International Refereed Journal

Karaelmas Journal of Educational Sciences

Journal Homepage: ebd.beun.edu.tr

The Investigation of the Group Invariance Property on Diverse Equating Methods[1]

Hande TANBERKAN SUNA[2], Şeref TAN[3]

Received: 06 November 2017, **Accepted:** 02 December 2017

ABSTRACT

In this study, we investigated the effect of group invariance property on Tucker linear equating, Levine linear equating, Braun-Holland linear equating, and equipercentile equating methods based on the classical test theory under the non-equivalent groups with anchor test. Two subforms of FVAT (Fast and Valid Aptitude Test) were applied to 3031 subjects that were used in the equating. The two subforms have 13 common items. The whole group was divided into gender subgroups to examine the effect of the group invariance. The results showed that Tucker and Braun-Holland linear equating methods produced equated scores below the acceptable limit of error and that these methods showed resistance to the assumption of group invariance. Furthermore, Levine linear equating and equipercentile equating methods generated equated scores above the acceptable limit of error.

Keywords: Test Equating, Group Invariance, Linear Equating, Equipercentile Equating.

EXTENDED ABSTRACT

Purpose and Significance

Tests are frequently used measurement tools in the education and psychology. There are cases where these tests are applied more than once. When a test is used more than once with multiple test forms, all stakeholders (such as students, practitioners, and teachers) in the process expect the test forms to be equivalent (Zhang, McDermorr, Fantuzzo & Gadsden, 2013). Because the concern of respondents in such situations is that test forms can be partially differentiated in terms of difficulty (Kolen & Brennan, 2014). The equating process is used to ensure that the different scores of test forms can be used interchangeable. Thus, as a concept of equating, it refers to a statistical process used to make more than one test form used in the same evaluation process comparable (von Davier, 2013). Regardless of the method used to test equating, there are some properties that must be met to equate test forms. Some of these properties are as follows: Symmetry property, equal construct property, equity property, equal reliability property and group invariance property. The group invariance property is related to have same results regardless of the respondent groups of the equating function. The test forms will be fair if this property is provided (van der Linden, 2000). For example, in a situation where the group invariance property is exactly provided, it is expected that there will be the same equating relation for two different groups (Kolen & Brennan, 2014). It has been necessary to investigate the effect of the group invariance property of the group invariance property due to these aspects. The aim of the study is to examine how the equating functions obtained from the subgroups in the cases where the RMSD indices differ in the equating methods based on the observed score.

Methods

[1] This study is derived from doctoral dissertation "The Effect of Group Invariance on Equating Functions of Diverse Equating Methods"

[2] Res. Assist, Bulent Ecevit University, handetanberkan@gmail.com

[3] Prof.Dr., Gazi University, Faculty of Education, sereftan4@yahoo.com

In present study, the role of group invariance on the equating functions obtained from the two subforms of the general ability test was examined. In this respect, the research is a comparative descriptive study because the role of group invariance on equating functions is examined by using different methods. Two subforms of FVAT (Fast and Valid Aptitude Test) were applied to 3031 respondents that were used in the equating. The Non-Equivalent-groups Anchor Test (NEAT) design was used according the structure of data. In this manner, the two subforms have 13 common items. The whole group was divided into gender subgroups to examine the effect of the group invariance.

Results

The results showed that Tucker and Braun-Holland linear equating methods generated equated scores below the acceptable limit of error and these methods showed resistance to the violation of group invariance. Furthermore, Levine linear equating and equipercentile equating methods generated equated scores above the acceptable limit of error. The findings were as follows: for the Tucker linear equating REMSD value was found as 0.0228; for the Levine linear equating method REMSD value was found as 0.2404; for the Braun-Holland equating REMSD value was found as 0.0571; for the unsmoothed equipercentile equating REMSD value was found as 0.1862; for the smoothed equipercentile equating REMSD value was found as 0.1403.

Discussion

As a result of the study, Tucker linear equating method in gender subgroups yield relatively minimum amount of error. The error value obtained after the equating with Braun-Holland linear equating method was also below the acceptable limit. These results were identical with the results of equating, performed without dividing the group into gender subgroups. It has been found that Tucker and Braun-Holland linear equating methods were to be the most accurate methods for the equating results obtained from the whole group. Accordingly, Tucker method yield more accurate equated scores in comparison with Levine, Dorans, Liu and Hammond (2008) also support this finding.

Grup Değişmezliği Özelliğinin Farklı Eşitleme Yöntemlerinde İncelenmesi[1]

Hande TANBERKAN SUNA[2], Şeref TAN[3]

Başvuru Tarihi: 06 Kasım 2017, **Kabul Tarihi:** 02 Aralık 2017

ÖZET

Bu çalışmada grup değişmezlik özelliğinin Klasik test kuramına dayalı gözlenen puan eşitleme yöntemlerinden Tucker lineer eşitleme, Levine lineer eşitleme, Braun-Holland lineer eşitleme, eşit yüzdelikli eşitleme yöntemleri üzerindeki etkisi araştırılmıştır. Bu doğrultuda FVAT (Fast and Valid Aptitude Test) testinin toplam 3031 kişiye uygulanan iki alt formu eşitleme çalışmasında kullanılmıştır. İki alt formun 13 ortak maddesi vardır ve eşitleme deseni olarak ortak maddeli eşit olmayan gruplar deseni seçilmiştir. Grup değişmezlik özelliğinin etkisini incelemek için tüm gruptan elde edilen veriler cinsiyet alt gruplarına bölünmüştür. Araştırmanın sonucunda Tucker ve Braun-Holland lineer eşitleme yöntemlerinin kabul edilen hata sınırının altında sonuçlar ürettiği ve grup değişmezlik özelliği varsayıma karşı daha dirençli olduğu; Levine lineer eşitleme ve eşit yüzdelikli eşitleme yöntemlerinin kabul edilen hata sınırının üstünde hata ile eşitleme yaptığı bulunmuştur.

Anahtar Kelimeler: Test Eşitleme, Grup Değişmezlik Özelliği, Lineer Eşitleme, Eşit Yüzdelikli Eşitleme.

