0,007	0,007
		Boots	0,008	0,007	0,007	0,007	0,006	0,007	0,007	0,007	0,007	0,007	0,007	0,006

EYE : Eşit yüzdelikli eşitleme
Ön Düz. : Polinominal loglinear öndüzgünleştirilmiş eşit yüzdelikli eşitleme yöntemi

Tablo 5'te göre eşitleme yöntemleri karşılaştırıldığında; 250, 1000 ve 5000 örneklem büyüklüğünde diğer eşitleme yöntemlerine göre daha az hata üreten eşitleme yöntemin doğrusal eşitleme yöntemi olduğu ve bu eşitleme yöntemi 1000 ve 5000 örneklem büyüklüğü altında delta yönteminin bootstrap yöntemine nazaran daha az hata ürettiği, 250 örneklem büyüklüğünde ise testte yer alan madde sayısının 40 ve 60 olması durumunda bootstrap yönteminin delta yöntemine nazaran daha az hata ürettiği tespit edilmiştir. Doğrusal eşitleme yöntemi, küçük örneklem olarak niteleyebileceğimiz 250 örneklem büyüklüğünde testte yer alan madde sayısının artması durumundan en fazla etkilenen yöntem olarak

karşımıza çıkmaktadır. Tüm eşitleme yöntemleri için testte yer alan kişi sayısı arttıkça eşitleme sonuçlarından elde edilen hata miktarının azaldığı, testte yer alan madde sayısı arttıkça doğrusal eşitleme yöntemi için delta yöntemi ile elde edilen hatalarda artış olduğu açıkça görülmektedir. 1000 ve 5000 örneklem büyüklüğü altında doğrusal eşitleme yönteminde bootstrap yöntemi ile elde edilen hata miktarları ise testte yer alan madde sayısı arttıkça azaldığı tespit edilmiştir. Ayrıca eşit yüzdelikli eşitleme ve polinominal loglinear ön düzgünleştirme yöntemleri için de madde sayısı arttıkça elde edilen hatalarda azalma eğiliminde oldukları görülmektedir. Testte yer alan madde sayısının artması durumundan en fazla etkilenen örneklem büyüklüğünün ise 250 örneklem büyüklüğü olduğu bir diğer dikkat çekici bulgu olarak karşımıza çıkmaktadır. Eşitleme hatalarını belirlemede kullanılan yöntem açısından bakılacak olursa, doğrusal eşitleme yönteminde delta yönteminin bootstrap yöntemine göre daha iyi performans gösterdiği, eşit yüzdelikli eşitlemede ise ele alınan bazı koşullara göre ufak farklar olsa da 250 örneklem büyüklüğünde delta yönteminin bootstrap yöntemine göre daha az hata ürettiği, 1000 ve 5000 örneklem büyüklüğünde ise bootstrap yönteminin daha az hata ürettiği tespit edilmiştir. Polinominal loglinear öndüzgünleştirme yöntemi ile yapılan eşitleme işleminde ise bootstrap yönteminin delta yöntemine göre daha az hata ürettiği görülmektedir.

Testte yer alan maddelerden bazılarının ikinci boyuta yüklenmesi sonucunda testin yapısının çok boyutlu olması durumunun yapılan eşitleme sonucuna etkisini azaltan koşulun ise testte yer alan kişi sayısı olduğu görülmektedir. Testte yer alan kişi sayısı arttıkça testin tek boyutlu olması durumunda elde edilen hatalar ile çok boyutlu olması durumunda yapılan eşitleme sonucu elde edilen hataların biribirine yakın olduğu görülmektedir. Ayrıca doğrusal eşitleme yönteminde delta hata puanlarının 1000 ve 5000 örneklem büyüklükleri altında testin yapısının çok boyutlu olması durumundan en az etkilenen yöntem olduğu, tek boyutlu eşitleme sonucu elde edilen eşitleme hataları ile çok boyutlu olması durumunda elde edilen eşitleme hatalarının hemen hemen aynı olduğu görülmektedir. Testin tek boyutlu yapısının bozulup çok boyutlu olması durumundan en fazla etkilenen örneklem büyüklüğünün ise 250 örneklem büyüklüğü olduğu, bu örneklem büyüklüğü altında en fazla etkilenen yöntemin ise doğrusal eşitleme sonucunda delta yöntemi ile elde edilen hatalar olduğu tespit edilmiştir. Burada yapılan yorum elde edilen hata miktarlarından ziyade tek boyutlu olarak hesaplanan hata miktarları ile çok boyutlu olması durumunda hesaplanan hata miktarlarının değişimi şeklinde olduğu gözden kaçırılmamalıdır. Ayrıca ikinci boyuta yüklenen maddelerin oranının eşitleme yöntemlerine göre sistematik bir şekilde farklılaşma olmadığı, farklı koşullarda farklı sonuçlar elde edildiği anlaşılmıştır.

