Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education) 34: 297-307 [2008]

FARKLI İŞLEYEN MADDELERİN BELİRLENMESİNDE SINIRLANDIRILMIŞ FAKTÖR ÇÖZÜMLEMESİNİN OLABİLİRLİK-ORANI VE MANTEL-HAENSZEL YÖNTEMLERİYLE KARŞILAŞTIRILMASI

COMPARISON OF RESTRICTED-FACTOR ANALYSIS WITH LIKELIHOOD-RATIO AND MANTEL-HAENSZEL METHODS IN DIF ANALYSES

Selda YILDIRIM*

ÖZET: Bu araştırmanın amacı farklı işleyen maddelerin belirlenmesinde, Mantel-Haenszel (M-H) ve Olabilirlik Oranı Analizi (OOA) gibi yöntemlere kıyasla, daha seyrek kullanılan Sınırlandırılmış Faktör Çözümlemesi (SFÇ) yöntemini M-H ve OOA ile karşılaştırmaktır. Söz konusu çalışmada PISA 2003 matematik maddeleri kullanılmış, İngilizce ve Türkçe formları arasında yanlı çalışma potansiyeli olan sorular tespit edilmiştir. Gerçek veri kullanılan bu analizlere ek olarak, elde edilen bulguların kontrol edilmesine ve açıklanmasına yönelik bir simülasyon çalışması da yapılmıştır. Sonuçta SFÇ yönteminin karşılaştırılan grup ortalamaları farklı ya da eşit olduğu durumlarda M-H ve OOA yöntemlerine göre, faklı işleyen maddeleri belirleme oranına göre daha doğru sonuçlar verdiği görülmüştür.

Anahtar sözcükler: sınırlandırılmış faktör çözümlemesi, farklı işleyen madde, mantel-haenszel, olabilirlik- oranı analizi

ABSTRACT: The purpose of this study is to compare the Restricted Factor Analysis (RFA) results to that of Mantel Haenszel (M-H) and Likelihood Ratio (IRT-LR) analyses. The data of PISA 2003 mathematics items from the English and Turkish versions were used in the study. Results of these empirical analyses were controlled and investigated further through the use of a simulation study. This study revealed that, in both occasions where the group means are equal or different, RFA produced more accurate results than M-H and IRT-LR.

Keywords: restricted factor analysis, differential ıtem functioning, mantel-haenszel, likelihood- ratio analysis

1. GİRİŞ

Bir ölçme aracındaki bir veya daha fazla maddenin (sorunun) doğru cevaplanma ihtimalinin, ölçülmek istenen yeterliğin dışında, maddeyi cevaplayan kişinin cinsiyeti gibi, bazı faktörlerden de etkilenmesi sonuçların geçerli ve adil olmasını engelleyebilmektedir (Beaton, 1998; Camili ve Congdon 1999). Alan yazınında cinsiyet, sosyo-ekonomik statü, dil, yaşanılan bölge gibi bazı faktörlerin yukarıda sözü edilen türden etkileri tespit edilmiştir (Zenisky, Hambleton ve Robin, 2003; Engelhard,1990; Harris ve Carlton 1993; Sireci ve Berberoğlu, 2000; Yurdugül ve Aşkar 2004a-b; Ercikan 2002; Çet, Yıldırım ve Berberoglu; 2006). Bu etkiler araştırılırken, temel olarak, ülke, cinsiyet gibi farklı gruplarda bulunan fakat ölçme aracının sonucuna göre aynı yeterlikte olan kişilerin araştırılan maddeyi doğru cevaplama ihtimalleri veya başka bir deyişle performansları karşılaştırılır. Eğer bu performanslar farklıysa söz konusu madde, ilgili gruplarda "farklı işleyen madde" (differential item functioning) ifadesiyle betimlenir.

Diğer yandan ülkelerin TIMSS (Third International Mathematics and Science Study), PISA (Programme for International Student Assessment) gibi uygulamalarla öğrenci performanslarını karşılaştırmak istemesi, farklı işleyen maddelerin tespitini çok daha önemli hale getirmiştir. Bu uygulamalarda ölçme aracı katılımcı ülkelerin dillerine çevrilmekte ve bunların denk oldukları varsayılmaktadır. Ancak alan yazınındaki bazı çalışmalar bu çevrilmiş ölçme araçlarının farklı işleyen maddeler içerebildiğini göstermektedir (Arim ve Ercikan, 2005).

Farklı işleyen maddelerin belirlenmesine yönelik analizler bir geçerlik çalışması olarak ortaya çıkmış olmakla birlikte, maddelerin farklı işleme sebeplerini tespit edebilecek bazı nitel çalışmaları da kapsayacak şekilde genişlemiştir (Ercikan, 2002; Gierl ve Khaliq, 2001). Böylece elde edilen bilgiler, daha geçerli ve adil ölçme araçlarının geliştirilmesinde kullanıldığı gibi, ilgili grupların tipik

* Yrd.Doç.Dr., Abant İzzet Baysal Üniversitesi, e-posta: cet_s@ibu.edu.tr

özelliklerini ortaya çıkaracak ipuçları da verebilmektedir (Hui ve Triandis, 1989). Bu bağlamda, farklı işleyen maddelerin ve bunların farklı işleme sebeplerinin belirlenmesine yönelik analizlerin, Türkiye'de yapılan ölçme ve değerlendirme çalışmalarının olağan bir parçası haline gelmesi birçok açıdan faydalı olacaktır. Örneğin ülkemizin çeşitli bölgelerindeki öğrencilerin farklılıklarının bilinmesi, ilgili öğrencilerin ihtiyaçlarına daha uygun müfredat oluşturulması, ders kitapları yazılması veya bazı araç ve gereçlerin hazırlanmasını sağlayabilecektir. Diğer yandan, Türkiye'nin de katıldığı TIMSS, PISA gibi uluslararası çalışmalarda kullanılan soruların Türkçeleştirilmesi ve daha adil olacak şekilde uyarlanması, farklı işlediği görülen soruların değerlendirmeye alınmayarak daha sağlıklı kararlara varılması gibi alanlarda da bu analizlerin sonuçlarından yararlanılabilecektir.