1. Giriş

Testler, eğitim ve psikoloji alanlarında sıklıkla kullanılan ölçme araçlarıdır. Testlerin sonuçları bireysel ya da kurumsal kararlar almak için kullanılır. Örneğin kişinin, eğitim görmek istediği bir kuruma girmek için belli bir puana sahip olması ya da sıralamada belli bir oran içinde kalması bireysel düzeyde bir kararlar alınmasına destek olur. Kurumsal düzeyde karar ise, kişinin belli bir yeterliliğe sahip olup olmadığına dair kurumun kişiye sertifika ya da bir resmi belge vermesi olarak değerlendirilebilir (Kolen & Brennan, 2014; Jurich, DeMars & Goodman, 2012). Bu kararların verilebilmesi için objektif ölçme sonuçlarına ihtiyaç duyulmaktadır. Ölçme sonuçlarına ulaşmak için Akademik Personel Lisansüstü Eğitimi ve Giriş Sınavı (ALES), Yabancı Dil Bilgisi Seviye Tespit Sınavı (YDS) gibi testler yıl içinde birden fazla kez uygulanmaktadır. Dolayısıyla adaylar teste farklı zamanlarda katılabilmektedir ve testi uygulayan kurum da süreci daha etkin yönetebilmektedir. Testlerin birden fazla kez uygulanmasının avantajlı ve dezavantajlı yönleri bulunmaktadır. Test içeriğinin değişiklik yapılmadan birden fazla uygulanmasının yarattığı olumsuzluk, kullanılan test maddelerinin paylaşılması tehlikesinin doğması; bu anlamda geçerlik ve güvenirliği düşürücü etkilerin söz konusu olmasıdır. Buna benzer şekilde aynı kişiler aynı testi farklı zamanlarda birden fazla kez yanıtladığında, aktarma ya da hatırlama etkisi söz konusu olmaktadır. Bu dezavantajlı yönler ile baş etmek için birbirine mümkün olduğunca benzer test formları yapılandırılmaktadır (Andrulis, Starr & Furst, 1978; Haladyna & Dowing, 2004). Bu nedenle sıklıkla, aynı psikolojik yapıyı ölçen ve yapısal olarak birbirine eşdeğer olması için gayret gösterilen testler birlikte kullanılmaktadır (Green, 1995; Kolen & Whitney, 1982).

Birden fazla testin kullanıldığı durumlarda, süreçteki tüm paydaşlar test formlarının eşdeğer olmasını beklemektedir (Zhang, McDermorr, Fantuzzo & Gadsden, 2013). Çünkü bu tür uygulamalarda yanıtlayıcıların endişesi, test formlarının zorluk açısından kısmen farklılaşabilmesi üzerine olmaktadır (Kolen & Brennan, 2014). Farklı test formlarının eşdeğer olmasını sağlamak amacıyla eşitleme süreci kullanılmaktadır. Dolayısıyla eşitleme bir kavram olarak, aynı değerlendirme sürecinde kullanılan birden fazla test formunu karşılaştırabilir hale getirmek için kullanılan istatistiksel bir süreci ifade eder (von Davier, 2013). Farklı bir ifadeyle eşitleme, bir test formunun birim sistemini, diğer bir formun birim sistemine dönüştürme sürecini içermektedir (Angoff, 1971). Eşitleme işlemi bir testin alternatif formu ya da formları olduğunda kullanılır ve farklı formlardan elde edilen puanlar birbirleriyle karşılaştırılabilir hale getirilir. Test geliştiricileri, her ne kadar kapsam ve istatistiksel özellikler açısından mümkün olduğunca

[1] Bu çalışma "Grup Değişmezliği Özelliğinin Farklı Eşitleme Yöntemlerinde Eşitleme Fonksiyonları Üzerindeki Etkisi" adlı doktora tezinden üretilmiştir.

[2] Arş.Gör., Bulent Ecevit University, handetanberkan@gmail.com
[3] Prof.Dr., Gazi Üniversitesi, Eğitim Fakültesi, sereftan4@yahoo.com

benzer test formları geliştirmeye çalışsalar da, geliştirilen formlar birbirinden az ya da çok farklı olacaktır; dolayısıyla bu farklılaşma durumu kaçınılmazdır (Kolen & Brennan, 2014). Eşitleme, test formları arasındaki eşitliği sağlayarak; test formlarını birbirleri yerine kullanılacak hale getirme sürecini ifade etmektedir (von Davier, 2013).

Test eşitleme yöntemleri, aynı psikolojik yapıyı ölçen test formları arasındaki istenmeyen zorluk farklılıklarını elimine etmek için kullanılan ve farklı test formlarını alan yanıtlayıcıların sonuçlarının karşılaştırılabilir olmasını sağlayan yöntemlerdir (Dorans & Holland, 2000). von Davier (2013)'e göre ise test eşitleme yöntemleri; klasik test kuramına dayalı yöntemler, gözlenen puana dayalı yöntemler, madde tepki kuramına (MTK) dayalı yöntemler ve birden fazla türün bir arada kullanıldığı melez (hibrit) yöntemler olarak sınıflandırılır. Test eşitlemede kullanılacak yöntemden bağımsız olarak, test formları arasında eşitleme yapılması için sağlanması gereken bazı özellikler bulunmaktadır. Bu özellikler, farklı kaynaklarda ifade edilmiştir (Angoff 1971; Harris & Crouse 1993; Holland & Dorans 2006; Lord, 1980). Bu özellikler simetri özelliği, eş yapı özelliği, eşitlik özelliği, eş güvenirlik özelliği ve grup değişmezlik özelliğidir.

