Türk Psikoloji Dergisi
2004, 19 (53), 69 - 82

Gruplararası Karşılaştırmalarda Ölçek Eşdeğerliğinin İncelenmesi: Madde ve Test Fonksiyonlarının Farklılaşması

Oya Somer*
Ege Üniversitesi

Özet

Gerek kültürlerarası karşılaştırmalarda, gerek belirli bir kültür içerisindeki gruplararası karşılaştırmalarda ölçme eşdeğerliğinin sağlanması temel metodolojik problemlerden birisini oluşturmaktadır. Ölçme eşdeğerliğinin incelenmesinde son yıllarda giderek yaygınlaşan madde-cevap kuramına dayalı modellerden yararlanılmaktadır. Madde ve Test Fonksiyonlarının Farklılaşması (Differential Item-Test Functioning, DIF-DTF) genel başlığı altında ele alınan bu modeller, gözlenen test puanları ile bunların altında yatan örtük özellik arasındaki ilişkinin, karşılaştırma grupları açısından eşit olup olmadığının incelenmesine dayalıdır. Ölçülen özellik bakımından aynı düzeyde bulunan, fakat farklı gruplara ait kişilerin maddeyi anahtarlanan yönde cevaplama olasılıkları farklılaştığında, madde fonksiyonlarının farklılaşması ya da maddenin karşılaştırma grupları için yanlı olması söz konusudur. Araştırmamızda 1807 kişilik bir öğrenci örneklemi üzerinde, bir kişilik alt ölçeğinin (Yumuşak Başlılık) 16 maddesinin kız ve erkek öğrenciler için madde fonksiyonlarının farklılıkları incelenmiştir. Yapılan analizler sonucunda 16 maddenin hepsinin iki parametreli modele uyum sağladığı, beş maddenin kız ve erkek öğrenciler arasında farklılık gösterdiği görülmüş, bu maddelerin özellikleri incelenmiş, toplam ölçek bazında nasıl ele alınabileceği tartışılmıştır.

Anahtar kelimeler: *Ölçek eşdeğerliği, Madde Fonksiyonlarının Farklılaşması (DIF), gruplararası karşılaştırmalar*

Abstract

Both in within and cross-cultural settings, measurement equivalence is one of the most important methodological problems in comparisons of group differences. Models based on Item Response Theory are being widely used in recent years in holding measurement equivalence. These models (generally take place under the title of Differential Item-Test Functioning-DIF, DTF) refer to the methods analyzing the relations between observed scores and the latent attribute measured by the test, across comparison groups. The existence of DIF-DTF is evidenced when these relations are different across comparison groups. DIF is defined as differences in the probability of endorsing an item between members of the reference and focal groups having the same latent trait level. In this study, DIF analyses of 16 items of a personality scale (agreeableness) were performed on a student sample (1807 subjects). According to the results, all of the 16 items fitted to the two-parameter logistic model, but 5 items of the agreeableness scale showed differential item functioning between girls and boys. The properties of these DIF items and how to handle them are discussed.

Key words: *Measurement equivalence, Differential Item Functioning (DIF), group comparisons.*

*Yazışma Adresi: Doç. Dr. Oya Somer, Ege Üniversitesi, Edebiyat Fakültesi, Psikoloji Bölümü Bornova, İzmir
E-posta: osomer@edebiyat.ege.edu.tr

Psikoloji alanındaki pek çok çalışma (kültürlerarası ya da kültür içinde) grup karşılaştırmalarını içermektedir. Gruplar arası karşılaştırmalar söz konusu olduğunda ilgilenilen değişken dışında, farklılığa neden olabilecek değişkenlerin kontrol altına alınması araştırma deseninin odak noktalarından birisini oluşturmaktadır. Deneysel desenlerde, ilgili değişkenler gruplara tam tesadüfi atama yoluyla kontrol altına alınmaya çalışılırken, yarı-deneysel desenlerde kontrol ilgili değişkenlerin (yaş, cinsiyet vb.) eşleştirilmesi ile sağlanmaya çalışılmaktadır. Ancak, Waller, Thompson ve Wenk'in (2000) ifade ettikleri gibi literature baktığımızda eşleştirilmesi gereken en önemli değişkenin çoğunlukla gözden kaçırıldığı görülmektedir. Bu, üzerinde karşılaştırma yapılmak istenen değişkendir. Karşılaştırma yapılan değişken açısından grupların eşleştirilmesi, yani ölçme yanlılığının ortadan kaldırılması ölçmede temel problemlerden birisidir. Gruplar, test puanlarının altında yatan örtük özellik (latent trait) üzerinde eşleştirilmeden yapılacak karşılaştırmalardan elde edilecek farklılıkların, ölçme yanlılığından mı, yoksa gerçek grup farklılıklarından mı kaynaklandığını yorumlamak mümkün değildir. Yanlı bir ölçme sonucunda, örtük değişken üzerinde bir farklılık olmadığı halde grupların test puanlarının birbiriyle farklılaşmasının mümkün olduğu gibi, gerçek farklılıkların maskelenmesi de söz konusu olabilmektedir.