4. Sonuç, Tartışma Ve Öneriler

Araştırmada, geleneksel test eşitleme yöntemlerinde hataların belirlenmesinde kullanılan delta ve bootstrap yöntemlerinin örneklem büyüklüğü, madde sayısı ve ikinci boyuta yüklenen madde oranı değişkenleri bakımından incelenmiştir. Ayrıca geleneksel eşitleme yöntemleri elde edilen hatalar bakımından karşılaştırılmıştır. Araştırmada Form Y tek boyutlu olarak ele alınırken FormX hem tek boyutlu hem de yerel bağımsızlık varsayımının ihmal edilmesi sonucu bazı maddelerin ikinci boyuta yüklenmesi durumunda iki boyutlu ve basit yapılı olarak ele alınmış ve eşitleme işlemi yapılmıştır.

Örneklem büyüklüğü bakımından geleneksel eşitleme yöntemleri incelendiğinde doğrusal eşitleme yönteminin diğer eşitleme yöntemlerine göre daha iyi performans gösterdiği görülmektedir. Burada dikkat çeken bir sonuç hataların elde edilmesinde kullanılan yöntem bakımından farklılıklar olması

durumudur. Doğrusal eşitleme yönteminde hataların delta yöntemi kullanarak hesaplanması durumunda diğer eşitleme yöntemlerine kıyasla 1000 ve 5000 örneklem büyüklüğünde en az hata üreten yöntem olurken 250 örneklem büyüklüğünde ise bu durumun tam aksine diğer yöntemlere oranla daha fazla hata ürettiği görülmektedir. 250 örneklem büyüklüğünde ise doğrusal eşitleme yöntemi eşitleme hatalarının bootstrap yöntemi kullanılarak hesaplanması durumunda diğer eşitleme yöntemlerine göre en az hata üreten eşitleme yöntemi olduğu tespit edilmiştir. Bu bulgular özetlendiğinde doğrusal eşitleme yönteminin diğer karşılaştırılan yöntemlere göre daha iyi performans sergilediği şeklinde yorumlanabilir. Bu bulgu Kelecioğlu ve Öztürk Gübeş (2013) ve Özkan (2015) araştırma bulguları ile örtüşürken, İnci (2014) araştırma bulguları ile çelişmektedir. İnci (2014) incelemiş olduğu tüm örneklem büyüklükleri için eşit yüzdelikli eşitlemenin doğrusal eşitlemeye göre daha az hata ürettiği sonucuna ulaşmıştır. Zhu (1998), tek grup desenine dayalı olarak yürüttüğü çalışmasında doğrusal eşitleme yönteminin en az random hata içeren yöntem olduğu bulgusuna ulaşmıştır. 1000 ve 5000 örneklem büyüklüğünde en az hata üreten yöntem olan doğrusal eşitleme yönteminde hataların elde edilmesinde kullanılan yöntem bakımından karşılaştırıldığında delta yönteminin bootstrap yöntemine nazaran daha az hata ürettiği, 250 örneklem büyüklüğü altında bootstrap yönteminin delta yöntemine göre daha az hata ürettiği görülmektedir. Mutluer (2021) yapmış olduğu araştırmada doğrusal eşitleme yöntemi altında delta yönteminin bootsrap yöntemine göre daha az hata ürettiğini ortaya koymuştur. Bu bulgu araştırmamız sonuçları ile kısmen örtüşmektedir. Tüm eşitleme yöntemleri için testte yer alan kişi sayısı arttıkça eşitleme sonuçlarından elde edilen hata miktarının azaldığı tespit edilmiştir. Kolen ve Brennan (2014), örneklem büyüklüğünün eşitleme hatası üzerinde etkisi olduğunu belirtmiştir. Bazı araştırmalar ise herhangi bir eşitleme işlemi için küçük örneklem büyüklüklerinin kullanılmasının, tahminin doğruluğunu tehlikeye atabileceğini ve eşitleme hatasını artırabileceğini savunmuştur (Kim vd., 2008; Livingston, 1993; Livingston ve Kim, 2009; Skaggs, 2005). Harris ve Crouse (1993), Kim ve Cohen (2002), Kilmen (2010), Lee ve Ban (2010), İnci (2014), Salmaner Doğan ve Tan (2022) yaptıkları çalışmalarda örneklem büyüklüğü arttıkça eşitleme yöntemlerinden elde edilen eşitleme hatalarının azaldığını ortaya koymuşlardır. Alan yazında yer alan bu bulgular araştırma sonuçlarımızı desteklemektedir. Sonuç olarak 1000 ve üzeri örneklemlerle yapılan eşitlemede, eşitleme hatasının az olması nedeniyle eşitleme sonuçlarının daha doğru kestirilmesine katkı sağlayacaktır. Ayrıca örneklem büyüklüğünün 1000 ve üzeri olması durumunda eşitleme yöntemlerinden elde edilen hataların birbirine oldukça yaklaştığı dikkat edilmesi gereken bir bulgu olarak karşımıza çıkmaktadır.