1.1 Grup Değişmezlik Özelliği

Grup değişmezlik özelliği diğer eşitleme özellikleri gibi eşitleme yapılırken sağlanması beklenen özelliklerden biridir. Bu özellik en genel anlamda eşitleme fonksiyonunun alt gruplar arasında değişmezliği ile ilgilidir. Eşitleme fonksiyonunun yanıtlayıcı gruplarına bağlı olmaksızın aynı sonucu vermesi beklenir. Bu özelliğin karşılanması, test formlarının adil olması açısından da önem teşkil etmektedir ve bu özelliğin karşılanmasının testlere yönelik adalet algısında ciddi bir rolü vardır (Van der Linden, 2000). Örneğin grup değişmezlik özelliğinin tümüyle sağlandığı bir durumda iki farklı grup için tamamen aynı eşitleme ilişkisinin bulunması beklenir (Kolen & Brennan, 2014). Bunu gerçekleştirmek için, her alt grup için eşitleme fonksiyonları ayrı ayrı hesaplanmakta ve ardından elde edilen fonksiyonlar karşılaştırılmaktadır (Powers, Turhan & Binici, 2012). Diğer bir ifadeyle, eşitleme sürecindeki dikkate alınan alt grupların test uygulama ve değerlendirme sürecindeki tüm faktörlerden aynı şekilde etkilediği kabul edilir. Ancak testin eşitlemeye uygun farklı formları için (yüksek güvenirlik katsayısı, aynı kapsam ve istatistiksel özellikler) bu durum, tümüyle sağlanması mümkün olmayan "ideal" bir durumu ifade ettiğinden grup değişmezlik özelliği asla tam olarak sağlanamasa da pratikte yüksek düzeyde sağlanması istenir. Grup değişmezliği özelliği aynı zamanda karşılaştırılan formların aynı yapıyı ölçme ve eşit güvenirlik katsayılarına sahip olma özelliklerini doğrulamayı da sağlar çünkü bu özelliklerle hem doğrudan hem de dolaylı olarak ilişki içindedir. Eğer formlar aynı yapıyı ölçmüyorlarsa ya da güvenirlikleri birbirinden ciddi ölçüde farklıysa sonuçların alt gruplar için değişmez olduğu iddia edilemez. Bu rolü sebebiyle, grup değişmezlik özelliğinin eşitlemenin en önemli özelliği olduğu kabul edilir (Harris & Kolen, 1986). Grup değişmezliği özelliğinin ihlal edilmesi, testin farklı kullanıcı grupları için karşılaştırılabilir olması ve adil olması gerekliliğine engel olmaktadır. Bu özelliğin sağlanmaması durumunda, karşılaştırılan alt gruplar için test formlarının güçlük düzeyi farklılaşmakta; bu durumda da bazı grupların aleyhine ya da lehine bir durum oluşmaktadır (Dorans, 2004; Huggins & Penfield, 2012; Powers, Turhan & Binici, 2012). Grup değişmezliğinin önemi özellikle alt gruplarda eşitleme fonksiyonlarının mümkün olduğunca benzer olmasında ortaya çıkmaktadır; bu nedenle grup değişmezliğinin değeri nicel olarak hesaplanabilmektedir. Dolayısıyla söz konusu beş özellik içinde grup değişmezlik özelliği, eşitleme için en önemli özellik olarak öne çıkmaktadır (Dorans & Holland, 2000). Buna karşın, grup değişmezliği özelliği istenen düzeyde sağlanmazsa dahi eşitleme fonksiyonları hesaplanabilir ve bu fonksiyonlar aracılığı ile puanlar eşitlenebilir. Bu nedenle grup değişmezliğinin eşitlemede sağlanmasının gerekmediğine yönelik görüşlerde mevcuttur. Ancak grup değişmezliğinin sağlanmadığı durumda yapılan eşitlemenin niteliği hakkında sorgulamalar yapılmaktadır (van der Linden, 2000).

Grup değişmezlik özelliğinin eşitleme için gerekli ancak yeterli bir koşul değildir (Dorans, Liu & Hammond, 2008). Eşitleme fonksiyonunun farklı alt gruplarda da aynı işleve sahip olması gerekliliği aşikâr olmasına rağmen test eşitleme literatüründe grup değişmezlik özelliğine verilen önemin farklılaştığı görülmüştür. Gerçek puana dayalı eşitleme yöntemlerinde grup değişmezlik özelliğinin test edilmesi ve sağlanması bir zorunlulukken, gözlenen puana dayalı eşitleme yöntemlerinde grup değişmezliğinin bir gereklilik olduğu vurgulansa da bu gerekliliğin sağlanmaması durumunda da eşitlemeye devam edildiği durumlar söz konusu olmaktadır. Dolayısıyla gözlenen puana dayalı eşitleme süreçlerinde grup değişmezliği gerekliliğinin sağlanmasının önemi hakkında genel bir görüş birliği bulunmamaktadır (Brennan, 2008). Grup değişmezlik özelliğinin, eşitleme sürecinde diğer özelliklerle de ilişkili olması,

sağlanmadığı durumda testler arasında kurulan ilişkinin "bağlantılılık"tan öteye gitmemesi ve yine bu özelliğin sağlanmadığı durumlarda önemli alt gruplar arasında uyum fonksiyonu geliştirme önerileri (Dorans & Holland, 2000) nedeniyle farklı eşitleme yöntemlerinde grup değişmezlik özelliğinin etkisinin araştırılması gerekli görülmüştür. Araştırmanın genel amacı, gözlenen puana dayalı eşitleme yöntemlerinde RMSD indekslerinin farklılaştığı durumlarda alt gruplardan elde edilen eşitleme fonksiyonlarının nasıl değişim gösterdiğini incelemektir. Bu genel amaç doğrultusunda aşağıdaki alt problemlere yanıt aranmıştır.

1. Klasik test kuramına dayalı gözlenen puan eşitleme yöntemlerinde alt gruplardan elde edilen eşitleme fonksiyonları, RMSD indekslerinin farklı düzeylerinde nasıl değişmektedir?

1.1. Tucker Lineer Eşitleme yönteminde erkek ve kadınlardan elde edilen eşitleme fonksiyonlarının sunduğu RMSD değerleri nasıl değişmektedir?

1.2. Levine Lineer Eşitleme yönteminde erkek ve kadınlardan elde edilen eşitleme fonksiyonlarının sunduğu RMSD değerleri nasıl değişmektedir?

1.3. Braun-Holland Lineer Eşitleme yönteminde erkek ve kadınlardan elde edilen eşitleme fonksiyonlarının sunduğu RMSD değerleri nasıl değişmektedir?

1.4 Düzgünleştirilmemiş eşit yüzdelikli eşitleme yönteminde erkek ve kadınlardan elde edilen eşitleme fonksiyonlarının sunduğu RMSD değerleri nasıl değişmektedir?

1.5. Düzgünleştirilmiş Eşit Yüzdelikli Eşitleme yönteminde erkek ve kadınlardan elde edilen eşitleme fonksiyonlarının sunduğu RMSD değerleri nasıl değişmektedir?

2. Yöntem

Bu araştırmada, genel yetenek testinin iki alt formundan elde edilen eşitleme fonksiyonları üzerinde grup değişmezliği özelliğinin rolü incelenmiştir. Farklı gruplar oluşturmak amacıyla veriler cinsiyet değişkenine göre ayrılmıştır ve bu gruplardaki eşitleme fonksiyonları farklı eşitleme yöntemleri kullanılarak çözümlenmiştir. Bu açıdan araştırma, farklı yöntemlerin kullanılmasıyla grup değişmezliğinin eşitleme fonksiyonları üzerindeki rolü incelendiğinden karşılaştırmaya dayalı betimsel araştırma niteliği gösterecek şekilde kurgulanmıştır.

2.1 Çalışma Grubu

Araştırmanın çalışma grubu, 2015-2016 yılları arasında Assessment Systems tarafından geliştirilmiş FVAT (Fast and Valid Aptitude Test) genel yetenek testinin iki paralel formunu cevaplayan toplam 3031 kişiden oluşmaktadır. FVAT'ın paralel olarak kullanılabilen iki formu olan v4 (eski form) ve v13 (yeni form)'ü yanıtlayanların cinsiyet dağılımları Tablo 1'de verilmiştir.

