Mersin Üniversitesi Eğitim Fakültesi Dergisi
Mersin University Journal of the Faculty of Education

Denk Olmayan Gruplarda Ortak Madde Deseni Kullanılarak Madde Tepki Kuramına Dayalı Eşitleme Yöntemlerinin Karşılaştırılması*

Comparison of IRT Equating Methods Using the Common-Item Nonequivalent Groups Design

Bilge GÖK**, Hülya KELECİOĞLU***

Özet: Bu araştırmada farklı koşullara göre türetilen test formlarını madde tepki kuramına dayalı kestirim yöntemlerini kullanarak eşitlemek ve bu yöntemlerden elde edilen sonuçları karşılaştırmak amaçlanmıştır. Araştırma iki ve üç parametreli lojistik modele uyumlu iki kategorili simülatif veriler kullanılarak yürütülmüştür. Eşitlemede "denk olmayan gruplarda ortak madde/test (NEAT) deseni" kullanılmıştır. Verilerin türetilmesinde WINGEN3 programından yararlanılmış ve araştırmada kullanılan 36 koşulun her biri için 50 tekrar yapılmıştır. Madde parametrelerinin kestirilmesi PARSCALE 4.1 ile ayrı kalibrasyon için test eşitleme ve ölçekleme IRTEQ ile yapılmıştır. Araştırmada elde edilen sonuçlar, eşitleme hatası (RMSE) ölçütüne göre değerlendirilmiştir. Araştırmanın sonunda, en düşük eşitleme hataları 3000 kişilik örneklemler, 80 maddelik testler, benzer yetenek dağılımına sahip gruplar ve ortalama-ortalama yönteminde elde edilmiştir. Ayrıca büyük örneklemler ile daha uzun testler kullanıldığında ve benzer yetenek dağılımına sahip gruplarda yöntemlerin daha az hatalı sonuç verdiği görülmüştür.
Anahtar Kelimeler: Test eşitleme, madde tepki kuramı, denk olmayan gruplar, ortak madde deseni.

Abstract: The purpose of this research was to equate the test for which were constructed in different conditions through scaling methods based on item response theory and to compare the results obtained from these methods. The research was conducted with using dichotomous simulated data which was consistent with two and three parameter logistic model. In order to equate two test forms "the common-item nonequivalent groups" was used in this research. WINGEN3 program was utilized for data generation and 50 replication were done for 36 different condition used in this research. PARSCALE 4.1 was utilized for the prediction of item parameters and IRTEQ was utilized for test equating and scaling in separate calibration. The results obtained from this simulation study were evaluated based on equating error (RMSE) criterions. The results revealed that, when the conditions evaluated generally, the best equating occurred in 3000-subjects samples, 80-item tests, groups have similar ability distribution, using and mean-mean methods. Moreover, the results indicated that methods had less equating errors when large sample sizes together with long tests were used in groups which had similar ability distributions under the conditions considered in this research.
Key Words: Test equating, item response theory, nonequivalent groups, common-item design.

GİRİŞ

Bireyleri okula ya da işe yerleştirmek amacıyla genellikle büyük ölçekli ve merkezi sınavlar uygulanır. Bu sınavların sonuçlarına dayalı olarak gerçekleştirilen yerleştirmeler de her yıl ya da yılda birkaç kez yapıldığından, bu sınavlar da belirli aralıklarla uygulanır. Bir işe ya da okula yerleştirmek için farklı zamanlarda yapılan sınavların amacı aynı olduğundan, her sınavda sorulan soruların güvenliğini sağlamak büyük bir sorun oluşturur. Bunun için aynı amaçla uygulanan testlerin pek çok formu geliştirilir. Aynı amaca yönelik olarak uygulanan sınavlarda

* Bu çalışma Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Eğitimde Ölçme ve Değerlendirme Ana Bilim Dalı için hazırlanan doktora tezinin bir kısmından özetlenmiştir.
** Arş. Gör. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, bilgeb@hacettepe.edu.tr
*** Prof. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, hulyaebb@hacettepe.edu.tr

farklı soruların sorulması, sınava giren adaylara bazı formların kolay, bazı formların zor gelmesine yol açabilir. Bir cevaplayıcının diğer bir cevaplayıcıdan daha zor bir test formu almasını engellemek ve aynı testin pek çok formu üzerinde karşılaştırılabilir puanlar elde etmek için test formlarının eşitlenmesine ihtiyaç duyulur (Cook & Eignor, 1991). Örneğin, aynı özelliği ölçen iki adet standartlaştırılmış test (A ve B) ve bu testlerin uygulandığı iki farklı grup (grup 1 ve grup 2) olduğunu varsayalım. Grup 1'de yer alan öğrenciler yalnızca A Testini, grup 2'de yer alan öğrenciler ise yalnızca B testini almış olsunlar. Grup 1 için A testine ilişkin ortalama puan 84; grup 2 için B testine ilişkin ortalama puan 80 olsun ve t testi sonucuna göre iki grubun ortalama puanları arasında anlamlı bir farklılık elde edilsin. Bu durumda, bu iki test ölçtüğü özellik açısından karşılaştırıldığında, grup 1'de yer alan öğrenciler grup 2'de yer alan öğrencilerden daha başarılıdır şeklinde bir sonuca varılabilir mi? İşte bu soruya verilecek cevap "test eşitleme" kavramının önemini ortaya koymaktadır. Bir seçme sınavında kesme puanının üzerinde puan alarak başarılı olan bir adaya aynı amaçla yapılan sınavın başka formu uygulandığında söz konusu aday, kesme puanının altında puan alabilir. Bu nedenle, her yıl tekrarlanan aynı amaca yönelik bu tür sınavların paralel olduğu varsayılan formlarının eşitlenmesi gerekir. Aynı özelliği ölçen farklı testlerden veya test formlarından elde edilen puanların karşılaştırılabilir olup olmadığının belirlenmesi, eğitimde önemli bir sorundur. Bu nedenle, test eşitlemeye en genel anlamda, aynı örtük özelliği ölçen testlerin farklı formlarından elde edilen puanları karşılaştırmak amacıyla ihtiyaç duyulur (Petersen, Marco & Stewart, 1982; Tsai, 1997; Wolkowitz, 2008).

Test Eşitleme

Test eşitleme, iki veya daha çok testten alınan puanlar arasındaki ilişkiyi ortaya çıkaran istatistiksel bir tekniktir. Puanlar arasındaki ilişkilere dayanarak, testler ortak bir ölçeğe yerleştirilir (Chu & Kamata, 2003). İki veya daha fazla testi ortak bir ölçeğe yerleştirmek için kullanılacak istatistiksel yöntem "test eşitleme" olarak adlandırılır ve eşitleme sonucunda formlardan elde edilen puanlar birbirlerinin yerine kullanılabilir (Hambleton & Swaminathan, 1985; Holland & Dorans, 2006; Kolen & Brennan, 2004).

Test geliştirme uzmanları içerik ve istatistiksel özellikler açısından benzer test formları oluşturmak için çabalasa da, bu test formları genellikle güçlük düzeyleri açısından farklılık gösterir. Güçlük düzeylerindeki bu farklılıkları açıklayabilmek için test formlarını eşitlemek gerekir (Dongyang, 2009). Genellikle büyük ölçekli sınavlarda içerik ve güçlük düzeyleri bakımından benzer olan farklı test formları kullanılır. Eğer eşitleme tam anlamıyla yerine getirilirse, bir cevaplayıcı tarafından alınan belirli bir test formunun, cevaplayıcının test puanı üzerinde sistematik bir etkiye sahip olması beklenmez ve elde edilen puanlar rahatlıkla karşılaştırılabilir (Nozawa, 2008). Böylece test sonuçlarına göre doğru kararlar alınabilir.

Eşitleme Desenleri

Eşitlenecek formlar oluşturulduktan sonra veri toplama desenine karar verilir. Eşitlemede veri toplama deseni "eşitleme deseni" olarak ifade edilir (Kolen & Brennan, 2004; Mao, von Davier & Rupp, 2006). Uygun bir eşitleme deseninin kullanılması eşitlemenin en önemli basamaklarından biridir (Holland & Dorans, 2006). Yaygın olarak üç farklı eşitleme deseni kullanılmaktadır. Bunlar; tek grup deseni, random grup deseni, denk olmayan gruplarda ortak madde/test (NEAT) desenidir (Crocker & Algina, 1986; Kolen & Brennan, 2004).