Ancak, bu çalışmalara başlarken göz önüne alınması gereken ilk nokta, farklı işleyen maddelerin tespitinde kullanılan çeşitli analiz yöntemlerinin, sayıltı ve algoritmalarının aynı olmamasından ötürü, farklı hata kaynakları içermekte olduğudur (Camili ve Shepard, 1994). Dolayısıyla sonuçlarına güvenilebilecek bir yöntemle analizlere başlamak önemlidir. Farklı işleyen maddelerin belirlenmesinde kullanılan yöntemler; 1) Klasik Test Kuramına dayalı yöntemler 2) Madde Tepki Kuramına dayalı yöntemler, 3) Kay-kare yöntemleri ve 4) Faktör çözümlemelerine dayalı yöntemler olmak üzere dört ana gruba ayrılmaktadır (Benito ve Ara 2000; Sireci ve Allalouf, 2003). Bu gruplarda yer alan yöntemlerin en tipikleri, sırasıyla, Delta Metodu (Angoff ve Ford, 1973), Olabilirlik-Oranı Analizi (OOA) (Thissen, Steinberg ve Wainer, 1988), Mantel-Haenszel yöntemi (M-H) (Hambleton ve Rogers, 1989) ve Sınırlandırılmış Faktör Çözümlemesi (SFÇ) (Oort, 1992) olarak belirtilebilir.

Alan yazınında yöntem karşılaştırması yapan birçok çalışma bulunmaktadır (Gao ve Wang 2005; Rogers ve Swaminathan, 1993; Gierl, Jodoin ve Ackerman, 2000; Benito ve Ara, 2000). Örneğin Gierl, Jodoin ve Ackerman (2000) bir testteki yanlı çalışan madde sayısı çok olsa dahi, her ikisi de bir kay-kare yöntemi olan Mantel-Haenszel (M-H) ve Lojistik Regresyon (LR) yöntemlerinin güvenilir sonuçlar verdiğini göstermişlerdir. Benito ve Ara (2000) Lojistik Regresyon (LR), Sınırlandırılmış Faktör Çözümlemesi (SFÇ), madde tepki kuramına dayalı bir yöntem olan İşaretli Alan İndeksi ve Mantel-Haenszel (M-H) yöntemlerini karşılaştırmış ve bu çalışmalarının sonucunda; M-H yönteminin farklı işleyen maddeleri tespit etmede daha güçlü bir yöntem olduğunu, bu yöntemler içinde görece az kullanılan SFÇ yönteminin de güvenilir sonuçlar verdiğini ve madde tepki kuramına dayalı olmayan yöntemlerin, madde tepki kuramına dayalı yöntemlere göre daha güçlü olduğunu belirtmişlerdir.

Bu çalışma Türkçeleştirilmiş maddelerden oluşan bir testte farklı işleyen maddelerin tespitinde yukarıda sözü edilen SFÇ yöntemi ile OOA, M-H yöntemlerini karşılaştırmak ve hangilerinin daha güvenilir sonuçlar verdiğini belirlemek amacıyla yapılmıştır. SFÇ bazı çalışmalarda oldukça güvenilir sonuçlar veren bir yöntem olarak belirtilmekle beraber diğer analizlere kıyasla görece az kullanılan bir yöntem olması nedeniyle seçilmiş ve diğer iki farklı gruba ait yöntemler olan M-H ve OOA ile uyumunun incelenmesi amaçlanmıştır. İki yöntem arasındaki uyum, bir maddenin her iki yöntem tarafından da, farklı işleyen ya da farklı işlemeyen olmak üzere aynı şekilde tespit edilmesi olarak tanımlanmaktadır (Camilli ve Shepard 1994). Çalışmada PISA 2003 uygulamasında yer alan Türkçe ve İngilizce matematik maddeleri kullanılmıştır.

Diğer yandan, gerçek bir uygulama verisine dayalı karşılaştırmadan elde edilecek bilgi, maddelerin gerçekte farklı işleyip işlemedikleri kesin olarak bilinemeyeceğinden, sınırlı kalmaktadır. Bu eksikliğin giderilmesi amacıyla, belirlenmiş bazı maddelerin farklı işleyeceği şekilde üretilmiş bir simülasyon verisi ile, gerçek veri kullanılarak yapılan çalışmadan elde edilen bulgular sınanmıştır. Böylece hem ampirik hem de simülasyon çalışmalarından elde edilen sonuçlar birleştirilerek daha kapsamlı bir karşılaştırma yapılmıştır.