Tablo 1

FVAT v4 ve v13 Formlarının Alt Gruplara Göre Örneklem Sayıları

Değişken	v4 (Eski Form)	v13 (Yeni Form)	Toplam
Kadın	676	722	1.398
Erkek	913	720	1.633
25 Yaş ve Altı	1046	669	1.715
26 Yaş ve Üstü	543	773	1.316
Toplam	1.589	1.442	3.031

2.2 Veri Toplama Aracı

Araştırmada veri toplama amacıyla kullanılan ölçme aracı olan FVAT; sözel anlama, sayısal yargılama, analitik düşünme ve dikkat olmak üzere genel yeteneği oluşturan dört temel unsuru içeren bir testtir. Bu unsurlar, özellikle iş hayatında yüksek performans göstermek için önemli görülen genel yetenek özelliklerini temsil etmektedir. Testte her biri çoktan seçmeli ve ikili (0-1 yöntemiyle) puanlanan toplam

40 soru bulunmaktadır ve sözel anlama, sayısal yargılama, analitik düşünme ve dikkat unsurlarının her biri 10 soru ile temsil edilmektedir. İki ölçme ve değerlendirme uzmanından testlerin kapsam geçerliği ve ölçtükleri psikolojik özellikler konularında görüş alınıp uygun olduğuna karar verilmiştir. İkili puanlanan maddelerden oluşan ölçme araçlarının faktör yapısını test etmek amacıyla kullanılan Açımlayıcı (Explanatory) Faktör Analizi sonucunda söz konusu dört unsurun tek bir genel yetenek boyutu altında toplandığı ve dolayısıyla testin tek boyutlu olduğu görülmüştür.

2.3 Verilerin Analizi

Araştırma kapsamında kullanılan FVAT'ın, 13 ortak maddeye sahip olan iki formu (v4 ve v13) kullanılarak ortak maddeli eşit olmayan gruplar desenine göre eşitleme fonksiyonları elde edilmiştir. FVAT'ın yapısı ve araştırmada kullanılan iki formunun 13 ortak maddeye sahip olması, bu eşitleme desenin uygulanmasını mümkün kılmaktadır. Eşitleme fonksiyonlarının elde edilmesi için CIPE (Kolen, 2004) eşitleme programı kullanılmıştır. Eşitleme yöntemleri açısından, verinin yapısına uygun olan Klasik Test Kuramı temelli gözlenen puan eşitleme yöntemleri kullanılmıştır. Her bir yöntemle yapılan eşitlemede; kadın-erkek gruplarından elde edilen eşitleme fonksiyonları karşılaştırılmıştır. Karşılaştırılan gruplarda eşitleme fonksiyonları arasındaki farkı belirlemek için literatürde sıklıkla kullanılan, Dorans ve Holland (2000) tarafından önerilen RMSD ve REMSD indeksleri hesaplanmıştır.

Dorans ve Holland (2000), grup değişmezlik göstergesi için Root Mean Square Difference (RMSD) indeksini önermektedir. Uygulanan test formları X ve Y olarak adlandırıldığında y bu formdan alınan bir puan olmak üzere RMSD indeksi

$$RMSD(y) = \frac{\sqrt{\sum_j w_j [e_{\boldsymbol{P}_j}(y) - e_{\boldsymbol{P}}(y)]^2}}{\sigma_{X\boldsymbol{P}}}$$

olarak tanımlanır (Von Davier & Wilson, 2008). Burada P evreni, P_j alt evreni, w_j ise ∑_j▒ w_j =1 olmak üzere P evreninden çekilen alt evrenlerin oranını temsil ederken; e_(P_j) (y),P_j alt evreni üzerindeki eşitleme (bağlantılılık) fonksiyonunu, e_P (y) ise P evreni üzerindeki eşitleme (bağlantılılık) fonksiyonunu temsil etmektedir. σ_XP ise P evrenindeki X formundan elde edilen puanların standart sapmalarının oranıdır. RMSD(y) değerlerini özetleyen tek bir sayı elde etmek için karekök almadan önce P evreninde Y dağılımının ortalamasını alarak farklı bir ölçüm elde edilir. Elde edilen bu yeni indeks, Root Expected Mean Square Difference (REMSD) olarak adlandırılır ve

$$MSD(y) = \frac{\sqrt{E_p\{\sum_j w_j [e_{\boldsymbol{P}_j}(Y) - e_{\boldsymbol{P}}(Y)]^2\}}}{\sigma_{X\boldsymbol{P}}} = \frac{\sqrt{\sum_j w_j E_p\{[e_{\boldsymbol{P}_j}(Y) - e_{\boldsymbol{P}}(Y)]^2\}}}{\sigma_{X\boldsymbol{P}}}$$

formülü ile hesaplanır. Burada Y, P evreninden alınan tesadüfi bir Y puanıdır ve E_p {.} bu dağılımın ortalamasıdır. Bu ortalamayı bulmak için kullanılan dağılım, P üzerinde Y'nin süreksiz dağılımıdır (Dorans & Holland, 2000). REMSD istatistikleri yorumlanırken DTM (difference that matters) ve SDTM (standardized difference that matters) değerleri ile karşılaştırılır. REMSD istatistiklerinin, DTM değerlerinden düşük çıktığı yerlerde alt gruplar arası fark önemsiz olarak kabul edilir (Dorans & Feigenbaum, 1994; Yang, Dorans & Tataneni, 2003). SDTM değeri, DTM değerinin, eski formun standart sapmasına bölünmesiyle standartlaştırılır ve yorumlarken daha çok SDTM değeri kullanılır. SDTM değerinin yorumu, DTM değerinin yorumlandığı şekildedir. (Dorans, 2003; Öztürk Gübeş & Kelecioğlu, 2017). Diğer bir ifadeyle, maddeler bazında hesaplanan RMSD ve karşılaştırma bazında hesaplanan REMSD değerlerinin SDTM'yi aşması kabul edilebilir hata düzeyinin üstüne çıkıldığının göstergesidir.

3. Bulgular

FVAT v4 ve FVAT v13 formuna ait betimsel istatistikler Tablo 2'de özetlenmiştir.