Tek grup deseni, en basit test eşitleme desenidir. Eşitlenecek iki test aynı cevaplayıcı grubuna verilir. Aynı cevaplayıcıların her iki testi de alması nedeniyle testlerin zorluk düzeyleri cevaplayıcıların yetenek düzeylerinden etkilenmez. Random grup deseninde gruplar random olarak ikiye ayrılır. Gruplara her iki test de uygulanır fakat her gruba sırasıyla farklı bir test uygulanır. Bu desenle, X ve Y formunu alan random eşdeğer grupların karşılaştırılmasına imkân sağlanmaktadır. Denk olmayan gruplarda ortak test/madde deseni, alanyazında en çok kullanılan desendir. Bu desen, denk olmayan gruplarda ortak maddeler üzerinde grupların performansını karşılaştırarak iki grup arasındaki eşitleme ilişkisini ortaya çıkarmada kullanılır. Bu desende ortak maddeler, grupların yeteneğindeki farklılıkları düzeltmede kullanıldıkları için eşitleme

fonksiyonunu belirlemede önemli bir rol oynamaktadır. Bu bakımdan, eşitleme çalışmaları yürütülürken ortak maddelerin özelliklerine ve sayısına/oranına dikkat edilmelidir. Bu araştırmada "denk olmayan gruplarda ortak madde/test deseni" kullanılmıştır.

Eşitleme Yöntemleri

Eşitleme yöntemleri yaygın olarak Klasik Test Kuramına (KTK) ve Madde Tepki Kuramına (MTK) dayalı eşitleme yöntemleri olarak sınıflandırılmaktadır. İki kuramın dayandığı varsayımlar ve matematiksel fonksiyonlar büyük ölçüde birbirinden farklıdır. Klasik Test Kuramına dayalı yöntemler; eşit yüzdelikli eşitleme (equipercentile), doğrusal (linear) eşitleme ve ortalama (mean) eşitleme olmak üzere üçe ayrılır (Barnard, 1996). Eşitleme yönteminin ve modelinin seçimi; eşitlemenin amacına, modelin doğruluğu ve uygunluğuna, cevaplayıcıların özelliklerine ve test verisine bağlıdır (Yang & Houang, 1996). Bu araştırmada madde tepki kuramına dayalı yöntemler kullanılmıştır.

Madde Tepki Kuramına Dayalı Yöntemler

Klasik test kuramına dayalı yöntemlerin tersine MTK'ye dayalı yöntemler, test formlarını eşitlemede her bir madde için doğrusal olmayan eğriler kullanır (Lord, 1980). Bu eğriler "madde karakteristik eğrileri" olarak adlandırılır. Madde karakteristik eğrisi, belirli bir yetenek düzeyindeki cevaplayıcının beklenen test puanını belirlemede kullanılır ve maddenin doğru cevaplanma olasılığı ile yetenek düzeyi arasındaki ilişkiyi gösterir. Madde karakteristik eğrisinin hangi matematiksel fonksiyonla gösterileceğini belirlemek için farklı modeller ortaya çıkmıştır. Bunlar, Rasch modeli, iki parametreli ve üç parametreli modellerdir.

Eşitleme sürecinde uygun olan MTK modeli seçildikten sonra, model parametreleri kestirilir. Tek grup veya denk grup deseni kullanılarak eşitleme yapıldığında, test formları aynı ölçekte olduğu için ek bir ölçeklemeye gereksinim duyulmaz. Denk olmayan gruplara dayalı eşitleme yönteminde ise, gruplar farklı olduğu için farklı test formlarından elde edilen parametreler aynı ölçekte olmayacaktır. Bu yüzden iki test formunu aynı ölçeğe yerleştirmek için doğrusal bir dönüştürme yapılmalıdır (Han, 2008; Kolen & Brennan, 1995). Farklı gruplardan elde edilen madde parametrelerini aynı ölçeğe dönüştürmenin "ayrı kalibrasyon/kestirim (separate calibration/estimation)"ve "eş zamanlı kalibrasyon/kestirim (concurrent calibration/estimation)" olmak üzere iki yolu vardır (Kolen & Brennan, 1995, 2004; Nozawa, 2008; Cao, 2008). Bu çalışmada ayrı kalibrasyon yöntemleri kullanılmış ve bu yöntemler aşağıda açıklanmıştır.

Ayrı Kalibrasyon/Kestirim Yöntemleri

Bu yöntem kullanıldığında denk olmayan gruplardan elde edilen madde parametreleri doğrusal olarak ilişkili fakat farklı bir ölçekte olacaktır (Hambleton & Murray, 1983; Akt. Hu, Rogers, Vukmiroviç, 2008). Parametreleri aynı ölçeğe yerleştirmek amacıyla denk olmayan gruplarda ortak madde deseninde yer alan ortak maddelerin madde parametreleri ölçek dönüştürmede kullanılır. Bu yöntemde iki farklı test formu için parametreler ayrı ayrı kestirilir. Elde edilen parametreler farklı ölçekte olduğundan karşılaştırılamaz. Bu karşılaştırmayı sağlamak amacıyla, ortak maddelerin a ve b parametrelerine dayalı olarak elde edilen A ve B sabitleri, bir formdaki θ değerini diğer forma dönüştürür. Ayrı kalibrasyon kullanıldığında A ve B sabitleri için, Test I'daki i kişisi için θ değerinin test J'deki karşılığı şu şekilde ifade edilir:

$$\theta_{Ji} = A\theta_{Ii} + B$$

θ_{Ji} ve θ_{Ii} : i kişisinin J ve I testine ait yetenek düzeyi

A: Denklemin eğimi

B: Denklemin sabiti.

Benzer şekilde iki testin madde parametreleri de dönüştürülür. Test I ve Test J'de yer alan madde parametrelerinin dönüşümü şu şekilde ifade edilir:

$$a_{Jj} = \frac{a_{Ij}}{A}$$

$$b_{Jj} = Ab_{Ij} + B$$

$$c_{Jj} = c_{Ij}$$

a_{Ij}, b_{Ij} ve c_{Ij} : j maddesi için I testinden (referans formdan) elde edilen madde parametreleri

a_{Jj}, b_{Jj} ve c_{Jj} : j maddesi için J testinden elde edilen sırasıyla madde ayırıcılık, madde güçlük ve şans parametreleri (Dönüştürülmüş madde parametreleri)

Yukarıdaki eşitlikte de görüldüğü üzere, c parametresi ölçek dönüşümünden bağımsızdır (Kolen & Brennan, 2004).

Bu kestirim sürecinde madde parametreleri iki farklı cevaplayıcı grubuna verilen iki farklı form için kestirildiğinde, yetenek parametreleri her iki grup için ortalaması 0, standart sapması 1 olacak şekilde ölçeklenir (Li, 2009). İki kategorili MTK modelleri için ortak madde deseninde ayrı kalibrasyonda ölçek dönüşümü için yaygın olarak kullanılan yöntemler şunlardır:

1. Moment yöntemleri
    a. Ortalama-ortalama (mean-mean) yöntemi (Loyd & Hoover, 1980),
    b. Ortalama-standart sapma (mean-sigma) yöntemi (Marco, 1977)
2. Karakteristik eğri dönüştürme (characteristic curve transformation) yöntemleri (Haebara, 1980; Stocking & Lord, 1983).

Ortalama-Ortalama (OO) ve Ortalama-Standart Sapma (OS) Yöntemleri

Ortak madde deseninde ölçek dönüştürmenin en basit yolu, ortak maddelerin madde parametrelerinin ortalaması ve standart sapmasını kullanmaktır. Her iki yöntem de ölçekleme sabitlerini hesaplamada kullanılır. Marco (1977) tarafından tanımlanan ortalama-sigma yönteminde, A ve B eşitleme katsayılarını kestirmek için ortak maddelerden elde edilen madde güçlük parametrelerinin ortalaması ve standart sapması kullanılır. A ve B katsayılarını kestirmek için kullanılan eşitlikler aşağıda verilmiştir:

$$A = \frac{\sigma(b_j)}{\sigma(b_I)} \qquad B = \mu(b_j) - A\mu(b_I)$$

$\sigma(b_I)$: I testindeki ortak maddelerin b parametrelerinin standart sapması

$\sigma(b_j)$: J testindeki ortak maddelerin b parametrelerinin standart sapması

$\mu(b_j)$: J testindeki ortak maddelerin b parametrelerinin ortalaması

$\mu(b_I)$: I testindeki ortak maddelerin b parametrelerinin ortalaması

Loyd ve Hoover (1980) tarafından tanımlanan ortalama-ortalama yönteminde ise A ve B eşitleme katsayılarını kestirmek için ortak maddelerden elde edilen madde ayırıcılık parametrelerinin ortalaması kullanılır. Bu yöntem için katsayılar şu şekilde hesaplanır:

$$A = \frac{\mu(a_I)}{\mu(a_j)}$$

$\mu(a_I)$: I testindeki ortak maddelerin a parametrelerinin ortalaması

$\mu(a_j)$: J testindeki ortak maddelerin a parametrelerinin ortalaması

İki yöntem, kestirim için farklı parametreleri kullandığından farklı sonuçlar verir. Bazen ortalama-standart sapma yöntemi ortalama-ortalama yöntemine tercih edilir; çünkü b parametresinin kestirimi a parametresinin kestiriminden daha kararlıdır. Bununla birlikte Baker ve Al-karni (1991) ve Ogasawara (2000) ortalama-ortalama yönteminin daha tercih edilebilir ve daha kararlı olduğunu çünkü ortalamaların standart sapmalardan daha kararlı olduğunu vurgulamışlardır. Diğer taraftan da Kolen ve Brennan (2004) ortalama-sigma yönteminin bazı durumlarda daha tercih edilebilir olduğunu belirtmektedirler. Bu nedenle, alanyazında her iki

yöntemle eşitleme yapılması ve iki yöntemden elde edilen ham/ölçek puanı dönüşümlerinin karşılaştırılması önerilmektedir.