2. YÖNTEM

2.1. Ölçme Aracı ve Örneklem

PISA, OECD tarafından öğrencilerin bilgi toplumunun gereklerine uygun yetişip yetişmediklerini değerlendiren ve üç yılda bir tekrarlanan bir çalışmadır. Çalışma 15 yaşındaki öğrencilere uygulanır. 2003 yılında yapılan PISA uygulamasının ana teması matematik okuryazarlığı olmuştur. Uygulamada 13 kitapçık kullanılmıştır. Bu araştırmada, bu kitapçıklar içinden matematik sorularının görece en fazla olduğu 2. kitapçık seçilmiştir. Bu kitapçığı cevaplayan 425 Amerikalı ve 319 Türk öğrenciden elde edilen veriler analiz edilmiştir.

2.2. İstatistiksel Analizler

2.2.1. Tek boyutluluk ve Yapısal Eşitlik

PISA 2003 uygulamasında 2. kitapçıktaki matematik yanıtları bazı sorularda 0 (yanlış), 1 (doğru) olarak kodlanırken bazılarında 0 (yanlış), 1 (kısmen doğru) ve 2 (tam doğru) gibi farklı şekillerde kodlanmıştır. Kitapçıktaki tüm sorularda istatistiksel analiz sonuçlarının aynı şekilde yorumlanabilmesi için, analizlerden önce, kısmen doğru ve tam doğru olarak değerlendirilen yanıtlar 1 bunun dışındaki her türlü yanıt ise 0 olarak yeniden kodlanmıştır.

Farklı işleyen madde analizinde kullanılan istatistiksel yöntemlerin tek boyutluluk ve yapısal eşitlik olmak üzere iki önkoşulu vardır. Tek boyutluluk farklı gruplardaki öğrencilere uygulanan ölçme araçlarının tek bir yapı ölçtüğünü, yapısal eşitlik ise ölçülen bu yapıların eş olmasını gerektirmektedir. Bu amaçla önce Amerikan ve Türk öğrenci verileri birleştirilmiş ve temel bileşenler faktör çözümlemesinde tek faktör altında toplanan sorular seçilmiştir (Ackerman, 1992). Faktör çözümlemesi SPSS 13 programı kullanılarak incelenmiş ve varimax yöntemi ile döndürülmüş eksenlerden elde edilen sonuçlar kullanılmıştır.

Faktör çözümlemeleri yöntemi ile seçilen bu soruların her iki grupta ayrı ayrı tek boyutlu bir modele uyup uymadığı doğrulayıcı faktör çözümlemeleri yöntemi ile ayrıca kontrol edilmiştir. Bu aşamada PRELIS 2 ve LISREL 8 programları kullanılmıştır (Jöreskog ve Sörbom, 2001; Jöreskog ve Sörbom, 2002).

2.2.2. Sınırlandırılmış Faktör Çözümlemesi (SFÇ)

SFÇ yöntemi ile analiz PRELIS 2 ve LISREL 8 programları kullanılarak yapılmıştır. SFÇ yönteminde kullanılan model,

$$\mathbf{X_i} = \tau_i + \Lambda_i \xi + \Delta_i \Psi + \delta_i$$

şeklinde yazılabilir. Bu modelde $\mathbf{X_i}$, herhangi bir kişinin i. maddeki puanı, ξ ve Ψ cinsinden ifade edilmiştir. Λ_i ve Δ_i, sırasıyla testin ölçtüğü yetenek (ξ) ve grup değişkeninin (Ψ) katsayılarıdır. Ayrıca δ_i modeldeki hata ve τ_i ise sabit terimdir. Bu modele göre ifade edecek olursak; i. madde üzerinde Ψ değişkeninin bir etkisi varsa, yani $\Delta_i \neq 0$ ise, i. madde Ψ değişkenine göre farklı işleyen bir maddedir (Oort, 1992).

2.2.3. Mantel-Haenszel Yöntemi (M-H)

Bu yöntemde önce, aynı puan kategorisinde fakat farklı gruplarda olan kişilerin incelenen maddeyi doğru cevaplama oranları, M-H kay-kare istatistiği kullanılarak yapılmıştır (Hambleton ve Rogers, 1989).

Mantel-Haenszel yönteminde karşılaştırılacak puan kategorileri alan yazınında önerildiği gibi örneklem büyüklüğü dikkate alınarak, yüzde toplam puanı üzerinden hesaplanmış ve bu amaçla toplam puanlar % 20'lik dilimlerde her iki ülke verisi birlikte kullanılarak belirlenmiştir (Donoghue ve Allen, 1993). Hesaplanan puan kategorileri M-H yöntemi için Waller (2005) tarafından yazılan EZDIF programında kullanılmıştır.

2.2.4. Olabilirlik Oranı Analizi (OOA)

OOA analizinde farklı gruplardan elde edilen madde karakteristik eğrilerinin aynı olmaması ilgili maddenin yanlı çalıştığına kanıt sayılır. Bu yöntem ile madde karakteristik eğrilerinin farklı olup olmadığı bu eğrileri belirleyen "a" (ayırdedicilik), "b" (madde güçlüğü) ve "c" (tahmin) parametrelerinin bir veya birkaçının farklılığı incelenerek araştırılmıştır. Bu amaçla, ilgili sorunun tüm parametrelerinin her iki grupta eşit olmasıyla sınırlı model, bir veya birkaç parametrenin farklı olabileceği serbest modelle karşılaştırılmıştır. Bu karşılaştırma her iki modelin olabilirlik oranları arasındaki farkın kay-kare dağılımına göre anlamlı bir büyüklük olup olmadığı incelenerek yapılmıştır (Thissen, et.all, 1988). Bu analiz, Thissen (2001) tarafından yazılan IRTLRDIF v.2.0b programı kullanılarak yapılmıştır.