Tablo 2
v4 ve v13 Formlarından Elde Edilen Verilere İlişkin Betimsel İstatistikler

	Ortalama	Standart Sapma	Çarpıklık	Basıklık	Mod	Medyan	Tüm test maddeleri ile ortak maddeler arası	
							Kovaryans	Korelasyon
v4	24.49	6.24	-0.3462	0.1612	26.00	25.00		
v13	22.94	6.29	-0.1094	-0.1020	22.00	23.00		
Vv4*	8.50	2.51	-0.4846	0.0700	-	-	13.73	0.87
Vv13**	8.16	2.23	-0.3623	-0.0070	-	-	11.56	0.82

** Vv4, v4 formundaki ortak maddeleri ifade etmektedir.*

*** Vv13, v13 formundaki ortak maddeleri ifade etmektedir.*

Tablo 2'ye göre FVAT testinin farklı formlarından elde edilen ortalamalar ve standart sapmalar birbirine yakındır. Bu sonuç, eşitlenecek olan test formları istatistiksel özellikler açısından benzer olmalıdır varsayımını da sağlamaktadır (Kolen& Brennan, 2014). Verilerin normal dağılımı çarpıklık/basıklık katsayılarıyla ve merkezi eğilim ölçüleriyle araştırılmıştır. v4 ve v13 formunun ortalama, mod ve medyan değerleri Tablo 2'de gösterilmiştir. Her iki form içinde merkezi eğilim ölçülerinin birbirine oldukça yakın olması dağılımların normal olduğunun bir göstergesidir (Tan, 2016). Çarpıklık ve basıklık katsayıları ise v4 ve v13 formlarından elde edilen verilerin normal dağılıma uygun olduğunu göstermektedir.

Tablo 3'de FVAT v4 ve v13 formu için maddelerin ortalama güçlüğü ile ortak maddelerin ortalama güçlüğü özetlenmiştir.

Tablo 3
v4 ve v13 Formlarındaki Tüm ve Ortak Maddelerin Ortalama Madde Güçlük İndeksleri

Test ve Form	Ortak Maddeler	Tüm Maddeler
FVAT v4	0.6542	0.6123
FVAT v13	0.6277	0.5742

Eşitleme yapılacak testlerdeki maddeler ile ortak maddeler kapsam ve istatistiksel açıdan birbirine benzer olmalıdır (Kolen & Brennan, 2014). Kapsam açısından benzerlik için belirtke tablosundan ve istatistiksel açıdan benzerlik için madde güçlük değerlerinden yararlanılmıştır. Bu bakış açısıyla ortak maddelerin ve testin tamamının ortalama madde güçlüğü hesaplanarak Tablo 3'de verilmiştir. FVAT testinin v4 formunun ortalama güçlüğü 0.6123 iken aynı formdaki ortak maddelerin ortalama güçlüğü 0.6542 olarak hesaplanmıştır. Benzer şekilde FVAT testinin v13 formunun ortalama madde güçlüğü 0.5742 ve bu formdaki ortak maddelerin ortalama madde güçlüğü 0.6277'dir. Bu istatistiklerin birbirine oldukça yakın olduğu söylenebilir.

3.1. Tüm Grup için Yapılan KTK'ya Dayalı Eşitleme Sonuçları

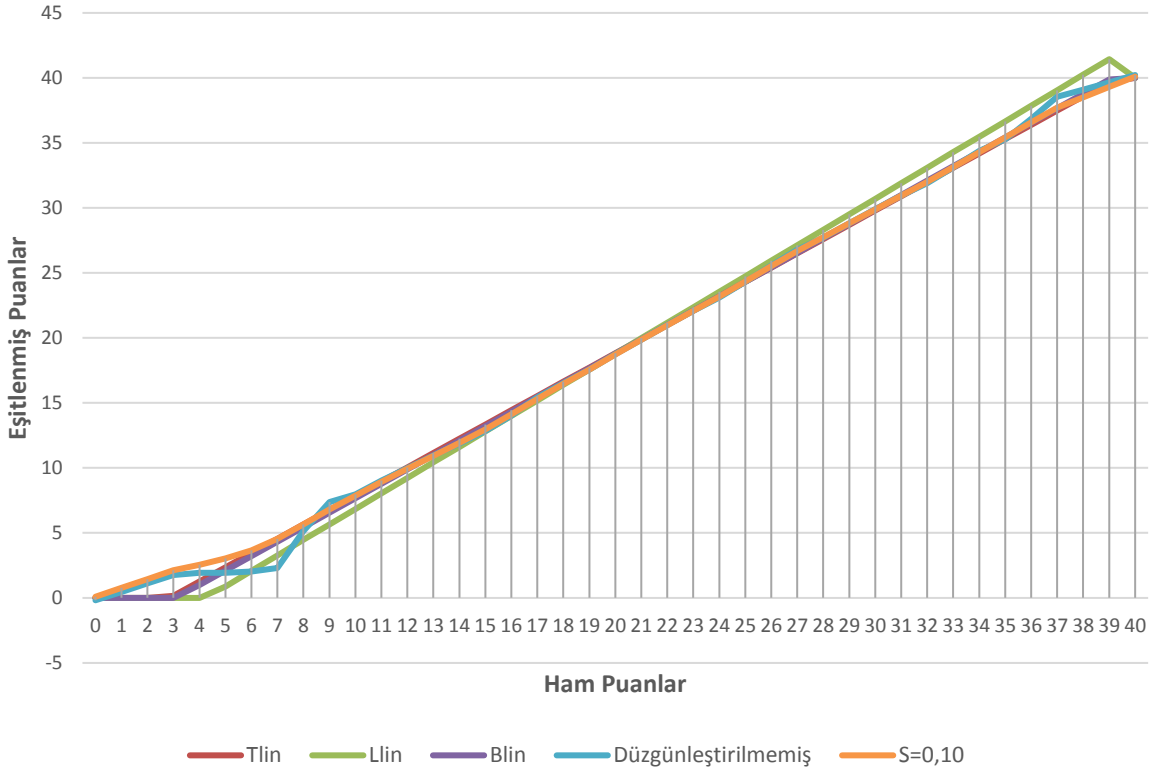
Klasik Test Kuramına dayalı yöntemler içinde eşit olmayan gruplar deseninde sık kullanılan eşitleme yöntemleri Tucker Lineer Eşitleme, Levine Lineer Eşitleme ve Braun-Holland Lineer eşitlemedir (Kolen ve Brennan, 2014). Temelde aynı kurama dayandıkları için benzer yapıları olan bu üç yöntemi ayrıştıran nokta varsayımlarıdır, bu nedenle yakın sonuçlar sağlamaktadırlar. Klasik Test Kuramına dayalı lineer yöntemler olan Tucker, Levine ve Braun-Holland ile elde edilen regresyon eşitliklerine ek olarak lineer olmayan ancak yine Klasik Test Kuramı'na dayalı olan Eşit Yüzdelikli Eşitleme yöntemiyle elde edilen eşitleme sonuçları Tablo 4'de verilmiştir. Her bir yöntem sütununda verilen puanlar ham puandan o yöntem kullanılarak eşitlenmiş puanları göstermektedir. Son düzgünleştirme işlemi için s=0.10 değerinin seçilme nedeni v4 formundan elde edilen momentlere en yakın değerleri sağlayan düzgünleştirme değeri olmasıdır. s=0.10'un seçilmesinin diğer bir nedeni düzgünleştirilmemiş eşit yüzdelikli eşitlemeden elde edilen ham puan dönüşümlerine en yakın değerleri vermesidir (Kolen & Brennan, 2014).