Karakteristik Eğri Dönüştürme Yöntemleri

Ortalama-ortalama ve ortalama-sigma yöntemleri, madde karakteristik eğrileri benzer ancak parametreleri farklı olan maddeler için hatalı sonuçlara yol açabilir (Norman-Dvorak, 2009). I ve J testlerindeki kestirimlerde b parametreleri arasındaki farkın büyük ancak madde karakteristik eğrilerinin benzer olduğu bir madde için, ortalama-sigma yöntemi, b parametresi kestirimleri arasındaki farklılıktan etkilenir. Bu problem, ölçek dönüştürme yöntemleri madde parametre kestirimlerinin tümünü aynı anda dikkate almadığı için oluşur. Bu probleme cevap olarak Haebara (1980), madde parametrelerinin tümünü aynı anda göz önüne alan bir yöntem ortaya atmıştır. Daha sonra Stocking ve Lord (1983), Haebara'nın yöntemine benzer bir yöntem geliştirmiştir.

Bu yöntemler, test karakteristik eğrilerine dayalı eşitleme katsayılarını belirleyen gerçek puan eşitleme yöntemleridir (Baker & Al-Karni, 1991). Karakteristik eğri dönüştürme yöntemleri, ortak maddelerin madde ya da test karakteristik eğrileri arasındaki farkı azaltmak için geliştirilmiştir ve sadece NEAT deseniyle toplanan verilere uygulanabilir. İki yöntem genellikle benzer kestirimler sağlar ve özellikle madde ayırt ediciliklerinin dönüştürülmesinde daha iyi sonuçlar verir.

Stocking ve Lord (1983) yöntemi ortak maddelerin madde karakteristik eğrilerinin toplamı arasındaki farkı azaltır. Yaygın olarak kullanılan eşitleme yöntemlerden biri olan Stocking ve Lord (SL) karakteristik eğri yönteminde kayıp fonksiyon şu şekilde hesaplanır:

$$L(\theta_i) = \left[\sum_{j=1}^{m} p_{ij}(\theta_i, a_{X_{1J}}, b_{X_{1J}}, c_{X_{1j}}) - \sum_{j=1}^{m} p_{ij}(\theta_i, a^*_{Y_{2J}}, b^*_{Y_{2J}}, c_{Y_{2J}}) \right]^2$$

Stocking-Lord ve Haebara yöntemleri genellikle benzer sonuçlar ürettiği ve Stocking-Lord, Haebara'ya nazaran test etme uygulamalarında daha yaygın olarak kullanılan yöntemlerden birisi olduğu için, bu çalışmada Stocking-Lord yöntemi tercih edilmiştir (Way & Tang, 1991; Lee & Fitzpatrick, 2008; He, 2011). Bu alanda yapılan çalışmalar, karakteristik eğri yöntemlerinin ortalama-ortalama ve ortalama-sigma yöntemlerinden daha iyi olduğunu ve daha kararlı sonuçlar üretme eğiliminde olduğunu ortaya koymuştur (Baker & Al-Karni, 1991; Way & Tang, 1991; Hanson & Beguin, 2002; Kolen & Brennan, 2004). Bununla birlikte karakteristik eğri yöntemleri, hesaplama bakımından oldukça karmaşık olduğu ve ortalama-ortalama ve ortalama-sigma yöntemlerinden hesaplanan ölçekleme katsayılarındaki uygunluk daha iyi olduğu için ortalama-ortalama, ortalama-sigma yöntemleri hâlâ yaygın olarak kullanılmaktadır.

Eşitleme Hatası

Eşitleme hataları, eşitleme doğruluğunu tanımlamak için kullanılır. Hataların miktarı, eşitleme yöntemine ve desenine bağlıdır. Random ve sistematik olmak üzere iki tür eşitleme hatası vardır. Random eşitleme hatası, cevaplayıcı örnekleminden kaynaklanan bir hatadır ve eşitlemenin standart hatası ile tanımlanır. Diğer bir eşitleme hatası ise, sistematik hatadır. Eşitlemenin koşullarının, varsayımların ihlal edilmesinden ve yanlılıktan kaynaklanır (Zeng, 1991). Random hataya göre, sistematik hatayı kontrol etmek daha zordur (Kolen & Brennan, 1995). Sistematik hata, dikkatli bir test geliştirme, eşitleme desenlerinin yeterli şekilde uygulanması ve uygun istatistiksel tekniklerin kullanılmasıyla kontrol edilebilir. Toplam hata ise random hata ve sistematik hatanın toplamıyla tanımlanır (Kolen & Brennan, 2004).

Araştırmanın Önemi ve Amacı

Eşitlemenin doğruluğu, eşitleme yapılacak koşullara en uygun desenin ve yöntemin seçilmesine bağlıdır. Bu nedenle, bu çalışmada eşitleme hatasını etkileyen faktörlerden test uzunluğu, örneklem büyüklüğü, yetenek dağımı ve model türü ele alınarak eşitleme hatasının hangi durumlarda daha az olacağı belirlenmeye çalışılmıştır. Çalışmada, değişkenler üzerinde kontrol sağlamak amacıyla simülasyon verileri kullanılmıştır.

Lehman ve Bailey (1968) ampirik bir çalışmanın mümkün olmadığı ya da çok masraflı olduğu durumlarda, simülasyon çalışmalarının rahatlıkla yapılabileceğini ifade etmişlerdir. Oh (2000), genellikle analiz için ihtiyaç duyulan veri miktarının ancak simülasyonla sağlanabildiğini, bu yüzden de simüle edilmiş verinin eşitleme çalışmalarında sıklıkla kullanıldığını ifade etmiştir. Türkiye'de yapılan ve sürekliliği olan seçme sınavlarında (YGS, KPSS, SBS gibi) yoklanan davranışlar değişmemekle birlikte, sınav güvenliği nedeniyle her yıl önceki yılların paraleli olduğu varsayılan yeni bir test uygulanır ve bu testlerin hiçbirinde bir önceki yıl ile veya farklı test formları arasında ortak maddeler yer almaz. Farklı yıllarda sınava giren ya da aynı sınavı tekrar alan bireylerin denkliğini sağlamak güç olduğundan, denk grup deseninin uygulanması da problemli olmaktadır. Ayrıca ülkemizde uygulanan sınavlarda ortak maddeler yer almadığından ortak madde deseni kullanılarak eşitleme çalışması yapmak güçtür. Bu nedenle bu araştırmada bir simülasyon çalışması yoluyla MTK'ye dayalı kestirim yöntemlerinin performansı karşılaştırılmıştır.

Ülkemizde yapılan eşitleme çalışmalarında, genellikle gerçek veri ve tek grup deseni kullanıldığı görülmektedir (Bozdağ & Kan, 2010, Öztürk, 2010; Kan, 2011). Bununla birlikte ülkemizde ortak madde deseni kullanılarak yatay eşitleme üzerine yapılan simülasyon çalışmalarının sayısı ise yok denecek kadar azdır. Bu çalışmada veriler türetilirken Türkiye'deki büyük ölçekli sınavların eşitleme sürecine katkı sağlamak amacıyla, bu sınavların örneklem büyüklüğü, soru sayısı, yetenek dağılımı gibi özellikleri dikkate alınmıştır. Türkiye'de yapılan sınavların koşullarına göre, eşitleme yöntemlerinin performansının belirlenmesinin ülkemizde büyük ölçekli sınavlar üzerinde yapılacak eşitleme çalışmalarına da örnek olacağı düşünülmektedir. Bu nedenle araştırmada farklı koşullara göre türetilen test formlarını MTK'ye dayalı kestirim yöntemlerini kullanarak eşitlemek ve bu yöntemlerden elde edilen sonuçları karşılaştırmak amaçlanmıştır.

YÖNTEM

Araştırmanın Türü

Bu araştırmada, belirlenen koşullara göre türetilen veriyi kullanarak madde tepki kuramına dayalı kestirim yöntemlerinin karşılaştırılması amaçlanmıştır. Böylece en az hatalı yöntemler ve koşullar belirlenerek kuramsal çalışmalara katkıda bulunulması planlanmıştır. Araştırma kestirim yöntemlerini karşılaştırma amacı taşıdığından simülasyon verileri ile yapılan bir temel araştırma niteliği taşımaktadır.