Farklı işleyen madde analizlerinde her madde için ayrı ayrı analiz yapılmasından kaynaklanabilecek birinci tip hatayı kontrol etmek için, 0.05 anlamlılık seviyesinde Benjamini-Hochberg (1995) yöntemi kullanılmıştır (Williams, Jones ve Tukey, 1999).

2.2.5. Simülasyon Çalışması

Simülasyon çalışmalarında, yanlı çalışması ve güçlüğü önceden belirlenen farazi maddeler için muhtemel öğrenci cevaplarından oluşan veri üretilir. Böylece farklı metotların önceden belirlenen madde özelliklerini tespit etme dereceleri incelenebilir. Bu çalışmada üretilen veriler öğrenci populasyonundaki puan dağılımının normal olduğu varsayımına dayanmaktadır. Simülasyon çalışmasında, ampirik çalışmadaki toplam soru sayısı, farklı işleyen soru sayısı, örneklem büyüklükleri, grup ortalamaları ve standart sapmaları gibi değerlere denk koşullarda veri üretilmiştir. Veri üretilirken Deng ve Lin (2000) tarafından önerilen algoritma kullanılmıştır.

Farklı işleyen maddeler, madde parametreleri arasındaki fark 0.75 olacak şekilde belirlenmiştir. Bu yaklaşım, farklı işleyen madde oluştururken, alan yazınındaki araştırmalarda da kullanılmıştır (Kim ve Cohen, 1992; Benito ve Ara, 2000; Lim ve Drasgow,1990).

3. BULGULAR

3.1. Betimsel İstatistikler

Her iki ülkenin birleştirilmiş verisi üzerinde yapılan temel bileşenler faktör çözümlemesinde bir faktör altında toplanan 22 madde, yapılacak diğer analizler için seçilmiştir. Seçilen maddelere ait betimsel istatistikler Tablo 1'de verilmektedir. Tablo 1'de Amerikan öğrencilerinin, seçilen 22 sorudaki ortalamalarının Türk öğrencilerin ortalamasından yarım standart sapma fazla olduğu ve standart sapmalarının ise yaklaşık aynı değerlerde olduğu görülmektedir.

Tablo 1: Tek Boyut Altında Seçilen Maddeler

	PISA 2003	
	Türkiye	**Amerika**
Madde sayısı	**22**	**22**
Öğrenci sayısı	**391**	**425**
Ortalama	**8.39**	**11.02**
Standart S.	**5.35**	**5.52**
Çarpıklık	**0.414**	**-0.043**
Basıklık	**-0.679**	**-1.016**

Cronbach alpha değerleri her iki ülke için 0.88 olarak bulunmuştur.

3.2. Tek Boyutluluk

Seçilen soruların her iki ülkede ayrı ayrı tek boyutluluk sağlayıp sağlamadığı doğrulayıcı faktör analizi ile incelenmiştir. Bu analizden elde edilen indisler Tablo 2'de verilmiştir.

Tablo 2: Seçilen Maddelerde Uyum İyiliği İndisleri

	PISA 2003	
	Türkiye	**Amerika**
NC	**2.7**	**1.85**
GFI	**0.96**	**0.97**
AGFI	**0.95**	**0.96**
RMSEA	**0.067**	**0.045**
NNFI	**0.92**	**0.95**
CFI	**0.93**	**0.95**

Bu değerlere bakıldığında 5 değerinden küçük bir NC (Normed Chi-Square), 0.10 değerinden küçük olan RMSEA (Root-Mean-Squared Error of Approximation), 0 ile 1 arasındaki AGFI (Adjusted-Goodness-of-Fit), NNFI (Non-Normed-Fit Index) ve CFI (Comparative-Fit-Index) indislerinin model veri uyumuna kanıt olabileceği görülmektedir (Kelloway, 1998).

3.3. Farklı İşleyen Maddelerin Analizi

Yapılan analizler sonucunda M-H yöntemi 13, OOA yöntemi 15 maddeyi farklı işleyen madde olarak tespit ederken SFÇ yöntemi 7 maddeyi farklı işleyen madde olarak tespit etmiştir. Bu 7 madde her üç yöntem tarafından da farklı işleyen madde olarak tespit edilmiştir.

Farklı işleyen madde analizi yöntemleri arasındaki uyum Tablo 3a'da özetlenmiştir. Karşılaştırılan her iki yöntemce aynı özellikleri tespit edilen madde sayısının tüm madde sayısına olan oranına bakılarak yapılan karşılaştırmada, M-H ile OOA yöntemlerindeki uyumun % 82, bu yöntemler ayrı ayrı SFÇ ile karşılaştırıldığında ise, M-H ile SFÇ yöntemlerindeki uyumun % 72, OOA ile SFÇ yöntemlerindeki uyumun ise % 64 olduğu görülmektedir.

Tablo 3a: Farklı İşleyen Maddelerin Analizinde Yöntemlerin Uyumu

Yöntem	**M-H**	**OOA**	**SFÇ**
M-H	1		
OOA	% 82	1	
SFÇ	% 72	% 64	1

3.4. Simülasyon Çalışması

Gerçek verideki ortalamalar ve standart sapmalar arasındaki ilişkiye uygun olarak, simülasyon çalışmasında gruplardan birinin ortalaması 1 ve diğerinin ortalaması 0.5 olacak şekilde veri üretilmiştir. Her iki grubun standart sapması 1 olarak alınmıştır. 429 Amerikan ve 319 Türk öğrencinin 22 soru cevapladıkları varsayımıyla üretilen veride, 8 soru farklı işleyecek şekilde sabitlenmiştir. Diğer yandan grup ortalamaları arasındaki farklılığın farklı işleyen madde analizi sonuçlarını etkileyen bir faktör olduğu alan yazınında belirtilmektedir (Hidalgo ve Pina, 2004). Bu nedenle ortalamaların aynı olduğu varsayımı ile simülasyon çalışması yinelenerek, her iki durumdaki sonuçlarda farklılık olup olmadığı gözlenmiştir. Simülasyon çalışmasında elde edilen sonuçlar aşağıda özetlenmiştir.