Tablo 4
KTK'ya Dayalı Eşitleme Yöntemleriyle Eşitlenmiş Puanlar ve Ham Puanlar

Ham Puanlar	T*lin**	L*lin***	B*lin****	Eşit Yüzdelikli Eşitleme	
				Düzgünleştirilmemiş	S=0.10
0	0	0	0	-0.18	0.09
1	0	0	0	0.46	0.77
2	0	0	0	1.11	1.45
3	0.16	0	0	1.75	2.13
4	1.26	0	1	1.94	2.55
5	2.36	0.88	2.11	1.94	3.03
6	3.46	2.07	3.22	2.03	3.66
7	4.55	3.27	4.33	2.31	4.55
8	5.65	4.46	5.44	5.13	5.66
9	6.75	5.65	6.55	7.38	6.81
10	7.85	6.84	7.67	7.96	7.87
11	8.94	8.04	8.78	9.02	8.92
12	10.04	9.23	9.89	9.98	9.93
13	11.14	10.42	11	10.94	10.92
14	12.24	11.62	12.11	11.88	11.92
15	13.33	12.81	13.22	12.84	12.97
16	14.43	14	14.33	14.03	14.12
17	15.53	15.19	15.44	15.43	15.31
18	16.63	16.39	16.55	16.5	16.46
19	17.72	17.58	17.66	17.58	17.58
20	18.82	18.77	18.77	18.75	18.72
21	19.92	19.97	19.88	19.85	19.85
22	21.02	21.16	20.99	20.99	20.97
23	22.11	22.35	22.1	22.07	22.07
24	23.21	23.54	23.21	23.1	23.18
25	24.31	24.74	24.32	24.32	24.34
26	25.41	25.93	25.43	25.57	25.53
27	26.51	27.12	26.54	26.77	26.68
28	27.6	28.31	27.66	27.72	27.75
29	28.7	29.51	28.77	28.82	28.82
30	29.8	30.7	29.88	29.91	29.89
31	30.9	31.89	30.99	30.95	30.94
32	31.99	33.09	32.1	31.89	32
33	33.09	34.28	33.21	33.1	33.13
34	34.19	35.47	34.32	34.38	34.29
35	35.29	36.66	35.43	35.28	35.42
36	36.38	37.86	36.54	36.84	36.57
37	37.48	39.05	37.65	38.54	37.71
38	38.58	40.24	38.76	39.1	38.5
39	39.68	41.44	39.87	39.66	39.3
40	40	40	40	40.22	40.1

Tlin***: Tucker Lineer Eşitleme
L**lin: Levine Lineer Eşitleme
****B**Lin***: Braun-Holland Lineer Eşitleme

Tablo 4'e göre eşitlenmiş puanlara bir örnek olması açısından, 21 ham puan Tucker lineer eşitleme yöntemine göre 19.92 puana, Levine Lineer Eşitleme yöntemine göre 19.97 puana, Braun-Holland lineer eşitleme yöntemine göre 19.88 puana, eşit yüzdelikli eşitleme yöntemine göre 19.85 puana ve düzgünleştirilmiş eşit yüzdelikli eşitleme yöntemine göre 19.85 puana karşılık gelmektedir. Diğer puan eşitlikleri de benzer şekilde yorumlanmaktadır. 21 ham puanı kullanılan veri seti için elde edilen norm puanıdır. Alt gruplar için yapılan eşitleme sonuçları sadece 21 puan için verilecektir.



Şekil 1. Ham Puanlar ve KTK'ya Dayalı Eşitleme Yöntemleriyle Elde Edilen Eşitlenmiş Puanlar

Şekil 16'da görüldüğü üzere, kullanılan dört yöntem özellikle 15-27 puan aralığında oldukça benzer sonuçlar sağlamakta ancak 0-14 ve 28-40 puan aralıklarında birbirinden kısmen farklı sonuçlar sağlamaktadırlar. Özellikle uç noktalarda eşitlenen puanlar arasındaki en büyük fark; Levine yöntemi ile S=0.10 kullanılarak yapılan son düzgünleştirme sonucu elde edilen Eşit Yüzdelikli Eşitleme yöntemi arasında olmuştur. En yakın eşitleme sonuçları ise Tucker yöntemi ile Braun-Holland yönteminden elde edilen eşitleme sonuçları arasındadır. Her iki formda da katılımcıların yığılım gösterdiği 14-28 puan aralığında diğer aralıklara göre gözlem sayısı çok daha fazla olduğu için bu aralıkta eşitlemenin daha hatasız olması, yöntemlerin benzer sonuçlar sağlaması beklenir bir durumdur.

3.2. Alt Gruplar için Yapılan KTK'ya Dayalı Eşitleme Sonuçları

Daha önce alt gruplara bölünmeksizin tüm veri setinde yapılan analizler, burada cinsiyet grupları için tekrarlanmıştır. Kadın ve erkek gruplarında yapılan eşitlemeler ve hata değerleri Tablo 5'de gösterilmiştir.

Tablo 5
v13 Norm Değeri Olan 21 Ham Puan için Eşitlenmiş Puanlar ve Hata Puanları

Ham Puan	T*lin*	Hata	L*lin*	Hata	B*lin*	Hata	Düzgünleştirilmemiş	Hata	S=0.10	Hata
Kadın										
21	19.86	1.14	20.57	0.43	19.71	1.29	19.52	1.48	19.43	1.57
Erkek										
21	20	1.00	19.61	1.39	19.91	1.09	20.22	0.78	20.25	0.75

Tablo 5'de görüldüğü üzere Tucker yöntemi kullanılarak eşitleme yapıldığında alt gruplar için hata puanları kadınlarda 1.14, erkeklerde ise 1.00 puandır. Levine yöntemi kullanılarak yapılan eşitlemede ise kadınlarda elde edilen hata miktarı 0.43, erkeklerde elde edilen hata miktarı ise 1.39 puandır. Braun-

Holland yönteminde ise kadınlarda 1.29 hata puanı bulunmuşken erkeklerde bu fark 1.09 puandır. Düzgünleştirilmemiş eşit yüzdelikli eşitlemede hata puanı kadınlarda 1.48, erkeklerde ise 0.78'dir. Düzgünleştirilmiş eşit yüzdelikli eşitlemede kadın grubunda 1.57 puan hata yapılmışken, erkeklerde 0.75 puan hata yapılmıştır.