Veri Toplama Deseni

Bu araştırmada iki test formunu eşitleyebilmek için "denk olmayan gruplarda ortak madde/test (NEAT) deseni" kullanılmıştır. NEAT deseni uygulamada yaygın olarak kullanılan desenlerden biri olmakla birlikte, en esnek ve en karmaşık desenlerden biridir (Sinharay & Holland, 2008). Bununla birlikte diğer desenlere göre daha az sınırlayıcıdır ve pratiklik açısından da tercih edilen bir yöntemdir (Zhu, 1998). Bu desende yeni test formu form X ile gösterilir, eski test formu ise form Y ile gösterilir. Form X ve form Y'de ortak maddeler (Z) vardır. İki cevaplayıcı grubu iki farklı populasyondan elde edilmiştir. Form X'i alan cevaplayıcı grubu grup 1, form Y'yi alan cevaplayıcı grubu ise grup 2 ile ifade edilir. Araştırmada standart normal dağılım için kullanılan NEAT deseni Şekil 1'de gösterilmiştir. Aynı durum çarpık dağılımlar için de geçerlidir.

Grup 1~ N(0,1) Form X	Grup 2~ N(0.5, 1) Form Y
X formuna ait maddeler	Y formuna ait maddeler
Ortak maddeler (Z)	Ortak maddeler (Z)

Şekil 1. *Araştırmada Kullanılan Eşitleme Deseni*

Simülasyon Faktörleri ve Koşulları

Çalışmada kullanılan simülasyon faktörleri; testlerdeki madde sayısı, her bir gruptaki cevaplayıcı sayısı, yetenek dağılımı ve veri türetmede kullanılan modelin türüdür. Uygulamada, bu faktörlerin eşitleme sürecini etkilediği ifade edilmektedir (Kang & Petersen, 2009). Simülasyon çalışmalarının koşulları, gerçek uygulamaları yansıttığı ölçüde bu uygulamalara genellenebilir. Bu nedenle, bu çalışmada ele alınan koşulların, ülkemizdeki büyük ölçekli sınavların özelliklerine uygun olmasına dikkat edilmiştir.

Araştırmada ele alınan ilk faktör test uzunluğudur. Test uzunluğu için her bir formda 30, 60 ve 80 maddenin yer aldığı üç koşul belirlenmiştir. Bu koşullar, ülkemizde uygulanan sınavlardaki ortalama madde sayıları temel alınarak belirlenmiştir (Bu sınavlar sırasıyla SBS, YGS ve ALES'tir). Her üç koşul için maddelerin en az %20'si ortak madde olarak alınmıştır. Bu nedenle n=30 koşulunda 10 madde, n=60 koşulunda 20 madde, n=80 koşulunda ise 30 madde ortak madde olarak düşünülmüştür. Genel olarak, ortak madde deseninde, maddelerin en az %20'si ortak madde olarak kabul edilmekle birlikte (Kolen & Brennan, 2004), yapılan çalışmalarda test uzunluğunun bir fonksiyonu olarak ortak maddelerin en fazla kaç olması gerektiği belirlenmemiştir (Sinharay & Holland, 2007). Bu araştırmada ortak madde sayısı testlere göre değişmekle birlikte, bir faktör olarak ele alınmamıştır.

Araştırmada ele alınan ikinci faktör, her bir gruptaki cevaplayıcı sayısıdır. Her bir grup için 1000 ve 3000 cevaplayıcı olmak üzere iki koşul belirlenmiştir. Bu konuda yapılan araştırmalar (Hanson & Beguin, 2002; Kim & Lee, 2004; Cao, 2008; Cui & Kolen, 2008; Zhao, 2008; Nozawa, 2009) seçilen bu örneklem büyüklüklerinin yeterli olduğunu göstermektedir. Alanyazında BILOG ve PARSCALE gibi programlarla yapılan analizlerde en az 1000 cevaplayıcı için madde parametrelerinin doğru bir şekilde kestirildiği de belirtilmiştir (Yen, 1987; Mislevy & Stocking, 1989).

Üçüncü faktör, grup 1 ve grup 2 için yetenek dağılımlarıdır. Bu faktör altında üç koşul ele alınmıştır. İlk yetenek dağılımı koşulunda Grup 1, ortalaması 0 standart sapması 1 olan standart normal dağılıma sahiptir. Grup 2'nin yetenek dağılımı ise grup 1'den 0.5 daha yüksektir ve grupların standart sapmaları eşittir. Ortalamalar arasında 0.5'lik farkın eşitleme üzerine grup farklılıklarının etkisini yansıtmak için yeterli büyüklükte olduğu gösterilmiştir (Li & Lissitz, 2000). Gerçek test etme uygulamalarında çarpık dağılımlara sıklıkla rastlanır (Kolen, 1985). Bu nedenle, ikinci yetenek dağılımı koşulunda, Grup 1'in dağılımı pozitif çarpık dağılıma; Grup 2 standart normal dağılıma sahiptir. Üçüncü yetenek dağılımı koşulunda ise Grup 1 negatif çarpık dağılıma; Grup 2 standart normal dağılıma sahiptir. Yapılan pek çok araştırmada çarpık yetenek dağılımlarının eşitleme yöntemleri üzerindeki etkisinin araştırılmasının önemli olduğu ifade edilmiştir (Kang & Petersen, 2009). Pozitif ve negatif çarpık dağılım koşullarında, yetenek parametreleri beta dağılımdan (pozitif çarpık dağılım, beta (6,14) dağılımdan, negatif çarpık dağılım beta (14, 6) dağılımdan) türetilmiştir. Dağılımlardan birinin pozitif çarpık, diğerinin ise negatif çarpık seçilmesinin nedeni ise Türkiye'de yapılan büyük ölçekli sınavların fen ve matematik gibi sayısal alt testlerinin pek çoğunda dağılım pozitif çarpık olmakla birlikte, Türkçe gibi sözel alt testlerinde ise genel olarak negatif çarpık dağılım gözlenmesidir (Bekci, 2007; Çetin, 2009; Öztürk, 2010). Çarpıklık katsayısı testlerde genellikle 0.30–0.40 arasında değiştiğinden, bu araştırmada da ortalama çarpıklık bu değerler arasındadır.

Dördüncü faktör, veriyi türetmede kullanılan modelin türüdür. Bu koşulda maddeler iki ve üç parametreli lojistik modele (2PLM ve 3PLM) göre türetilmiştir. Form X, form Y ve ortak maddeler için iki ve üç parametreli lojistik modele göre türetilen maddelerin sayısı Tablo 1'de verilmiştir:

Tablo 1. *İki ve Üç Parametreli Lojistik Modele Göre Türetilen Madde Sayısı*

Maddenin türetildiği model	Form X	Ortak maddeler	Form Y	Toplam
2PLM	20	10	20	30
3PLM	20	10	20	30
2PLM	40	20	40	60
3PLM	40	20	40	60
2PLM	50	30	50	80
3PLM	50	30	50	80

Araştırmada ele alınan dört simülasyon faktörü ve bu faktörlere ait koşullar çarpılarak koşulların tüm olası kombinasyonları incelenmiştir (3x2x3x2 olmak üzere toplam 36 kombinasyon).

Verilerin Türetilmesi (Simülasyon Çalışması)

Araştırmada hangi eşitleme yönteminin hangi koşullar altında daha iyi sonuçlar verdiğini ortaya koymak amacıyla Monte-Carlo simülasyon çalışması yürütülmüştür. Madde cevapları iki kategorili olarak Wingen3 programı ile türetilmiştir (Han, 2007). Araştırmada ele alınan her koşul için iki ve üç parametreli lojistik modellere dayalı veriler elde edilmiştir. Ortak maddelerin güçlük düzeyleri form X ve form Y'ye benzerdir.

Veri türetme süreci üç basamakta gerçekleştirilmiştir. Her bir basamak aşağıda açıklanmıştır:

a. Yetenek parametreleri: Yetenek dağılımları her bir grup için standart normal dağılımdan (θ~N(0,1)) örneklenmiştir. Ayrıca Grup 1 için pozitif ve negatif çarpık dağılım da kullanılmıştır.

b. Madde parametreleri: Araştırmada gerçeğe yakın parametre değerleri elde etmek için bu konuda yapılan diğer araştırmalarda kullanılan parametre değerlerinden yararlanılmıştır (Kim & Lee, 2006; Cao, 2008). Eşitleme için iki test formu türetilmiştir. Formlar ayrı ayrı türetilen ortak ve kendilerine özgü maddelerden oluşmaktadır.