Diğer değişkenimiz olan testte yer alan madde sayısı açısından değerlendirecek olursak karşılaştırılan yöntemler için farklılıklar olduğu görülmektedir. Doğrusal eşitleme yönteminde eşitleme hatalarını belirlemek için kullanılan yöntemin delta olması durumunda madde sayısı arttıkça hatalarda artış olduğu, bunun aksine bootstrap yönteminde ise testte yer alan madde sayısı arttıkça eşitleme hatalarında azalma olduğu görülmektedir. Ayrıca eşit yüzdelikli eşitleme ve polinominal loglinear ön düzgünleştirme yöntemleri için de madde sayısı arttıkça elde edilen hatalarda azalma eğiliminde oldukları tespit edilmiştir. Testte yer alan madde sayısının artması durumundan en fazla etkilenen yöntemin ise doğrusal eşitleme yöntemi olduğu bulunmuştur. Bu noktadan hareketle eşitleme yapacak olan araştırmacıların eşitlenecek test formlarının özelliklerine göre kullanacakları yöntemleri belirlemeleri gerektiği açıkça görülmektedir.

Yerel bağımsızlık varsayımının ihlali sonucu bu varsayımı ihlal eden maddelerin başka bir boyuta yüklenmesiyle testin yapısının çok boyutlu olması durumunun eşitleme sonuçlara etkisini azaltan koşulun

örneklem büyüklüğü olduğu tespit edilmiştir. Örneklem büyüklüğü arttıkça testin tek boyutlu olması durumunda elde edilen hatalar ile çok boyutlu olması durumunda yapılan eşitleme sonucu elde edilen hataların biribirine yakın olduğu görülmektedir. 1000 ve 5000 örneklem büyüklüğünde hem delta hem de bootstrap yöntemleri ile hesaplanan hataların doğrusal eşitleme, eşit yüzdelikli eşitleme ve polinominal loglinear öndüzgünleştirme yöntemlerinde ikinci boyuta yüklenen madde oranları fark etmeksizin testin tek boyutlu olması durumundan daha az veya benzer hatalar üretme eğiliminde oldukları görülmektedir. Ek olarak doğrusal eşitleme yönteminde delta hata puanlarının 1000 ve 5000 örneklem büyüklükleri altında testin yapısının çok boyutlu olması durumundan en az etkilenen yöntem olduğu tespit edilmiştir. Testin çok boyutlu olması durumunun test eşitleme sonuçlarına etkisinin en fazla olduğu koşulun 250 örneklem büyüklüğü olduğu görülmektedir. 250 örneklem büyüklüğünde ise bootstrap yönteminin delta yöntemine göre tek boyutlu hesaplanan hatalara daha yakın hata ürettiği tespit edilmiştir. Testin tek boyutlu yapısının bozulup çok boyutlu olması durumunda ikinci boyuta yüklenen madde oranı bakımından değerlendirecek olursak elde edilen eşitleme hatalarında sistematik bir bulguya rastlanılmadığı, araştırmada ele alınan test eşitleme yöntemleri, eşitleme hatalarını belirlemede kullanılan yöntem, örneklem büyüklüğü ve testte yer alan madde sayısı koşullarına göre değişkenlik gösterdiği görülmektedir. Buna karşılık en istikrarsız hata hesaplamaları 250 örneklem ve 60 madde koşulları altında tespit edilmiştir. Birinci boyutla en fazla farklılaşan koşul ise polinominal loglinear öndüzgünleştirme yöntemi altında 250 örneklem büyüklüğünde testte yer alan madde sayısının 60 olması durumunda ikinci boyuta yüklenen madde oranının %50 olması koşulunda tespit edilmiştir. Elde edilen hata miktarı ise tek boyutlu olması durumundan daha az kestirilmiştir.

Özetle, örneklem sayısı küçüldükçe elde edilen hata miktarlarında artış olduğu, en az hata içeren koşulun 5000 örneklem büyüklüğü ve testte yer alan madde sayısının 20 olması durumunda doğrusal eşitlemede delta yönteminde elde edildiği, en iyi performansa sahip eşitleme yönteminin doğrusal eşitleme yöntemi olduğu ve eşitleme hatalarını belirlemede kullanılan yöntem olarak da delta yöntemi olduğu sonucuna ulaşılmıştır.

Bu çalışma, simülasyon veriler üzerinden yürütülmüştür. Gerçek veri setleri üzerinden test eşitleme çalışması yapılarak araştırma tekrarlanabilir. Araştırmamızda kullanılan random grup deseni yerine farklı eşitleme desenleri kullanarak benzer araştırmalar yapılabilir. Ayrıca formlar arasında güçlük farkı, yerel bağımsızlık varsayımın derecesi, iki boyut arasında korelasyonların farklılaşması durumu gibi kontrol edilebilecek değişkenler de eklenerek bu çalışma daha da geliştirilebilir. Geleneksel eşitleme yöntemleri ile MTK test eşitleme yöntemleri ve Çok Boyutlu Madde Tepki Kuramı'na dayalı test eşitleme yöntemleri karşılaştırılabilir. Son olarak incelenen hata kestirim yöntemlerine ek olarak çoklu veri atama (multiple imputation) yöntemi de eklenerek alanyazına katkı sağlayan araştırmalar yapılabilir.