KTK'ya dayalı dört yöntem aracılığıyla elde edilen eşitleme fonksiyonlarından elde edilen hataların cinsiyet grubuna göre RMSD ve REMSD değerleri Tablo 6'da verilmiştir. Tüm yöntemler için hesaplanan ve özet bir hata indeksi olan REMSD değeri ise her bir yöntem için verilmiştir.

Tablo 6
KTK'ya Dayalı Yöntemlerin Cinsiyet Alt Gruplarında RMSD ve REMSD Değerleri

PUAN	Tucker-Lineer Eşitleme	Levine-Lineer Eşitleme	Braun Holland-Lineer Eşitleme	Eşit Yüzdelikli Eşitleme-Düzgünleştirilmemiş	Eşit Yüzdelikli Eşitleme-S=0.10
0	0.0210	**0.7431**	**0.4182**	0.0362	0.0156
1	0.0202	**0.6026**	**0.2923**	0.0729	0.0473
2	0.0192	**0.4620**	**0.1653**	**0.1227**	0.0780
3	0.0199	**0.3324**	0.0393	**0.1825**	**0.1116**
4	0.0192	**0.2764**	0.0455	**0.2344**	**0.1359**
5	0.0181	**0.2718**	0.0434	**0.3045**	**0.1603**
6	0.0181	**0.2597**	0.0418	**0.3783**	**0.1746**
7	0.0177	**0.2481**	0.0397	**0.3994**	**0.1689**
8	0.0171	**0.2360**	0.0372	**0.3432**	**0.1452**
9	0.0160	**0.2238**	0.0352	**0.2849**	**0.1176**
10	0.0160	**0.2117**	0.0352	**0.2985**	**0.0983**
11	0.0160	**0.1997**	0.0332	**0.0821**	**0.0911**
12	0.0150	**0.1876**	0.0317	0.0636	**0.0985**
13	0.0150	**0.1754**	0.0299	**0.1257**	**0.0902**
14	0.0140	**0.1629**	0.0284	**0.1279**	0.0740
15	0.0140	**0.1517**	0.0267	0.0238	0.0526
16	0.0140	**0.1396**	0.0254	0.0257	0.0363
17	0.0131	**0.1274**	0.0238	0.0488	0.0200
18	0.0121	**0.1155**	0.0226	0.0061	0.0104
19	0.0131	**0.1033**	0.0214	0.0588	0.0430
20	0.0121	**0.0912**	0.0204	0.0741	0.0620
21	0.0112	0.0787	0.0194	0.0557	0.0652
22	0.0111	0.0665	0.0188	0.0704	0.0694
23	0.0112	0.0554	0.0171	0.0587	0.0778
24	0.0103	0.0432	0.0172	**0.0996**	**0.0864**
25	0.0095	0.0312	0.0169	**0.0851**	0.0755
26	0.0103	0.0191	0.0175	0.0412	0.0446
27	0.0096	0.0071	0.0176	0.0191	0.0131
28	0.0087	0.0056	0.0184	0.0160	0.0137
29	0.0096	0.0178	0.0187	0.0367	0.0320
30	0.0090	0.0299	0.0201	0.0429	0.0398
31	0.0085	0.0412	0.0207	0.0437	0.0382
32	0.0081	0.0530	0.0224	0.0310	0.0279
33	0.0085	0.0652	0.0232	0.0071	0.0151
34	0.0081	0.0773	0.0251	0.0243	0.0116
35	0.0079	**0.0895**	0.0272	0.0270	0.0393
36	0.0081	**0.1020**	0.0282	**0.1057**	0.0744
37	0.0079	**0.1142**	0.0304	**0.1547**	**0.0884**
38	0.0078	**0.1236**	0.0315	**0.1103**	0.0638
39	0.0078	**0.2338**	0.0338	0.0660	0.0383
40	0.0078	**0.3744**	**0.1450**	0.0216	0.0128
REMSD	0.028	**0.2404**	0.0571	**0.1862**	**0.1403**

**RMSD>SDTM, REMSD>SDTM değerleri kalın yazılmıştır.*

Tucker lineer eşitleme yöntemine göre REMSD değeri 0.028 (REMSD<SDTM), Levine lineer eşitleme yöntemi için REMSD değeri 0.2404 (REMSD>SDTM), Braun-Holland lineer eşitleme yöntemine göre REMSD değeri 0.0571 (REMSD<SDTM), Düzgünleştirilmemiş eşit yüzdelikli eşitleme için REMSD değeri 0.1862 (REMSD>SDTM) ve düzgünleştirilmiş eşit yüzdelikli eşitleme için REMSD değeri 0.1403 (REMSD>SDTM) olarak hesaplanmıştır. Bu sonuçlara göre, DTM değerini aşmayan eşitleme sonuçları sağlayan iki yöntem Tucker ve Braun-Holland lineer eşitleme yöntemleridir. Grup değişmezlik özelliğinin üst düzeyde sağlandığı alt gruplara bölünmeden her iki formun eşitlenmesi sürecine kıyaslandığında; verilerin cinsiyet alt gruplarına bölünmesi tüm yöntemlerde hata miktarlarını değişen miktarlarda artırmıştır. Grup değişmezliği özelliği bağlamında grupların birbirine yakın olduğu durumu temsil eden cinsiyet gruplarında Levine yöntemi ve Eşit Yüzdelikli Eşitleme yöntemi (düzgünleştirilmemiş ve S=0.10 düzgünleştirmesi halleri ile); Tucker ve Braun-Holland yöntemlerine kıyasla daha hatalı eşitlemeler yapılmasına neden olmuşlardır. Bu sonuç doğrultusunda, Levine yöntemi ve Eşit Yüzdelikli Eşitleme yöntemleri ile yapılan eşitlemelerde, düşük düzeyde de olsa grup değişmezliği özelliğinin azaldığı durumlarda hata miktarının arttığı görülmüştür. Tucker ve Braun-Holland yöntemleri ise, verilerin cinsiyet alt gruplarına bölünerek grup değişmezlik özelliğinin azalmasına karşın REMSD değerlerini SDTM'nin altında tutabilmiş, bu azalmaya daha dirençli olduklarını göstermişlerdir.