Form X'in kendine özgü maddelere ait parametreler şu şekilde belirlenmiştir: Madde ayırıcılık parametreleri (a_i), ortalaması 0 standart sapması 0.5 olan log-normal dağılımdan türetilmiştir. Madde güçlük parametreleri (b_i) ortalaması 0 standart sapması 1 olan normal dağılımdan türetilmiştir. Şans parametresi (c_i) ise Kim ve Lee (2006) ile Cao (2008) tarafından da kullanıldığı gibi $\alpha=8$, $\beta=32$ olduğu beta dağılımından türetilmiştir.

Form Y için farklı maddelere ait madde parametreleri güçlük parametresi dışında form X'dekiyle aynı madde parametre dağılımlarından türetilmiştir. Eşitlemenin amacı, formlardaki güçlük farklılıklarını istatistiksel olarak düzeltmek olduğu için, madde güçlük parametrelerinin dağılımı form Y için form X'e göre 0.5 arttırılmıştır.

c. MTK modeli: Her bir form için madde parametreleri ve her bir grup için yetenek parametreleri belirlendikten sonra, iki ve üç parametreli MTK modellerine göre veriler türetilmiştir. Ülkemizde ve yurt dışında yapılan büyük ölçekli sınavlarda genellikle cevaplayıcıların şansla doğru cevabı bulmaları olası olduğu için üç parametreli model kullanılırken, şansla doğru cevabı bulma ihtimalinin olmadığı durumlarda yöntemlerin nasıl sonuçlar verdiğini görmek amacıyla 2PLM kullanılmıştır.

Eşitleme Sürecinin Uygulanması

Araştırmada kullanılan eşitleme desenine uygun olarak madde parametrelerini kestirmek için ayrı kestirim yöntemleri kullanılmıştır. Araştırmada kullanılan ayrı kalibrasyon yöntemlerinde her iki test formu için madde cevapları ayrı ayrı kalibre edilmiştir. İkinci form kalibre edilirken ortak maddeler için madde parametreleri ilk formun kalibrasyonunda oluşturulan değerlere sabitlenmiştir. Böylece ilk formun ölçeğine tüm madde parametreleri yerleştirilmiş ve ortak bir ölçek oluşturulmuştur. Bunun sonucunda da, Grup 2'nin yetenek dağılımı ortalaması 0 standart

sapması 1 olacak şekilde ölçeklenmiş ve madde parametreleri kestirilerek ortak bir ölçeğe yerleştirilmiştir.

Verilerin Analizi ve Değerlendirme Ölçütü

Madde parametrelerinin kestirilmesi PARSCALE 4.1 ile, ayrı kalibrasyon için test eşitleme ve ölçekleme IRTEQ ile yapılmıştır. Araştırmada bilinen evren parametreleri ile çalışıldığından, madde parametreleri expected a posteriori (EAP) ile kestirilmiştir.

Araştırmada daha kararlı kestirimler elde etmek amacıyla her bir koşul için ve her bir yöntem için veri türetme ve eşitleme süreci 50 kez tekrarlanmıştır. Alanyazında her bir koşul için farklı eşitleme yöntemlerinden elde edilen sonuçları karşılaştırmak amacıyla 50 tekrarın yeterli olduğu (Harwell, Stone, Hsu & Kirisci, 1996; Hanson & Beguin, 2002; Hu ve diğerleri, 2008) görülmüştür.

Eşitleme sonuçlarının doğruluğunu değerlendirmek için RMSE (Root mean square error-Eşitleme hatası) kullanılmıştır.

$$RMSE(\tau_j) = \sqrt{\frac{\sum (\tau^*_{jr} - \tau_j)^2}{R}}$$

τ_j : j parametresinin gerçek değeri

τ_{jr}^* : tekrar edilen veri seti (r=1,….R) için j parametresinin kestirilen değeri

R: tekrar sayısı

RMSE indeksi, toplam eşitleme hatasını (random ve sistematik hatanın toplamı) ifade eder. Parametre veya yetenek kestirimlerindeki değişkenliği değerlendirmek için 50 kez yapılan tekrarlardan elde edilen RMSE değerlerinin ortalaması alınmıştır.

$$\sum_{i=1}^{I} RMSE(\tau_j) / I$$

I: rapor edilen parametreye bağlı olarak kişi veya madde sayısı

Ortalama RMSE değerleri küçüldükçe daha doğru eşitleme sonuçları elde edilir. İyi bir yöntem küçük yanlılığa ve küçük RMSE değerine sahip olmalıdır (Chu & Kamata, 2003). Çalışmada ilk olarak eşitlemede kullanılan yöntemlerin (OO, OS ve SL) eşitleme hatalarına bakılmıştır. Daha sonra çalışmada kullanılan koşulların etkilerini belirlemek amacı ile her bir yöntem için ANOVA yapılmıştır. Koşullar ve aralarındaki tüm etkileşimlerle bir model kurulduğunda hataların serbestlik derecesi sıfır olduğundan dolayı analiz yapılamamıştır. Bu nedenle ana etkiler ile 2, 3 ve 4 yönlü etkileşimler test edilmiştir. Çok fazla sayıda anlamlılık testi uygulandığından, ortaya çıkan hatayı kontrol etmek amacı ile Bonferroni düzeltmesi kullanılmıştır (anlamlılık düzeyi .002). Buna ek olarak, değişkenlerin yöntemlerin performansı üzerindeki etkilerini göstermek amacı ile eta kare değerleri de rapor edilmiştir. ANOVA'da anlamlı bulunan etkiler için post-hoc testleri yapılmıştır.

BULGULAR ve YORUM

Tablo 2'de OO, OS ve SL yöntemlerinin araştırmada ele alınan koşullara göre -yetenek dağılımı, örneklem büyüklüğü oranı, test uzunluğu ve model türüne- göre madde parametrelerinin eşitleme hataları ve bu hatalara ilişkin ortalama değerler verilmiştir.

Tablo 2. *Koşullara Göre Madde Parametrelerinin Eşitleme Hataları (RMSE)*

Koşullar		*a parametresi*			*b parametresi*		
		OO	*OS*	*SL*	*OO*	*OS*	*SL*
Yetenek Dağılımı	normal	0.141	0.168	0.146	0.071	0.197	0.133
	pozitif	0.337	0.333	0.344	1.275	1.306	1.520
	negatif	0.382	0.409	0.387	1.167	1.181	1.421
Örneklem Büyüklüğü	1000	0.314	0.316	0.313	0.785	0.845	0.981
	3000	0.259	0.290	0.271	0.891	0.945	1.068
Test Uzunluğu	30	0.295	0.307	0.313	0.956	0.993	1.177
	60	0.287	0.292	0.281	0.817	0.880	0.967
	80	0.278	0.311	0.280	0.741	0.811	0.930
Model	2PLM	0.293	0.318	0.287	0.762	0.859	0.968
	3PLM	0.280	0.289	0.297	0.913	0.930	1.080
Ortalama		0.287	0.303	0.292	0.838	0.895	1.024

Tablo 2'de koşullara göre en düşük ortalama hatanın a parametresinde OO yönteminde; en yüksek ortalama hatanın OS yönteminde; b parametresinde ise en düşük ortalama hatanın OO yönteminde en yüksek ortalama hatanın SL yönteminde olduğu görülmektedir. Elde edilen bu sonuç Baker ve Al-Karni (1991) yaptıkları çalışmada OO yönteminin OS yöntemine göre daha iyi eşitleme sonuçlarına yol açtığını dair elde edilen sonuç ile tutarlılık göstermektedir. Bununla birlikte elde edilen bu sonuç, Ogasawara (2001) tarafından yapılan çalışmada elde edilen en iyi yöntemin SL olduğu bulgusuyla çelişmekte, en çok hataya sahip yöntemin ise OS olduğuna dair araştırma bulgusuyla benzerlik göstermektedir. Way ve Tang (1991) tarafından yapılan çalışmada ise OS yönteminin OO ve SL yöntemlerinden daha az kararlı kestirimler sağladığı bulunmuştur. Her üç eşitleme yönteminde ve her iki parametrede de en düşük ortalama hata yetenek dağılımının normal olduğu durumda; en yüksek hatanın ise a parametresinde yetenek dağılımının negatif carpık olduğu durumda, b parametresinde ise pozitif carpık olduğu durumda ortaya çıkmıştır.