Kaynaklar

Aiken, L. R. (2000). *Psychological testing and assesment.* Allyn and Bacon.

Albano, A. D. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software, 74*(8), 1-36. https://doi.org/10.18637/jss.v074.i08

Baykul, Y. (2015). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması.* Pegem Akademi.

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

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of statistical Software, 48*, 1-29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P. (2016). mirtCAT: Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications. *Journal of Statistical Software*, *71(5)*, 1–39. https://doi.org/10.18637/jss.v071.i05.

Cook, L. L., & Eignor, D. R. (1991). An NCME module on IRT Equating methods. *Educational Measurement: Issues and Practice, 10*(3), 191-199.

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

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). Harper & Row.

Cui, Z. (2006). *Two new alternative smoothing methods in equating: The cubic B-spline presmoothing method and the direct presmoothing method* (Publication No. 3229654) [Doctoral dissertation, University of Iowa]. ProQuest Dissertations & Theses Global.

Cui, Z., & 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

Cui, Z., & Kolen, M. J. (2009). Evaluation of two new smoothing methods in equating: The cubic B-spline presmoothing method and the direct presmoothing method. *Journal of Educational Measurement, 46(2)*, 135-158.

De Gruijter, D. N., & Leo, J. T. (2007). *Statistical test theory for the behavioral sciences*. Chapman and Hall/CRC.

DeMars, C. (2010). *Item response theory.* Oxford University Press.

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. CRC Press.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). Principles and practices of test score equating. *ETS Research Report Series, 2010*(2)*,* i-41. https://doi.org/10.1002/j.2333-8504.2010.tb02236.x

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists multivariate applications book series.* Lawrence Erlbaum Associates.

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *Journal of Educational Research, 94,* 275-283. https://doi.org/10.1080/00220670109598763

Felan, G. D. (2002). Test equating: Mean, linear, equipercentile, and ıtem response theory. *Annual Meeting of the Southwest Educational Research Association,* 1-24.

Finch, H. (2006). Comparison of the performance of varimax and promax rotations: Factor structure recovery for dichotomous items. *Journal of Educational Measurement, 43*, 39-52.

Finch, H., & French, B. F. (2019). A comparison of estimation techniques for IRT models with small samples. *Applied Measurement in Education, 32(2)*, 77-96.

Finch, H., French, B. F., & Immekus, J. C. (2014). *Applied psychometrics using SAS*. IAP.

Haberman, S. J. (1974). Log-linear models for frequency tables with ordered classifications. *Biometrics*, 589-600.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3)*,* 38-47.

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

Hanson, B.A. (1990). *An investigation of methods for improving estimation of test score distributions (Research Rep. No. 90-4)*. American College Testing.

Hanson, B. A., Zeng, L., & Colton, D. A. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating (No. 94)*. American College Testing Program.

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

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). American Council on Education/Praeger.

İnci, Y. (2014). *Örneklem büyüklüğünün test eşitlemeye etkisi* (Yayın No. 363203) [Yüksek lisans tezi, Hacettepe Üniversitesi]. YÖK. https://tez.yok.gov.tr/UlusalTezMerkezi/

Kahraman, N. (2013). Unidimensional interpretations for multidimensional test items. *Journal of Educational Measurement, 50(2),* 227-246. https://doi.org/10.1111/jedm.12012

Kahraman, N., & Kamata, A. (2004). Increasing the precision of subscale scores by usingout-of-scale information. *Applied Psychological Measurement, 28,* 407–426.

Kahraman, N., & Thompson, T. (2011). Relating unidimensional IRT parameters to a multidi-mensional response space: A review of two alternative projection IRT models for subscalescores. *Journal of Educational Measurement,48,* 146–164.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). American Council on Education & Praeger.

Kelecioğlu, H., & Öztürk Gübeş, N. (2013). Comparing linear equating and equipercentile equating methods using random groups design. *International. Online Journal of Educational Sciences, 5*(1), 227-241.

Kılıç, S. (2011). Neyin Peşindeyiz? Kutsal p değerinin mi (istatistiksel önemlilik) yoksa klinik önemliliğin mi? *Journal of Mood Disorders (1),* 46-48.