Tablo 3b, karşılaştırılan grupların ortalamaları aynı ve farklı olduğu durumlarda yöntemler arasındaki uyum yüzdelerini göstermektedir. Her iki durumda da OOA ve M-H yöntemlerinin birbiriyle uyumunun, SFÇ yönteminin OOA ve M-H yöntemleri ile uyumundan fazla olduğu görülmektedir. Ortalamalar farklı olduğunda SFÇ yönteminin OOA ve M-H ile uyumu daha da azalmaktadır.

Tablo 3b: Simülasyon Çalışmasında Yöntemler Arasındaki Uyum Yüzdeleri

Yöntem	Uyum (yüzde)	
	Farklı ortalama	Aynı ortalama
OOA ve M-H	% 95	% 91
M-H ve SFÇ	% 59	% 82
OOA ve SFÇ	% 64	% 77

Tablo 3c grupların ortalamalarının farklı, Tablo 3d ise grupların ortalamalarının aynı olduğu durumlarda, farklı işleyen ya da işlemeyen olarak tespit edilen madde sayılarını göstermektedir. Parantez içindeki sayılar doğru tespit edilen madde sayılarıdır.

Tablo 3c: Ortalamalar Farklı İken Yöntemlerin Tespit Ettiği Madde Sayıları

Yöntem	Madde sayısı					
	Farklı işleyen		Farklı işlemeyen		Toplam	
OOA	4 (2)	% 50	18 (12)	% 67	22 (14)	% 64
M-H	5 (3)	% 60	17 (12)	% 71	22 (15)	% 68
SFÇ	10 (8)	% 80	12 (12)	% 100	22 (20)	% 91

Tablo 3d: Ortalamalar Aynı İken Yöntemlerin Tespit Ettiği Madde Sayıları

Yöntem	Madde sayısı					
	Farklı işleyen	(%)	Farklı işlemeyen	(%)	Toplam	(%)
OOA	11 (6)	% 55	11 (9)	% 81	22 (15)	% 68
M-H	9 (6)	% 67	13 (11)	% 85	22 (17)	% 77
SFÇ	6 (6)	% 100	16 (14)	% 88	22 (20)	% 91

Görüldüğü üzere her iki durumda da SFÇ yöntemi, M-H ve OOA yöntemlerine göre daha fazla maddeyi (20 madde) doğru tespit etmiştir. Özellikle grup ortalamalarının farklı olduğu durum dikkat çekicidir. Bu durumda M-H ve OOA yöntemleri farklı işleyen maddelerin çoğunluğunu yakalayamamıştır. Ortalamaların aynı olduğu durumda ise M-H ve OOA yöntemleri aslında farklı işlemeyen birçok maddeyi farklı işleyen olarak tespit etmiştir.

4. YORUM / TARTIŞMA

Bu çalışmanın sonuçları PISA gibi çeşitli dillere uyarlanan uluslararası uygulamalarda kullanılan testlerde farklı işleyen maddeler olabileceğini göstermiştir. Gerçek veri ile yapılan analizlerde her üç yöntemin de farklı işleyen madde olarak bulduğu maddeler dikkate alınırsa incelenen maddelerin % 32'sinde öğrenci performanslarını testin ölçtüğü yeterlik dışında etkileyen unsurların olabileceği görülmüştür.

Gerçek veri kullanılarak yapılan karşılaştırmada, M-H ile OOA yöntemlerinin uyumlu oldukları görülmüştür. Bu sonuç, alan yazınındaki çalışmaların sonuçları ile benzerlik göstermektedir (Thissen ve

diğerleri, 1988). Ayrıca SFÇ ile tespit edilen tüm maddeler M-H ve OOA yöntemleriyle de farklı işleyen madde olarak tespit edilmiş, ancak M-H ve OOA yöntemlerinin SFÇ yönteminin tespit ettiği maddelerin dışında bazı maddeleri de farklı işleyen madde olarak tespit ettiği görülmüştür. Bu durumda SFÇ yönteminin farklı işleyen bazı maddeleri tespit edemediği veya M-H ve OOA yöntemlerinin farklı işleyen olarak bazı maddeleri yanlış tespit ettiği söylenebilir. Ancak gerçek veride analizlerden önce, hangi maddelerin farklı işleyen madde olduğu kesinlikle bilinemeyeceği için, bu iki durumdan hangisinin olduğu söylenememektedir. Bu eksikliği gidermek için yapılan simülasyon çalışmasından elde edilen bulgular hangi yöntemlerin maddeleri daha doğru tespit ettiği ile ilgili bilgiler vermiştir.