Yetenek dağılımına göre madde parametreleri incelendiğinde normal ve çarpık yetenek dağılımlarında en az hataya sahip yöntem OO olurken; en yüksek hataya sahip yöntem ise OS/SL olduğu görülmüştür. Keller ve Keller (2008) 3PLM'yi kullanarak yaptıkları araştırmada çarpık yetenek dağılımına sahip gruplarda OS yönteminin en çok hataya sahip yöntem olduğunu bulmuşlardır. Gruplar arasındaki farkların fazla olduğu pozitif ve negatif çarpık dağılımlar için, moment yöntemlerinin (OO ve OS) SL yönteminden daha iyi sonuçlar verdiği görülmektedir.
Örneklem büyüklüğü açısından bakıldığında da her iki parametre için de en iyi yöntem OO olurken, en yüksek hataya sahip yöntem SL olmuştur. Elde edilen bu sonuç Hanson ve Beguin (2002) tarafından yapılan araştırmada 3PLM kullanıldığında Stocking-Lord ve Haebara yöntemlerinin, ortalama-ortalama ve ortalama-standart sapma yöntemlerinden daha düşük hata ürettiğini dair bulunan sonuç ile tutarlılık göstermemektedir. Genel olarak yöntemler test uzunluğuna göre incelendiğinde ise a ve b parametresi için en iyi yöntem OO olurken, en hatalı yöntemin 30 test uzunluğunda OS 60 ve 80 maddelik test uzunluklarında ise SL yöntemi olduğu görülmüştür. Araştırmadan elde edilen bu sonuç, Kim ve Lee (2006), Kim ve Kolen (2006), Baker ve Al-Karni (1991), Speron (2009), Hanson ve Beguin (2002) tarafından yapılan araştırmada en iyi yöntemin SL en hatalı yöntemin ise OS yöntemi olduğuna dair elde edilen sonuç ile hem benzerlik hem de farklılık göstermektedir.

Eşitleme yöntemlerinin, a ve b parametresinin eşitleme hataları aralarında fark olup olmadığını belirlemek amacı ile yapılan ANOVA sonucunda, yöntemler arasında istatistiksel olarak anlamlı fark olmadığı bulunmuştur (RMSEa $F_{(2,107)}$= .134, p=.874, RMSEb $F_{(2,107)}$= .909, p=.406). Yöntemler arasında istatistiksel olarak fark çıkmamasının nedeni, 0.00-1.00 aralığındaki verilere ANOVA uygulanmasından kaynaklanıyor olabilir.

Çalışmada kullanılan koşulların etkilerini belirlemek amacı ile her bir yöntemin eşitleme hataları ayrı ayrı analiz edilmiştir. Tablo 3'de anlamlı bulunan etkilerin yöntemlere göre F ve eta değerleri verilmiştir.

Tablo 3. *Madde Parametrelerine Dayalı Eşitleme Hataları İçin ANOVA Sonuçları*

Hatalar			YÖNTEMLER					
			OO		OS		SL	
	Etkiler	sd	F	η^2	F	η^2	F	η^2
RMSEa	Yetenek dağ(YD)	2	39.650*	.832	25.731*	.763	47.630*	.856
	Orneklem Büyüklüğü	1	5.615*	.260	.861	.051	3.738	.189
	Test uzunluğu (TU)	2	.171	.021	.180	.022	.980	.109
	Model	1	.333	.020	1.062	.062	.251	.015
	YD*OB	2	4.004*	.334	3.479	.303	2.202	.216
	YD * TU		.407	.092	.289	.067	.990	.198
	YD*model		4.710*	.371	3.587	.310	2.210	.216
	OB * TU	2	.229	.028	1.371	.146	.994	.110
	OB * model	1	18.241*	.533	8.841*	.356	15.846*	.498
	TU * model		.868	.098	.706	.081	2.345	.227
RMSEb	Yetenek dağ(YD)	2	325.224*	.976	181.011*	.958	220.734*	.965
	Orneklem Büyüklüğü	1	6.147*	.278	3.677	.187	2.057	.114
	Test uzunluğu (TU)	2	8.669*	.520	4.133*	.341	6.535*	.450
	Model	1	12.568*	.440	1.870	.105	3.464	.178
	YD*OB	2	6.358*	.443	3.661*	.314	3.822*	.323
	YD * TU		2.135	.348	1.231	.235	1.626	.289
	YD*model	2	4.046*	.336	.139	.017	.107	.013
	OB * TU		1.145	.125	1.173	.128	.563	.066
	OB * model	1	.370	.023	.811	.048	.000	.000
	TU * model		.401	.048	.092	.011	.663	.077

p< .002

Tablo 3'de de görüldüğü gibi yetenek dağılımı her iki parametrede de her üç yöntemde de manidardır. Yetenek dağılımlarının benzer olduğu standart normal dağılıma sahip gruplardan elde edilen hata değerleri, yetenek dağılımları arasındaki farkların fazla olduğu pozitif ve negatif çarpık dağılımlardan elde edilen değerlerden manidar düzeyde düşüktür. Araştırmada elde edilen bu sonuç Kim ve Lee (2006)'nin araştırma sonucuyla benzerlik göstermektedir. Dağılımın pozitif veya negatif çarpık olduğu durumlarda, madde parametrelerini kestirmek problemli olabilir. Çarpık dağılımlar kendi içinde karşılaştırıldığında, a parametresi için negatif çarpık yetenek dağılımlarında hata değerlerinin yüksek olduğu, b parametresi için ise pozitif çarpık dağılımlardan hatanın daha yüksek olduğu görülmektedir. Yöntemler örneklem büyüklüğü açısından incelendiğinde, örneklem büyüklüğü etkisinin sadece OO yönteminde manidar olduğu, diğer yöntemlerde ise manidar olmadığı görülmektedir. a parametresi için örneklem büyüklüğü arttıkça hata değerlerinin düştüğü gözlenirken, b parametresinde hata değerleri artmaktadır. Bu durumun sonucu olarak örneklem büyüklüğünün a parametresi üzerinde b parametresine göre daha büyük ve olumlu yönde bir etkiye sahip olduğu söylenebilir. Araştırmadan elde edilen bu sonuç Swaminathan ve Gifford (1983), Wingersky ve Lord (1984), Skaggs ve Stevenson (1989)'un üç parametreli modele göre türetilen veriler üzerinde yaptıkları araştırmada elde ettikleri sonuç ile benzerlik göstermektedir. Bununla birlikte standart normal dağılımda a parametresi için hata değerlerinin tüm yöntemlerde düşük ve benzer performans gösterdiği görülmüştür. Araştırmadan elde edilen bu sonuç Keller, Kim, Nering ve Keller (2007) tarafından yapılan araştırmada bulunan sonuç ile benzerlik göstermektedir.

Araştırmada ele alınan yöntemler test uzunluğuna göre incelendiğinde, a parametresinde hata değerleri arasında manidar fark olmadığı, b parametresinde ise anlamlı fark olduğu Tablo 4'te görülmektedir. Genel olarak tüm yöntemlerde test uzunluğu arttıkça hata değerlerinin düştüğü görülmektedir. Hata değerleri, test uzunluğu 30'dan 60'a çıktığında 60'dan 80'e

çıktığına göre daha fazla düşmüştür. Elde edilen bu sonuç Han (2008) tarafından üç parametreli modele uyumlu veriler üzerinde yapılan çalışmada elde edilen sonuçla benzerlik göstermektedir.

Araştırmadan elde edilen sonuçlar model türüne göre incelendiğinde, a parametresinde hata değerleri arasında manidar fark olmadığı, b parametresinde ise sadece OO yönteminde manidar fark olduğu Tablo 4'te görülmektedir. Bu çalışmada, a parametresine ilişkin sonuçlarda üç parametreli modelin iki parametreli modelden daha iyi sonuçlar verdiğini, b parametresinde ise durumun tersi olduğu sonucu ortaya çıkmıştır. Hem iki hem de üç parametreli modeldeki madde parametreleri incelendiğinde ise a parametresi b parametresinden daha kararlı kestirimlere yol açmıştır. Araştırmadan elde edilen bu sonuç Ironson (1983) ile Li, Tam ve Tompkins (2004)'in yaptıkları araştırmada elde ettikleri sonuçlarla tutarlılık gösterirken, Lord (1983) ile tutarlılık göstermemektedir. Lord (1983) tarafından yapılan çalışmada a parametresinin b parametresinden daha zayıf kestirildiği görülmüştür. Bu konuya ilişkin olarak alanyazında zayıf bir şekilde kestirilen madde güçlüklerinin MTK'ye dayalı eşitleme sonuçları üzerinde ciddi bir etkiye sahip olduğu ifade edilmektedir (Stocking & Lord, 1983). Kaskowitz ve De Ayala (2001) yaptıkları çalışmada a ve b parametreleri için 3PLM'nin 2PLM'den daha yüksek hataya sahip sonuçlar verdiğini bulurken, Domaleski (2006) madde parametreleri açısından model türüne göre yöntemler arasındaki farkın oldukça küçük olduğunu bulmuştur.

SONUÇ ve ÖNERİLER

Bu araştırmada madde tepki kuramına dayalı eşitleme yöntemlerinden OO, OS ve SL yöntemlerinin test uzunluğu, örneklem büyüklüğü, yetenek dağılımı ve model türüne göre karşılaştırılması amaçlanmıştır. Bu amaçla Türkiye'deki gerçek veri uygulamalarına benzer koşullarda veriler türetmeye çalışılarak ele alınan koşullarda hangi yöntemin en az hatalı sonuçlar verdiği incelenmiştir.