Kilmen, S. (2010). *Madde Tepki Kuramına dayalı test eşitleme yöntemlerinden kestirilen eşitleme hatalarının örneklem büyüklüğü ve yetenek dağılımına göre karşılaştırılması* (Yayın No. 279926) [Doktora tezi, Ankara Üniversitesi]. YÖK. https://tez.yok.gov.tr/UlusalTezMerkezi/

Kim, H.Y. (2014). *A comparison of smoothing methods for the anchor item nonequivalent groups design* (Publication No. 3638390) [Doctoral dissertation, University of Iowa]. ProQuest Dissertations & Theses Global. https://doi.org/10.17077/etd.qysisl6w

Kim, S. H., & Cohen, A. S. (2002). A Comparison of Linking and Concurrent Calibration Under the Graded Response Model. *Applied Psychological Measurement, 26*(1), 25-41. https://doi.org/10.1177/0146621602026001002

Kim, S. Y. (2018). *Simple structure MIRT equating for multidimensional tests* (Publication No. 10750515) [Doctoral dissertation, University of Iowa]. ProQuest Dissertations & Theses Global.

Kim, S., von Davier, A. A. , & Haberman, S. (2008). Small-sample equating using a synthetic linking function. *Journalof Educational Measurement, 45*(4), 325–342. https://doi.org/10.1111/j.1745-3984.2008.00068.x

Kline, R. B. (2015). *Principles and practice of structural equation modeling.* Guilford publications.

Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice, 7*(4), 29-37.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices.* Springer Science and Business Media.

Kolen, M. J., & Hendrickson, A. B. (2013). Scaling, norming, and equating. In *APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology* (pp. 201-222). American Psychological Association.

Lee, G., & Lee, W. C. (2016). Bi-factor MIRT observed-score equating for mixed-format tests. *Applied Measurement in Education, 29*(3), 224-241. https://doi.org/10.1080/08957347.2016.1171770

Lee, G., Lee, W., Kolen, M. J., Park, I. –Y., Kim, D. I., & Yang, J. S. (2015). Bi-factor MIRT true-score equating for testlet-based tests. *Journal of Educational Evaluation, 28,* 681-700.

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

Lee, W., & Brossman, B. G. (2012). Observed score equating for mixed-format tests using a simple-structure multidimensional IRT framework. In M. J. Kolen, & W. Lee (Eds.), *Mixed-format tests:*

Psychometric properties with a primary focus on equating (Vol. 2; CASMA Monograph No. 2.2.; pp. 115-142). Center for Advanced Studies in Measurement and Assessment, University of Iowa. Retrieved from https://education.uiowa.edu/sites/education.uiowa.edu/files/documents/centers/casma/publications/casma-monograph-2.2.pdf

Liu, C. (2011). *A comparison of statistics for selecting smoothing parameters for loglinear presmoothing and cubic spline postsmoothing under a random groups design*. (Publication No. 3461186) [Doctoral dissertation, University of Iowa]. ProQuest Dissertations & Theses Global.

Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement, 30*(1), 23–29.

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

Livingston, S. A., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement, 46*(3*),* 330–343. https://doi.org/10.1111/j.1745-3984.2009.00084.x

Lord, F. M., & Novick M. R. (1968). *Statistical theories of mental test scores.* Addison-Wesley.

McDonald R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24,* 99-114.

McLeod, L. D., Swygert, K. A., & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 189– 216). Erlbaum.

Mutluer, C. (2021). *Klasik test kuramına ve madde tepki kuramına dayalı test eşitleme yöntemlerinin karşılaştırması: Uluslararası öğrenci değerlendirme programı (PISA) 2012 matematik testi örneği* (Yayın No. 658052) [Doktora tezi, Gazi Üniversitesi]. YÖK. https://tez.yok.gov.tr/UlusalTezMerkezi/

Ogasawara, H. (2001). Standard errors of ıtem response theory equating/linking by response function methods. *Applied Psychological Measurement, 25* (1), 53– 67. https://doi.org/10.1177/01466216010251004.

Özçelik, D. A. (2010). *Eğitim programları ve öğretim.* Pegem Akademi.

Özkan, M. (2015). *Teog kapsamında uygulanan matematik alt testi ile matematik mazeret alt testinin istatistiksel eşitliğinin sınanması* (Yayın No. 396176) [Yüksek lisans tezi, Ankara Üniversitesi]. YÖK. https://tez.yok.gov.tr/UlusalTezMerkezi/.

Parshall, C. G., Houghton, P. D. B., & Kromrey, J. D. (1995). Equating error and statistical bias in small sample linearequating. *Journal of Educational Measurement,32*(1)*,* 37–54. https://doi.org/10.1111/j.1745-3984.1995.tb00455.x.