Simülasyon çalışmasında, grupların ortalamalarının farklı ve aynı olduğu her iki durumda da, SFÇ ile M-H ve OOA yöntemleri arasındaki uyum, M-H ve OOA arasındaki uyuma göre daha az olmuştur. Ayrıca ortalama farklı olduğunda SFÇ yönteminin M-H ve OOA yöntemleri ile uyumunun daha da azaldığı görülmüştür. Bu sonuç, gerçek veri ile yapılan karşılaştırmada gözlenen sonuçla benzerdir. Ancak gerçek veriden elde edilen sonucun tersine, M-H ve OOA yöntemleri ortalamaların farklı olduğu durumda, SFÇ yönteminden daha az maddeyi farklı işleyen madde olarak göstermiştir. Bu durum, farklı işleyen madde analizlerinin sonucunu, gerçek veride, ortalama farklılığının dışında, başka faktörlerin de etkileyebileceğini göstermektedir.

Bu çalışmanın en dikkat çekici sonuçlarından biri de, SFÇ yönteminin maddelerin nasıl işlediği ile ilgili M-H ve OOA yöntemlerine göre daha doğru bilgi vermesi olmuştur. Özellikle grup ortalamalarının farklı olduğu durumda, SFÇ farklı işleyen maddelerin hepsini tespit ederken, M-H ve OOA yöntemleri bu maddeleri tespit etmede yetersiz kalmıştır. SFÇ yönteminin maddeleri doğru tespit etme oranı % 91 iken, M-H ve OOA yöntemlerinin maddeleri doğru tespit etme oranı sırasıyla % 68 ve % 64 olmuştur. Ortalamaların aynı olduğu durumda ise M-H ve OOA yöntemleri farklı işlemeyen birçok maddeyi farklı işleyen madde olarak göstermiştir. Bu durumda ise SFÇ yönteminin maddeleri doğru tespit etme oranı yine % 91 iken, M-H ve OOA yöntemlerinin maddeleri doğru tespit etme oranı sırasıyla % 77 ve % 68 olmuştur. SFÇ yönteminin gruplar arasındaki ortalama farklılığından diğer yöntemler kadar etkilenmemesi, bu yöntemin kullandığı algoritmada, karşılaştırılan gruplardaki kişilerin ortalamasını da bir değişken olarak analize dahil etmesi olabilir (Oort, 1992).

Alan yazınındaki pek çok çalışma (Hambleton ve Rogers, 1989; Benito ve Ara, 2000) M-H yönteminin çok iyi sonuçlar verdiğini belirtmekle birlikte, bazı çalışmalar ortalama veya standart sapma farklı olduğunda M-H yönteminin sonuçlarının olumsuz etkilendiğini göstermektedir (Narayanan ve Swaminathan, 1996). Benzer şekilde bu çalışmadaki simülasyon bulguları da, karşılaştırılan grupların ortalamalarının farklı olmasının M-H yönteminin sonuçlarını etkileyeceği ve M-H yönteminin farklı çalışan maddeleri tespit etmede yeterli olamayacağı görüşünü desteklemektedir.

Aynı durum OOA yönteminin sonucunda da görülmektedir. M-H gibi OOA yönteminin de simülasyon çalışmasında karşılaştırılan gruplar arasındaki ortalamalar farklı olduğunda, farklı işlediği halde birçok maddeyi farklı işlemeyen madde olarak tespit ettiği görülmüştür. OOA yöntemindeki bu sonuç gruplardaki ortalama farklılığından veya gruplardaki kişi sayısının yeterince fazla olmamasından kaynaklanabilir. Alan yazınında madde tepki kuramı analizlerinin küçük örneklem gruplarından olumsuz etkilendiği ve farklı işleyen maddelerin, özellikle maddeler arasında küçük farklar varsa, büyük örneklem gruplarında daha iyi tespit edileceği belirtilmektedir (Lim ve Drasgow, 1990; Thissen, et.all, 1988).

5. SONUÇ VE ÖNERİLER

Alan yazınında, farklı işleyen maddelerin belirlenmesinde farklı analiz yöntemlerinin farklı algoritmalar kullandığı, bu nedenle farklı hata kaynaklarının olabileceği ve bunu kontrol etmek için birden fazla yöntemin kullanılması ve bu yöntemlerin ortak olarak tespit ettiği maddelerin farklı işleyen madde olarak değerlendirilmesi önerilmektedir (Hambleton, Clauser, Mazor ve Jones, 1993; Hidalgo ve Pina, 2004). Ancak farklı işleyen maddeleri belirleyen sonuçlarına güvenilir bir analiz için birden fazla yöntemin kulanılması zor olmakta ve zaman almaktadır. Bu durum test geliştirme aşamasında farklı işleyen maddelerin belirlenmesinin rutin bir işlem olmasına engel olmaktadır. Bu çalışmadaki bulgular

SFÇ yönteminin bu duruma çözüm olarak kullanılabileceğini göstermektedir. Bulgulara göre, SFÇ yöntemi M-H ve OOA ile kıyaslandığında, tek başına kullanıldığı zaman, sonuçlarına daha fazla güvenilebilecek bir yöntemdir.

SFÇ yönteminin diğer yöntemlerden farklı bir başka özelliği de, bir maddenin farklı işlemesine sebep olabilecek birden fazla değişkenin aynı analizde yer alabilmesidir. Bu özelliğinden dolayı da farklı işleyen madde analizleri için oldukça kullanışlı bir yöntemdir (Oort, 1992).