Araştırmadan elde edilen sonuçlar doğrultusunda; örneklem büyüklüğü ve test uzunluğu arttıkça daha kararlı madde parametre kestirimleri elde edilmiştir. Bu doğrultuda, en kararlı parametre kestirimlerine örneklem büyüklüğü 1000'in üzerinde olduğunda ve test uzunluğu da 60'ın üzerinde olduğu durumlarda ulaşılabileceği söylenebilir. Benzer yetenek dağılımına sahip gruplarda ise farklı yetenek dağılımına sahip gruplara göre yöntemlerin hata değerlerinin daha düşük olduğu bulunmuştur. Madde parametreleri açısından yöntemler genel olarak incelendiğinde a ve b parametresi için en az hatalı yöntemin OO yöntemi olduğu görülürken, en fazla hataya sahip yöntemin ise a parametresi için OS, b parametresi için ise SL yöntemi olduğu görülmüştür.

Araştırmadan elde edilen sonuçlar genel olarak değerlendirildiğinde, en iyi eşitlemelerin 3000 kişilik örneklemler, 80 maddelik testler, benzer yetenek dağılımına sahip gruplar ve OO yöntemi kullanılarak elde edilebileceği ifade edilebilir. Ayrıca araştırmada ele alınan koşullar doğrultusunda, büyük örneklemler ile daha uzun testler kullanıldığında ve benzer yetenek dağılımına sahip gruplarda yöntemlerin daha az hatalı olduğu sonucuna da ulaşılmıştır.

Bu araştırmadan elde edilen sonuçların başka çalışmalara genellenmesi sadece bu çalışmada ele alınan koşullarla (örneklem büyüklüğü, test uzunluğu, yetenek dağılımı ve model türü) sınırlıdır. Bu kapsamda değerlendirildiğinde, araştırmada kullanılan koşulların birbirleriyle etkileşimi de araştırmanın sonuçlarını etkileyen unsurlardan bir diğeridir. Araştırmada madde parametrelerinin değerlendirilmesinde yetenek dağılımının eşitleme sonuçları en önemli etkiye sahip olduğu tespit edilmiştir. Bu doğrultuda, araştırmadan elde edilen sonuçlar belirli bir eşitleme yöntemini seçerken önemle üzerinde durulması gereken unsurun yetenek dağılımı olduğunu göstermiştir. Grupların denk olduğu ya da benzer olduğu durumlarda yöntemler denk ya da benzer olmayan gruplara göre daha iyi performans göstermiştir. Bu çalışmanın sonucunda yöntemlerin farklı koşullarda farklı sonuçlar verdiği görülmüştür. Farklılığın miktarı ve hangi yöntemin tercih edileceği iki grubun yetenek dağılımının benzerliği, örneklem büyüklüğü, test uzunluğu ve model türüne bağlıdır. Bu doğrultuda, araştırmada ele alınan bu koşullar dikkate alınarak kullanılacak eşitleme yönteminin belirlenmesinin oldukça önemli olduğu sonucuna

varılmıştır. Yöntemlerdeki farklılıklar sonucunda, elde edilen sonuçlar bir yöntemi diğerine tercih etmek için yeterince kararlı değildir ve her koşulda etkili tek bir yöntem yoktur.

Çalışmanın sonuçları farklı koşul ve durumlarda diğer yöntemlerden üstün olan tek bir yöntem olmadığı gibi, hangi yöntemi seçmenin en iyi sonuçları vereceği konusunda da açık bir kanıt yoktur; fakat eşitleme çalışmalarında elde edilen sonuçların daha önceki çalışmaların sonuçlarıyla tutarlılığı bir yöntemi seçmek için en bilgilendirici yoldur. Hanson ve Beguin (2002) pek çok eşitleme yöntemini birlikte kullanmanın ve sonuçları karşılaştırmanın uygulamada oldukça etkili olduğunu ifade etmişlerdir. Bu durum farklı koşullarda en uygun yöntemin seçilmesine yardımcı olacaktır.

Bu çalışma, tüm eşitleme sürecinin sadece bir parçasını içermektedir. Test etme programlarında ham-ölçek puanı dönüştürmeleri eşitlemenin son basamağı olarak yürütülür ve ölçek puanları beklenen toplam puanlar yerine rapor edilir. Bu yüzden test geliştirme uzmanları ölçek puanını etkileyen faktörler üzerinde çalışabilir. Bu çalışmada formlar benzer güçlüktedir. Formların güçlükleri farklı olduğunda yöntemlerdeki farklılıklar incelenebilir.

Bu konuda yapılacak araştırmalarda ortak madde uzunlukları, denk gruplar ve daha büyük örneklemler (örneğin gerçek veriye uygun olarak 10.000 kişi gibi) simülasyon koşulu olarak ele alınıp, bu koşulların yöntemler üzerindeki etkisi incelenebilir. Eşitleme yöntemleri birbirleriyle karşılaştırılırken ve değerlendirilirken RMSE (toplam hata) dışında random eşitleme hatası ya da eşitlemenin standart hatası gibi farklı değerlendirme indeksleri kullanılabilir. Bu indeksler kullanıldığında, bu çalışmada ele alınan yöntemlerin performansı değişebilir. Bu nedenle bundan sonra yapılacak araştırmalarda random hata ya da standart hata kullanılarak yöntemlerdeki farklılıklar araştırılabilir.

Araştırmada MTK'ye dayalı eşitleme yöntemlerinden OO, OS ve SL yöntemleri karşılaştırılmıştır. Bu konuda araştırma yapacak kişilere farklı eşitleme desenleri ve farklı eşitleme yöntemlerini kullanmaları önerilebilir. Bu çalışmada eşitleme yöntemlerinin karşılaştırılması simülasyon verisi kullanılarak yapılmıştır. Simülasyon çalışması ile birlikte gerçek veri kullanılarak da benzer çalışmalar yapılabilir ve farklı türde iki veri setinden elde edilen sonuçlar karşılaştırılabilir.

KAYNAKLAR

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

Barnard, J. J. (1996). *In search for equity in educational measurement: traditional versus modern equating methods*. Paper presented at ASEESA's national conference at the HSRC Conference Centre, Pretoria, South Africa.

Bekci, B. (2007). *Orta Öğretim Kurumları Öğrenci Seçme ve Yerleştirme Sınavı'nın değişen madde fonksiyonlarının cinsiyete ve okul türüne göre incelenmesi*. Yüksek Lisans tezi, Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Bozdağ, S. & Kan, A. (2010). Şans başarısının eşitlemeye etkisi. *H.Ü. Eğitim Fakültesi Dergisi, 39*, 91-108.

Cao, L. (2008). *Mixed format test equating: Effects of test dimensionality and common-item sets.* Unpublished doctorate thesis, University of Maryland.

Chu, K. L. & Kamata, A. (2003). *Test equating with the presence of DIF*. Paper presented at the annual meeting of American Educational Research Association, Chicago.

Chui, Z. & Kolen, M. (2008). Comparison of parametric and nonparametric bootstrap methods for estimating random error in equipercentile equating. *Applied Psychological Measurement, 32*(4), 334-347.

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

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.

Domaleski, C. S. (2006). *Exploring the efficacy of pre-equating a large scale criterion-referenced assessment with respect to measurement equivalence.* Unpublished doctorate thesis, Georgia State University, The College of Education, Atlanta, GA.

Dongyang, L. (2009). *Developing a common scale for testlet model parameter estimates under the common-item nonequivalent groups design*. Unpublished doctorate thesis, University of Maryland.

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

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

Han, K. T. (2007). *WinGen2: Windows software that generates IRT parameters and item responses [computer program]*. Amherst, MA: Center for Educational Assessment, University of Massachusetts Amherst.

Han, K. T. (2008). *Impact of item parameter drift on test equating and proficiency estimates.* Unpublished doctorate thesis, University of Massachusetts, Amherst.

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

Harwell, M., Stone, C. A., Hsu, T.-C. & Kirişci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101-125.

He, Y. (2011). *Evaluating equating properties for mixed-format tests*. Unpublished doctoral dissertation, University of Iowa, Iowa City.

Holland, P. W. & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (pp. 187–220). Westport, CT: Praeger Publishers.

Hu, H., Rogers, T. W. & Vukmirovic, Z. (2008). Investigation of IRT-based equating methods in the presence of outlier common items. *Applied Psychological Measurement, 32*(4), 311-333.

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

Kan, A. (2011). Test Eşitleme: OKS testlerinin istatistiksel eşitliğinin sınanması. *Eğitim ve Bilim, 36*(160), 38-51.