R Core Team (2019). *R: A Language and environment for statistical computing version 4.1. 1.* R Foundation for Statistical Computing. https://www.R-project.org/

Salmaner Doğan, R., & Tan, Ş. (2022). Madde tepki kuramında eşitleme hatalarının belirlenmesinde kullanılan delta ve bootstrap yöntemlerinin çeşitli değişkenlere göre incelenmesi. *Gazi University Journal of Gazi Educational Faculty (GUJGEF), 42*(2), 1053-1081. https://doi.org/10.17152/gefad.913241

Sansivieri, V., Wiberg, M., & Matteucci, M. (2017). A review of test equating methods with a special focus on IRT-based approaches. *Statistica, 77*(4), 329-352. https://doi.org/10.6092/issn.1973-2201/7066

Sass D. A., Schmitt T. A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research*, 45, 73-103. https://doi.org/10.1080/00273170903504810

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement,42*(4)*,* 309–330. https://doi.org/10.1111/j.1745-3984.2005.00018.x

Sunnassee, D. (2011). *Conditions affecting the accuracy of classical equating methods for small samples under the NEAT design: A simulation study.* (Publication No. 3473486) [Doctoral dissertation, University of North Carolina]. ProQuest Dissertations & Theses Global.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics (6th ed.)*. Pearson.

Tao, W., & Cao, Y. (2016). An extension of IRT-based equating to the dichotomous testlet response theory model. Applied Measurement in Education, 29(2), 108-121. https://doi.org/10.1080/08957347.2016.1138956

Team, R. C. (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria: Available at: https://www. R-project. org/.

Uğurlu, S. (2020). *Comparison of equating methods for multidimensional tests which contain items with differential item functioning* (Yayın No. 656957) [Doktora tezi, Hacettepe Üniversitesi]. YÖK. https://tez.yok.gov.tr/UlusalTezMerkezi/

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement, 12*(3), 239-252. https://doi.org/10.1177/014662168801200303

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). The chain and post-stratification methods for observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement, 41(1)*, 15-32. https://doi.org/10.1111/j.1745-3984.2004.tb01156.x

Zhang, Z. (2022). Estimating standard errors of IRT true score equating coefficients using imputed item parameters. *The Journal of Experimental Education, 90*(3), 760-782. https://doi.org/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. https://doi.org/10.1111/jedm.12210

Zhu, W. (1998). Test equating: What, why, how? *Research Quarterly for Exercise and Sport, 69*(1), 11-23. https://doi.org/10.1080/02701367.1998.10607662

Extended Abstract

Introduction

The test, which is defined as a measurement tool that is intended to numerically describe the degree or amount of a construct under standardized conditions (Kane, 2006), is widely used in education and psychology. Test developers and practitioners take many precautions to ensure the high validity and reliability of the scores obtained from large-scale tests. One of these measures is to create multiple test forms to prevent students from copying answers from each other. These forms are called alternative or parallel forms/tests. Parallel tests are defined as tests that measure the same construct or latent trait and have the same true score and error variances (De Gruijter & Leo, 2007). Test administrators act on the assumption that these test forms are parallel. The scores of students who take these tests at different times are used in the same assessments, and this creates a concern for all stakeholders involved with the tests against a sense of equality. Test equating is a statistical process that regulates the differences between test forms with similar content and difficulty levels so that the scores obtained from these forms can be used interchangeably (Kolen & Brennan, 2014).

Different equating designs can be used depending on the situation of obtaining the data to be used in the equating study. Single-group design, balanced single-group design, random-group design, and non-equivalent-groups common item design are equating designs used in data collection (Kolen & Brennan, 2014). In the random-group design, also known as the equivalent-groups design, respondents from a common population are randomly assigned to test forms, and two groups take different test forms (Cook & Eignor, 1991). In this study, a random group design was used.

Equating methods based on Classical Test Theory are divided into three categories: mean equating, linear equating, and equipercentile equating (Kolen & Brennan, 2014). Mean equating is the least rigid of the traditional equating methods (Sansivieri vd., 2017) and is perhaps the simplest approach to equalizing scores from two test forms (Finch, French, & Immekus, 2014). This equating method focuses only on the means of the two test forms administered. In contrast to mean equating, which assumes a constant difference between the two forms across the score scale (Sunnassee, 2011), linear equating is a method that allows the differences in difficulty between the two test forms to vary across the score scale (Kolen & Brennan, 2014). Linear equating is based on the assumption that the distributions of scores on forms X and Y are the same, except for differences in means and standard deviations (Crocker & Algina, 2008). Equipercentile equating is a method that requires the least assumptions among traditional equating methods, yet it is more complex than other methods (Finch, French, & Immekus, 2014). In the most general definition, this method determines which scores on two forms have the same percentile rank (Crocker & Algina, 1986).

Local independence, an assumption of Item Response Theory, means that an individual responds to items independently. In other words, the probability of an individual answering one item does not statistically affect the probability of answering other items (Crocker & Algina, 1986). In a test, if individuals' responses to items are not locally independent, another dimension causes dependence. In the data sets of this study, the violation of the local independence assumption was considered the cause of dimensionality.

There are two types of equating errors, random and systematic, in the literature (Felan, 2002; Kolen & Brennan, 2014). Random equating error, also known as the standard error of equating, is a type of error related to the sample characteristics of the test forms to be equated. As the sample size to which the test forms are to be equated increases, the standard error of equating becomes smaller and insignificant for very large samples. The other type of error is systematic equating error, which is also defined as equating bias. The factors underlying this error arise from neglecting the conditions and assumptions necessary to use equating methods (Kolen & Brennan, 2014).