Alan yazınındaki çalışmalar, testteki madde sayısı, maddelerin zor ya da kolay olması, farklı işleyen madde sayısı, farklı işleyen maddelerin tek biçimli olması ya da olmaması, farklı işleyen maddenin derecesi, karşılaştırma yapılan gruplardaki kişi sayısı ve bu grupların ortalamalarının aynı ya da farklı olması vb. faktörlerin farklı işleyen maddelerin belirlenmesinde sonuçları etkilediğini ortaya koymaktadır (Hidalgo ve Pina, 2004; Narayanan ve Swaminathan, 1996; Gao ve Wang, 2005). Bu nedenle gerçek veri ile yapılan çalışmalarda, farklı işleyen maddelerin analizinde kullanılan yöntemler her zaman güvenilir sonuçlar vermeyebilir (Lim ve Drasgow, 1990; Thissen ve diğerleri, 1988; Benito ve Ara, 2000). Alan yazınında bu faktörlerin dikkate alındığı, SFÇ yöntemi ile ilgili araştırmalar azdır. Bu çalışmada SFÇ yönteminin ortalamalar arasındaki farklılık faktöründen M-H ve OOA kadar etkilenmediği görülmüştür. SFÇ yönteminin sonuçlarını bu faktörlerin ne kadar etkilediği ile ilgili çalışmaların yapılması, bu yöntemin tek başına kullanıldığında, güvenilirliği ile ilgili daha fazla bilgi verecektir.

KAYNAKLAR

Ackerman, T.A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91.

Allalouf, A., Hambleton, R.K. ve Sireci, S.G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36*(3), 185 – 198.

Angoff, W.H. ve Ford, S.F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10*, 95-105.

Arim, R.G. ve Ercikan, K. (2005 April). *Comparability Between the US and Turkish Versions of the Third International Mathematics and Science Study's Mathematics Test Results.* Paper presented at National Council on Measurement in Evaluation, Montreal, Canada.

Beaton, A.E. (1998). Comparing cross-national student performance on TIMSS using different test items. *International Journal of Educational Research, 29*, 529-542.

Benito J.G. ve Ara M.J.N. (2000). A comparison of X^2, RFA and IRT based procedures in the detection of DIF. *Quality ve Quantity, 34*, 17-31.

Benjamini, Y. ve Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society.series B, 57*, 89-300.

Camilli, G. ve Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational and Behavioral Statistics, 24*(4), 323-341.

Camilli, G. ve Shepard, L.A. (1994). *Methods for identifying biased test items.* California: Sage Publications.

Çet, S., Yıldırım, H.H. ve Berberoğlu, G. (2006 June-July). *Differential item functioning (DIF) analysis of PISA 2003 mathematics items across Gender and SES groups.* Paper presented at the Third International Conference on the Teaching of Mathematics, İstanbul, Turkey.

Deng, L.Y. ve Lin, D.K.J. (2000). Random Number Generation for the New Century.*The American Statistician, 54*(2), 145-150.

Donoghue, J.R. ve Allen, N.L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational Statistics, 18*(2), 131 –154.

Engelhard, G. (1990). Gender differences in performance on mathematics items: evidence from the United States and Thailand. *Contemporary Educational Psychology, 15*, 13-26.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing, 2*(3ve4), 199-215.

Gao, L. ve Wang, C. (2005). *Using five procedures to detect DIF with passage-based testlets*. A Paper prepared for the Poster Presentation at the Graduate Student Poster Session at the Annual Meeting of the National Council of Measurement in Education, Montreal, Quebec.

Gierl, M. J., Jodoin, M. ve Ackerman T. (2000). *Performance of Mantel-Haenszel, Simultaneous Item Bias Test, and Logistic Regression when the proportion of DIF items is large.* Paper Presented at the Annual Meeting of the American Educational Research Association, New Orleans, Louisiana, USA.

Gierl, M. ve Khaliq, S. (2001). Identifying soruces of differential item functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement, 38*, 164-187.

Hambleton, R., Clauser, B., Mazor,K. ve Jones, R. (1993). Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment, 9*(1), 1-18.

Hambleton, R.K. ve Rogers, H.J. (1989). Detecting potentially biased test items: comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2*(4), 313-334.

Harris, A. M. ve Carlton, S. T. (1993). Patterns of gender differences on mathematics items on the scholastic aptitude test. *Applied Measurement in Education, 6*, 137-151.

Hidalgo, M.D. ve Pina, S.A.L. (2004). Differential item functioning detection and effect size: a comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement*, *64*(6), 905-915.

Hui, C.H. ve Triandis, H.C. (1989). Effects of culture and responce format on extreme response style. *Journal of Cross-Cultural Psychology, 20*(3), 296-309.

Jöreskog, K. ve Sörbom, D. (2001). *LISREL 8: User's reference guide*. Chicago: Scientific Software International Inc, USA.

Jöreskog, K. ve Sörbom, D. (2002). *PRELIS 2:User's reference guide*. Chicago: Scientific Software International Inc, USA.

Kelloway, E. K. (1998). *Using LISREL for structural equation modeling.* London, New Delhi: Sage Publications.

Kim, S. ve Cohen, A.S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29*(1), 51-66.

Lim, R.G. ve Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology, 75*, 164-174.

Mislevy, R. J. ve Bock, R. D. (1984). *BILOG: Maximum likelihood item analysis and test scoring with logistic models.* Mooresville, IN: Scientific Software.

Narayanan, P. ve Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20*, 257-274.

Oort, F.J. (1992). Using restricted factor analysis to detect item bias. *Methodika, 6*, 150-166.

Rogers, J. ve Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*(2), 105-116.

Sireci, S.G. ve Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing*, *20*(2), 148-166.

Sireci, S.G., Bastari, B. ve Allalouf, A. (1998 August). *Evaluating construct equivalence across adapted tests*. Paper presented at American Psychological Association, San Francisco, CA.