Kang, T., & Petersen, N. S. (2009). *Linking item parameters to a base scale.* Paper presented at the National Council on measurement in education, San Diego, CA.

Kaskowitz, G. S., & De Ayala, R. J. (2001). The effect of error in item parameter estimates on the test response function method of linking. *Applied Psychological Measurement, 25*, 39-52.

Keller, L. A., & Keller, R. R. (2008). *A comparison of transformation methods and calibration methods on the classification of students over time.* Paper presented at the meeting of the Psychometric Society, Dover, NH.

Keller, R. R., Kim, W., Nering, M. & Keller, L. A. (2007). *What breaks the equating? A preliminary investigation into threats to a five-year equating chain.* Paper presented at the 2007 AERA Annual Meeting, Chicago, IL.

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

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

Kim, S., & Lee, W. (2006). An extension of four IRT linking methods for mixed-format tests. *Journal of Educational Measurement, 43*(1), 53-76.

Kolen, M. J. (1985). Standart errors of Tucker equating. *Applied Psychological Measurement, 2*, 209-223.

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

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

Lee, G. & Fitzpatrick, A. R. (2008). A new approach to test score equating using ıtem response theory with fixed c-parameters. *Asia Pacific Education Review, 3*, 248–261.

Lehman, R. S., & Bailey, D. E. (1968). *Digital computing: Fortran IV and its applications in behavioral science*. New York: John Wiley.

Li, D. (2009). *Developing a common scale for testlet model parameter estimates under the common- item nonequivalent groups design*. Unpublished doctorate thesis, University of Maryland.

Li, Y. H. & Lissitz, R. W. (2000). An evaluation of multidimensional IRT linking. *Applied Psychological Measurement, 24*, 115-138

Li, Y. H., Tam, H. P. & Tompkins, L. J. (2004). A comparison of using the fixed common-precalibrated parameter method and the matched characteristic curve method for linking multiple-test items. *International Journal of Testing, 4*(3), 267-293.

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

Lord, F.M. (1983). Statistical bias in maximum likelihood estimators of item parameters. *Psychometrika, 48,* 477-482

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

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

Mislevy, R. J. & Stocking, M. L. (1989). A consumer's guide to LOJIST and BILOG. *Applied Psychological Measurement, 13,* 57-75.

Nozawa, Y. (2008). *Comparison of parametric and nonparametric IRT equating methods under the common-item nonequivalent groups design*. Unpublished doctorate thesis, The University of Iowa, Iowa City.

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

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

Oh, S. (2000). *Comparison of traditional and item response theory equating using arm and shoulder girdle muscular strength and endurance tests*. Doctorate thesis, University of Georgia, Athens, Georgia.

Öztürk, N. (2010). *Akademik personel ve lisansüstü eğitimi giriş sınavı puanlarının eşitlenmesi üzerine bir çalışma.* Yüksek Lisans Tezi, Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear scores equating models. In P. W. Holland & D. B. Rubin (Eds.) *Test Equating*, 71-136. New York: Academic Press.

Sinharay, S. & Holland, P. W. (2008). The missing data assumptions of the nonequivalent groups with anchor test (neat) design and their implications for test equating (ETS RR-09-16 Research report). Princeton NJ: Educational Testing Service.

Skaggs, G. & Stevenson, J. (1989). A comparison of pseudo-bayesian and joint maximum likelihood procedures for estimating item parameters in the three-parameter IRT model. *Applied Psychological Measurement, 13*(4), 391-402.

Speron, E. (2009). *A comparison of metric linking procedures in Item Response Theory.* Unpublished doctorate thesis, University of Illinois**,** Chicago, Illinois.

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Swaminathan, H. & Gifford, J. A. (1983). Estimation of parameters in the three- parameter latent trait model. In D. Weiss (Ed.), *New horizons in testing.* New York: Academic Press.

Tsai, T. H. (1997). *Estimating minimum sample sizes in random groups equating*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Way, W. D., & Tang, K. L. (1991, April). *A comparison of four logistic model equating methods*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Wingersky, M. S. & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8,* 347–364.

Wolkowitz, A. A. (2008). *A comparison of classical test theory and item response theory methods for equating number-right scored to formula scored assessments.* Unpublished doctorate thesis, University of Kansas.

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

Yen, W. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52*, 275-291.

Zeng, L. (1991). *Standard errors of linear equating for the single-group design*. ACT Research Report Series. 91–4.

Zhao, Y. (2008). *Approaches for addressing the fit of item response theory models to educational test data.* Unpublished doctorate thesis, University of Massachusett.

EXTENDED ABSTRACT

Large scaled or centralized examinations are generally used for placement to schools or employment of individuals. Large scaled examinations are done periodically since they are applied every year or a few times in a year. As the aims of these exams are the same which are done at different times of the year to maintain the security of questions has been a crucial issue. Therefore many forms of these collateral tests are developed. Asking different questions for these collateral tests may cause candidates to find some forms easy or difficult. These exams which are done in two different periods have similar content however different levels of difficulty may cause mistakes in taking important decisions about the students. In order to avoid this problem and to compare two points of these different forms of collateral tests, and to be able to use these points interchangeably test scores should be equalized. Equating studies which increase the validity of decisions based on test scores are crucial. However equating error influence the accuracy of equating. The accuracy of equating depends on choosing the most appropriate design and method according to the conditions. Hence in this study, test length, sample size, ability distribution and model type are considered and in which situation the equating error can be decreased is determined. In order to control variables in the study simulation data was used. Another reason for using simulation data, equating with common item nonequivalent group design is impossible because in Turkey there is no test implemented with common items. During the data generation in order to provide contribution to the equating process of the large scaled tests in Turkey, sample size, number of the questions, ability distribution features are taken into consideration. To determine the performance of the methods of equating according to the conditions of the tests in Turkey can be a sample for the equating studies of large scaled tests for our country.

In this study by using the generated data according to the identified conditions it is aimed to compare the prediction methods based on item response theory. As a result the methods and conditions which consist the least errors have been planned to be determined and contribution to theoretical studies would be provided.

The study holds the qualification of basic research since it includes the aim of comparing scaling methods. To equate two test forms in this study common-item nonequivalent groups design has been used. Simulation factors used in this study are; numbers of items in the tests (30, 60, 80), the numbers of sample in each group (1000, 3000), ability distribution (similar, negative and positive skewed) and the model type (two and three parameter logistic model) used in data generation. As the equating design in the study horizontal equating based on common items in all factors has been used. Two categories item answers are generated using Wingen3 simulation program.

Item response theory based equating methods are applied to 50 datasets that are generated for each simulation condition. In this study as the calibration method mean-mean (MM), mean-sigma (MS) and Stocking-Lord (SL) was used. PARSCALE 4.1 was utilized for the prediction of item parameters and IRTEQ was utilized for test equating and scaling in separate calibration. In this study since the studies are done with known population parameters, item parameters are predicted with expected a posteriori.

In order to provide more stable predictions, for each condition and method, data generation and equating process were repeated 50 times. The results of the study are evaluated according to the equating error RMSE criterion. Previously the differences between the equating errors of the methods used in the study were investigated. Later, to determine the effects of simulation conditions used in this study ANOVA was conducted for each method. When a model is established with conditions and all interactions the degree of freedom was zero as a result the analysis could not be done. For this reason, with main effects 2, 3 and 4 ways interactions were tested. Since many significance tests were applied to control the error Bonferroni correction was conducted (.002). Additionally to show the effects of the variables on

the performances of the methods, eta squared values were reported as a measure of effect size. Post-hoc tests were done for the values that were found significant in ANOVA.

According to the results of the study the bigger sample sizes and longer tests the more stable item parameters predictions were obtained. In this direction, it can be said that the most stable parameter prediction can be reached if the sample size is over 1000 and the test length are over 60. When the methods which are appropriate for similar ability distribution groups are used error values were found lower. When the methods were investigated according to item parameters, for a and b parameters the least error causing method was MM and the most error causing method was MS for a parameter, SL for b parameter. When the results are evaluated in a general it can be said that the best equalization values can be obtained for 3000 people with 80 items similar ability distribution by using MM method. Besides according to the conditions undertaken in the study with bigger size samples, longer and similar ability distributed tests methods were found to be less erroneous. To generalize the results of this study with other studies is limited with the conditions in this study. Another factor which affects the results of this study is the interaction between the conditions of this study. In evaluating the item parameters ability distribution has the most effects on equating results. In this direction, ability distribution can be considered crucial while choosing a specific equating method. In situations where groups are similar, methods showed better performances compared with dissimilar or unequal groups. As the result of this study it is seen that methods have different results in different conditions. The amount of difference and which method to choose depend on the similarity of ability distribution of groups, sample size, test length and model type. To determine the equating method according to the conditions taken into consideration in this study is significant. Due to the differences in the methods, the results do not show enough stability to prefer one method to the other and it can be said that there is no effective single method for every condition.