4. Tartışma ve Sonuç

Bu çalışmada eşitleme özelliklerinden grup değişmezlik özelliğinin farklı yöntemlerle elde edilmiş eşitleme fonksiyonlarını nasıl etkilediği bir geniş ölçekli genel yetenek testi verileri aracılığı araştırılmıştır. Çalışmanın sonucunda cinsiyet alt gruplarında Tucker lineer eşitleme ile en düşük REMSD değerinin elde edildiği ve dolayısıyla Tucker lineer eşitleme yönteminin en az hatalı sonucu verdiği bulunmuştur. Braun-Holland lineer eşitleme yöntemi ile yapılan eşitleme sonucu elde edilen hata değeri de yine kabul edilen sınırın altında kalmıştır. Bu sonuçlar grubu cinsiyet alt gruplarına bölmeden yapılan eşitleme sonuçları ile paralellik göstermektedir. Tüm grup için elde edilen eşitleme sonuçlarında da Tucker ve Braun-Holland lineer eşitleme en az hatalı yöntemler olarak gösterilmiştir. Dorans, Liu ve Hammond (2008) cinsiyet alt gruplarında Levine yönteminin Tucker yöntemine göre daha az grup değişmez olduğunu söylemişlerdir. Bu bulgu, çalışmada elde edilen sonuçları desteklemektedir. von Davier ve Han (2004) çalışmalarında kendi veri setleri için Tucker lineer eşitleme yönteminin Levine lineer eşitleme yöntemine göre daha az hatalı sonuçlar ürettiğini belirtmişlerdir; ancak genel bir değerlendirme yapıldığında lineer yöntemler arası farkın az olduğunu vurgulamışlardır. Eşitleme sonuçları değerlendirilirken Dorans ve Holland'ın (2000) önerdiği RMSD ve REMSD indekslerinden faydalanılmıştır. Bu indekslerin yorumlanması SDTM değeri ile karşılaştırılarak yapılmıştır. DTM=0.50 değerinin formun standart sapmasına bölünmesiyle elde edilen değer SDTM'dir. SDTM değerinin REMSD değerinden büyük çıkmadığı durumlarda grup değişmezlik özelliğinin sağlandığı ifade edilir. Levine lineer eşitleme, düzgünleştirilmiş ve düzgünleştirilmemiş eşit yüzdelikli eşitlemeden kabul edilen DTM değerinin üstünde hata puanları elde edilmiştir. Başka bir deyişle araştırmanın sonucunda bu eşitleme yöntemlerinin sağladığı eşitlenmiş puanların grup değişmezlik özelliğine karşı dirençli olduğu bulunmuştur. von Davier, Holland ve Thayer'e (2004) göre RMSD ve REMSD indeksleri testlerin eşitlenebilirliğini değerlendirmek için kullanılan önemli kriterlerdir. Eşitleme fonksiyonlarının gruba duyarlı olması testlerin güçlüklerinin gruplara göre değiştiğinin bir göstergesidir (Öztürk Gübeş ve Kelecioğlu, 2017). Grup değişmezlik özelliğinin test eşitleme sürecinin en önemli özelliği olarak kabul edilmesi (Harris ve Kolen, 1986) dolayısıyla araştırmacılara çalışma gruplarında farklılık oluşturabilecek alt gruplarda grup değişmezlik özelliğinin incelenmesi önerilmektedir. Ayrıca seçilen gruplama değişkeni ve eşitleme yöntemlerinin etkisi de araştırılabilir.

Kaynaklar

Andrulis, R. S., Starr, L. M., & Furst, L. M. (1978). The effect of repeaters on test equating. *Educational and Psychological Measurement*, 38, 341-349.

Angoff, W. H. (1971). *Scales, norms and equivalent scores*. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed.). American Council on Education: Washington, DC.

Brennan, R. L. (2008). A discussion of population invariance. *Applied Psychological Measurement*, 32(1), 102-114.

Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement*, 41, 43–68.

Dorans, N. J. (Ed.). (2003). Population invariance of score linking: Theory and applications to advanced placement program examinations (ETS Research Report RR-03-27). Educational Testing Service: Princeton, NJ.

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. (ETS Research Report RM-94-10). Educational Testing Service: Princeton, NJ.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37(4), 281-306.

Dorans, N. J., & Liu, J., Hammond, S. (2008). Anchor test type and population ınvariance: an exploration across subpopulations and test administrations. *Applied Psychological Measurement*, 32(1), 81-97.

Green, B. F. (1995). Comparability of scores from performance assessments. *Educational Measurement: Issues and Practices*, 14, 13-24.

Haladyna, T. M., & Dowing, S. M. (2004). Construct-ırrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practices*, 23(1), 17-27.

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

Harris, D. J., & Kolen, M. J. (1986). Effect of examinee group on equating relationships. *Applied Psychological Measurement*, 10, 35-43.

Holland P.W., & Dorans, N. J. (2006). "*Linking and equating*." In RL Brennan (ed.), Educational Measurement, (4th ed.). Westport, CT: Greenwood.

Huggins, A. C., & Penfield, R. D. (2012). An NCME ınstructional module on population ınvariance in linking and equating. *Educational Measurement: Issues and Practice*, 31: 27–40.

Jurich, D. P., DeMars, C. E., & Goodman, J. T. (2012). Investigating the impact of compromised anchor items on IRT equating under the nonequivalent anchor test design. *Applied Psychological Measurement*, 36(4), 291-308.

Kolen, M. J. (2004). Common Item Program for Equating (CIPE) [computer program]. Version 1100. http://www.education.uiowa.edu/centers/casma adresinden ulaşılmıştır.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking*. Springer Verlag: New York, NY.

Kolen, M. J., & Whitney, D. R. (1982). Comparison of four procedures for equating the tests of general educational development. *Journal of Educational Measurement*, 19, 279-293.

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

Öztürk Gübeş, N., & Kelecioğlu, H. (2017). Investigating group invariance of equating results. *Elementary Education Online*, 16(1), 217-227.

Powers, S., Turhan, A., & Binici, S. (2012). *Population invariance of vertical scaling results*. National Council of Measaurement in Education, Pearson: Vancouver, BC.

Tan, Ş. (2016). *SPSS ve excel uygulamalı temel istatistik-1*. PEGEM Akademi: Ankara.

van der Linden, W. J. (2000). A test-theoretic approach to observed-scored equating. *Psychometrika*, 65, 437-456.

von Davier, A. A., & Han, N. (2004). Population invariance and linear equating for the non-equivalent groups Design. (ETS Research Report Series No: 04-47). Princeton, NJ: Educational Testing Service.

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 ınvariance. *Journal of Educational Measurement*, 41(1), 15-32.

von Davier, A. A. (2013). Observed-score equating: An overview. *Psychometrika*, 78(4), 605–623.

Yang, W. L., Dorans, N. J., & Tateneni, K. (2003). Effect of sample selection on advanced placement multiple-choice score to composite score linking. (ETS RR-03-27). Educational Testing Service: Princeton, NJ.

Zhang, X., McDermott, P. A., Fantuzzo, J. W., & Gadsden, V. L. (2013). Longitudinal stability of IRT and equivalent-groups linear and equipercentile equating. *Psychological Reports: Measures and Statistics*, 113, 1303–1325.