Sireci, S.G. ve Berberoğlu, G. (2000). Using bilingual respondents to evaluate translated-adapted items. *Applied Measurement in Education, 13*(3), 229-248.

Swaminathan H. ve Rogers, J.H. (1990). Detecting differential item functioning using logistic regression Procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Thissen, D. (2001). IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning. Retrieved October 15, 2005, from http://www.unc.edu/~dthissen/dl.html

Thissen, D., Steinberg, L. ve Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer ve H. Braun (Eds.), *Test Validity*. (pp. 147-169). Hillsdale, NJ: Erlbaum.

Waller, N.G. (2005). EZDIF: *A computer program for detecting uniform and nonuniform differential item functioning with the Mantel-Haenszel and logistic regression procedures*. Retrieved April 10, 2005, from http://peabody.vanderbilt.edu /depts/ psych_and_hd/ faculty/wallern/

Williams, V.S.L., Jones, L.V. ve Tukey, J.W. (1999). Controlling error in multiple comparisons with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics, 24*(1),42-69.

Yurdugül, H. ve Aşkar P. (2004a) Ortaöğretim kurumları öğrenci seçme ve yerleştirme sınavının cinsiyete göre madde yanlılığı açısından incelenmesi. *Eğitim Bilimleri ve Uygulama Dergisi, 3*(5), 3-20.

Yurdugül, H. ve Aşkar, P. (2004b) Ortaöğretim kurumları öğrenci seçme ve yerleştirme sınavının öğrencilerin yerleşim yerlerine göre diferansiyel madde fonksiyonu açısından incelenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *27*, 268-275.

Zenisky, A.L., Hambleton, R.K. ve Robin, F. (2003). Detection of DIF in large-scale state sssessments: a study evaluating a two-stage approach. *Educational and Psychological Measurement, 63*(1), 51-64.

EXTENDED ABSTRACT

Educational or psychological tests may include items that operate differently for certain groups. It is important to identify these items because these items may lead to unfair results for groups being compared. The reason for such items to operate differently may be gender, culture or language differences between groups. In the measurement literature the analyses to determine such items are called differential item functioning (DIF) analyses. And the items detected trough these analyses are called items functioning differentially among the groups, or shortly DIF items. DIF analyses should become an essential part of test development studies in Turkey. Both in national assessments, such as student selection and placement test for high school, or international assessments, such as Third International Mathematics and Science Study (TIMSS) or Programme for International Student Assessment (PISA), there is a potential of sources to cause DIF. For example; language, gender or cultural differences may cause some items to show DIF.

While conducting a DIF study, the selection of an appropriate DIF analysis method is a crucial step, because in the litarature there are many methods for investigating the DIF phenomenon. The development of these methods has been a critical area of research and many studies with DIF phenomenon have been investigated through the use of different algorithms and theories which led to different classifications of DIF methods. One of these classifications is as follows: Classical Test Theory (CTT)-based methods, Item Response Theory (IRT)-based methods, Chi-square based methods and Factor Analysis (FA)-based methods. Different methods based on this classification have been investigated in simulation studies. For example, in their simulation study Benito and Ara (2000) have reported that restricted factor analysis (RFA), a FA-based method led to very similar results with other DIF methods. On the other hand it has rarely been used in the studies compare to other DIF methods such as Mantel-Haenszel (M-H), a chi-square method and Likelihood Ratio (IRT-LR), an IRT-based method. For this reason, in this study RFA was selected and the agreement between RFA, M-H and IRT-LR was investigated using both real and simulated data.

The aim of this study is to compare the RFA results to that of M-H and IRT-LR methods and to determine which methods give more reliable results by examining the agreement between these methods. In the first part of the study, the results of DIF analyses were compared through the use of real data. However, the result of this comparison was restricted, because DIF analyses using real data can not guarantee to know exactly whether the items detected were truly DIF items. Therefore, a simulated data in which some items were fixed to show DIF were use to check the results of various mehods. Using real and simulated data provided more extensive comparison in the study. The real data used in the study came from PISA 2003 English and Turkish mathematics items of second booklet which were answered by 425 American and 319 Turkish students respectively.

The results of the investigated DIF methods in real data showed that the agreement of RFA with M-H and IRT-LR were 72 % and 64 % respectively while the agreement of M-H and IRT-LR was 82 %. And the results demonstrated that all the items which were detected as DIF items in RFA were also detected in IRT-LR and M-H analyses. Then, DIF analyses were conducted using data which were simulated according to investigated item number and the mean difference between groups as in real data. In the literature studies state that the difference between means of group is an effective factor for the results of DIF analyses (Hidalgo ve Pina, 2004). Therefore, in real data although the means of American and Turkish students are different, in simulation study, the DIF methods were compared separately in both cases where the means of groups were equal and different. Whether the means of compared groups were equal or different, RFA was best to detect items with DIF that truly function differently across groups, and items with no true DIF.

As a summary, RFA was better than M-H and IRT-LR at least in two dimensions. When the means of the groups were different, M-H and IRT-LR failed in detecting most of the DIF items. On the other hand when group means were equal, all three methods detected most of the DIF items. However M-H and LR misspecified some additional items as showing DIF. As a conclusion, it can be argued that RFA is robust against the groups mean differences compared to M-H and IRT-LR. Among the factors

which may affect the DIF results, are the number of DIF items in the test, difficulty level of items, the degree of DIF in items, the number of individuals in each group, the difference between the group means. This study controlled the difference between the group means in detecting the DIF results. Further studies controlling the other effects may contribute to the literature on DIF studies.