In practice, two general methods have been developed to estimate the standard error. These methods are called bootstrap and delta methods (Kolen and Brennan, 2014). In the bootstrap method, many samples are taken from the available data and equating functions are estimated at each sample. Standard errors are calculated using the data from these many resamplings. The second method is the delta method, which uses the sampling statistics of procedures to estimate standard errors and is analytical because it results in an equation. The delta method is a widely used statistical method for estimating the standard errors of equating functions. Although the pattern and behavior of the standard errors of traditional equating methods have been widely investigated, and the results are well known (Tsai et al., 2001), there are no studies comparing the errors of traditional equating methods in terms of the methods by which the errors are obtained. Our research is supposed to contribute to the field in this respect.

This study aims to examine the delta and bootstrap methods used to determine the errors in traditional test equating methods in terms of sample size, number of items and the proportion of items loaded on the second dimension.

Method

It was aimed to compare the delta and bootstrap methods used to determine the errors in traditional test equating methods in terms of sample size, number of items and the proportion of items loaded on the second dimension, and it was planned to find the method that gives the least equating error by deriving data specific to the conditions determined for this purpose. The study is a simulation study due to these aspects. Random groups design was used in this study. In the random group design, respondents from a common population are randomly assigned to test forms, and two groups take different test forms (Cook & Eignor, 1991).

In the study, simulation conditions were determined as the number of items (20, 40, 60), sample size (250, 1000 and 5000) and percentage of items loaded on the second dimension (0%, 15%, 30% and 50%). A total of 36 (3x3x4) conditions were considered, including three conditions for sample size, three for test length, and four for the percentage of items loaded on the second dimension. Under these conditions, each test form was generated with 100 replications.

R statistical software, version 4.1.1.1 (R Core Team, 2019), was used to obtain the data to be used in our study. While generating data under R statistical software, the "simdata" command with the "for" loop in the "mirt" package (Chalmers, 2012) was used to generate univariate normal distributions for item responses of individuals and latent traits for one-dimensional data, and the "generate_pattern" command with the "for" loop in the "mirtCAT" package (Chalmers, 2016) was used to generate item responses of individuals and bivariate normal distributions for latent traits for two-dimensional data using 100 replications. For the current study, a two-dimensional structure called simple structure was considered. In our research, the correlation between the two dimensions was fixed at .5 in the datasets generated as a simple two-dimensional structure.

In the unidimensional data set, item discrimination parameters were generated from a log-normal distribution with a mean of 0 and a standard deviation of 0.2, while item difficulty parameters were generated from a normal distribution with a mean of 0 and a standard deviation of 1. θ parameters for individuals were derived from a normal distribution with a mean of 0 and a standard deviation of 1. The data sets in which the assumption of local independence was violated were generated according to the distribution obtained from the first form of the PISA 2018 Mathematics test Turkey sample. The item discrimination parameters of the items loaded on the first dimension were derived from a log-normal distribution with a mean of 0.53 and a standard deviation of 0.178; the discrimination parameters of the items loaded on the second dimension were derived from a log-normal distribution with a mean of 0.70 and a standard deviation of 0.21; and the item difficulty parameter was derived from a normal distribution with a mean of 0.327 and a standard deviation of 0.5. After the item parameters were created, ability parameters were created using bivariate normal distributions BN(0, 0, 0, 1, 1, 0.5). All data were generated under a unidimensional 2-parameter logistic model.

Test equating was performed using the "equate" package (Albano, 2016) in R software. For each equating procedure, 100 replications were also used to obtain bootstrap equating errors.

Results

At sample sizes of 250, 1000 and 5000, the equating method that produces less error than the other equating methods is the linear equating method, and the delta method produces less error than the bootstrap method at sample sizes of 1000 and 5000, and the bootstrap method produces less error than the delta method when the number of items in the test is 40 and 60 at sample sizes of 250. The linear equating method appears to be the most affected by the increase in the number of items in the test in the 250 sample size, which we can be characterized as a small sample. For all equating methods, it is clearly seen that the amount of error obtained from the equating results decreases as the number of participants in the test increases, and the errors obtained with the delta method for the linear equating method increase as the number of items in the test increases. Under the conditions of 1000 and 5000 sample sizes, the amount of error obtained with the bootstrap method in the linear equating method are found to decrease as the number of items in the test increase. It is also seen that the errors obtained for equipercentile equating and polynomial log-linear presmoothing methods tend to decrease as the number of items increases. Another striking finding is that the sample size most affected by the increase in the number of items in the test is the sample size of 250. In terms of the method used to determine the equating errors, it was found that the delta method performed better than the bootstrap method in linear

equating, the delta method produced fewer errors than the bootstrap method at sample size 250, and the bootstrap method produced fewer errors at sample sizes 1000 and 5000, although there were slight differences according to some of the conditions considered in equipercentile equating. It is seen that the bootstrap method produces less error than the delta method in the equating process performed with the polynomial log-linear presmoothing method.

It is seen that the condition that reduces the effect of the multidimensionality of the test structure on the equating result as a result of loading some of the items to the second dimension is the number of participants in the test. As the number of people in the test increases, it is seen that the errors obtained in the case of the test being unidimensional and multidimensional are close to each other. It was determined that the sample size most affected by the violation of the unidimensional structure of the test was the sample size of 250, and the most affected method under this sample size was the errors obtained by the delta method as a result of linear equating. In addition, it was understood that the ratio of items loaded on the second dimension did not differ systematically according to the equating methods, and different results were obtained under different conditions.

Conclusion, Suggestions and Recommendations

When traditional equating methods are examined in terms of sample size, it is seen that the linear equating method performs better than other equating methods. In the linear equating method, when the errors are calculated using the delta method, this method produces the slightest error at sample sizes of 1000 and 5000 compared to other equating methods, while at sample size 250, on the contrary, it produces more errors compared to other methods. At 250 sample size, the linear equating method was found to be the least error-producing equating method compared to other equating methods when equating errors were calculated using the bootstrap method. While this finding coincides with the findings of Kelecioğlu and Öztürk Gübeş (2013) and Özkan (2015), it contradicts the findings of İnci (2014). At sample sizes of 1000 and 5000, the linear equating method, which is the method that produces the least error, produces less error than the delta method than the bootstrap method when compared in terms of the method used to obtain errors, and under a sample size of 250, the bootstrap method produces less error than the delta method. In his research, Mutluer (2021) found that the delta method produces less error than the bootstrap method under the linear equating method. For all equating methods, it was found that the amount of error obtained from the equating results decreased as the number of participants in the test increased. Harris and Crouse (1993), Kim and Cohen (2002), Kilmen (2010), Lee and Ban (2010), İnci (2014), Salmaner Doğan and Tan (2022) found that the equating errors obtained from equating methods decrease as the sample size increases.

In terms of the number of items in the test, there are differences between the methods. In the linear equating method, if the method used to determine equating errors is delta, there is an increase in errors as the number of items increases, whereas in the bootstrap method, there is a decrease in equating errors as the number of items in the test increases. In addition, for equipercentile equating and polynomial log-linear presmoothing methods, it has been found that the errors obtained tend to decrease as the number of items increases. The linear equating method has been found to be the most affected by the increase in the number of items in the test.

It has been determined that the condition that reduces the effect of the multidimensionality of the test structure on the equating results is the sample size. As the sample size increases, it is seen that the errors obtained as a result of equating when the test is one-dimensional and multidimensional are close to each other. In the linear equating method, error scores for delta were found to be the least affected by the multidimensionality of the test structure under sample sizes of 1000 and 5000. It is seen that the condition where the effect of the multidimensionality of the test on the test equating results is the highest is the sample size of 250. At 250 sample size, the bootstrap method has been found to produce errors closer to the ones for one-dimensional than the delta method.

This study was conducted on simulated data. The research can be repeated by performing a test equating study on real data sets. Similar studies can be conducted using different equating designs instead of the random group design used in our study. In addition, this study can be further improved by adding variables that can be controlled, such as the difference in difficulty between forms, the degree of local independence assumption, and the difference in correlations between the two dimensions. Traditional equating methods can be compared with IRT test equating methods and test equating methods based on Multidimensional Item Response Theory. Finally, in addition to the error estimation methods examined, multiple imputation method can also be added to the study to contribute to the literature.

Yayın Etiği Beyanı

Bu araştırmanın planlanmasından, uygulanmasına, verilerin toplanmasından verilerin analizine kadar olan tüm süreçte "Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesi" kapsamında uyulması belirtilen tüm kurallara uyulmuştur. Yönergenin ikinci bölümü olan "Bilimsel Araştırma ve Yayın Etiğine Aykırı Eylemler" başlığı altında belirtilen eylemlerden hiçbiri gerçekleştirilmemiştir. Bu araştırmanın yazım sürecinde bilimsel, etik ve alıntı kurallarına uyulmuş; toplanan veriler üzerinde herhangi bir tahrifat yapılmamıştır. Bu çalışma herhangi başka bir akademik yayın ortamına değerlendirme için gönderilmemiştir.

Araştırmacıların Katkı Oranı Beyanı

Birinci Yazar Mehmet Fatih Doğuyurt %70, İkinci Yazar Prof. Dr. Şeref Tan %30 oranında katkı sağlamıştır.

Çatışma Beyanı

Araştırmanın yazarları arasında herhangi bir çıkar çatışması bulunmamaktadır. Ayrıca yazarlar, diğer kişi, kurum ya da kuruluşlarla herhangi bir çıkar çatışması içinde olmadıklarını beyan ederler.