Kültürel, organizasyonel, etnik, cinsiyete dayalı ve benzeri grup karşılaştırmalarında öncelikle ölçme eşdeğerliğinin sağlanmasını temel bir gereklilik olarak ortaya çıkmaktadır (Geisinger, 1994; Van de Vijver ve Leung, 2000). Hulin, Drasgow ve Parsons (1983), gözlenen test puanları ile bunların altında yatan örtük özellik arasındaki ilişki, karşılaştırma grupları açısından eşit olduğunda ölçme eşdeğerliğinin sağlandığını ifade etmektedirler. Ölçme eşdeğerliğinin bozulduğuna dair kanıtlar madde-test fonksiyonlarındaki farklılığın (Differential Item Functioning-DIF) incelenmesi ile elde edilebilmektedir. Ölçme literatüründeki "madde yanlılığı" terimi "madde fonksiyon farklılığı" terimi ile büyük ölçüde örtüşse de, DIF daha ziyade maddenin iki ya da daha fazla grup için gösterdiği farklı psikometrik özelliklere işaret ederken, madde yanlılığı DIF analizleri sonucunda çıkarsanan, madde hakkındaki sosyal ve etik değer yargılarını kapsamaktadır (Camilli ve Shepard, 1994; Waller, Thompson ve Wenk, 2000). Madde cevap kuramına dayalı, madde-test fonksiyonlarının farklılaşması (DIF, DTF) analizleri ile, farklı grupların ölçülen yapı ile ilgili özellikleri ve gösterdikleri farklılıklar hakkında ayrıntılı bilgi elde edilebilmektedir. Hambleton, Robin ve Xing (2000), madde cevap kuramı modellerinin günümüzde test geliştirme, değerlendirme ve test veri analizlerine ilişkin uygulamalarda merkezi bir konuma geldiğine dikkat çekmektedirler.

Bu çalışmanın amacı, bir kişilik ölçeği üzerinde örnek bir uygulama yaparak, madde cevap kuramına dayalı madde-test eşdeğerliğinin incelenmesi ve Türkiye'de yapılan psikoloji araştırmalarında bu yeni yöntemlerden yararlanılmasının yaygınlaşmasına katkıda bulunulmasıdır.

Makalede sırasıyla DIF yöntemlerinin temelini oluşturan madde cevap kuramı (IRT) kısaca ele alınacak, DIF kavramı tanıtılacak, örnek bir uygulama yapılarak, DIF analizlerinin yorumları ve ölçme eşdeğerliğinin sağlanması konuları tartışılacaktır.

Madde Cevap Kuramı (Item Response Theory-IRT)

Madde Cevap Kuramı, kişilerin ölçülmekte olan özellik üzerindeki düzeyleri ile, test maddelerinin özellikleri arasındaki ilişkileri inceleyen bir grup modeli içermektedir. Kuram, kişilerin gözlenen tepkilerinin ölçülmek istenen kuramsal yapılara, olasılıksal bir yolla bağlanmasını sağlamaktadır. Madde cevap kuramı modelleri, kişilerin ve maddenin belirli özelliklerine göre, kişinin bir maddeye göstereceği belirli bir tepkinin olasılığını veren matematiksel fonksiyonları içermektedir. Modelde, kişilerin belirli bir test maddesine belirli bir tepkiyi gösterme olasılığının, kişinin test maddelerinin altında yatan örtük özellik (latent trait -q) üzerindeki konumu ile bağlantılı olduğu varsayılmakta ve bu doğrudan gözlenemeyen özellikle, gözlenen tepkiler arasında bağlantı kurulmaktadır.

Madde cevap modellerinin üzerine inşa edildiği temel kavram Madde Karakteristik Eğrisi'dir (Item Characteristic Curve-ICC veya Item Trace Line). Madde karakteristik eğrisi yoluyla, ölçülmekte olan özellik (q) üzerinde birbirinden farklı konumlarda bulunan deneklerin bir maddeye doğru cevap verme olasılıkları elde edilmektedir. Bu olasılığın elde edilmesinde bir, iki ya da üç parametre kullanan farklı madde cevap modelleri bulunmaktadır (farklı modellerin özellikleri ve klasik kuram ile ilişkisi için bkz. Hambleton ve Swaminathan, 1985; Hambleton, Swaminathan ve Rogers, 1991; Lord, 1980; Somer, 1998; 1999).

Bu çalışmada kişilik ve tutum çalışmalarında çoğunlukla tercih edilen iki parametreli lojistik model kullanılmıştır. Reise ve Waller (2003), kişilik ölçekleri üzerinde yaptıkları araştırmaları sonucunda 2 ve 3 parametreli model uyumları arasında anlamlı bir farklılık olmadığını ifade etmişlerdir. İki parametreli modelde güçlük parametresi "b_i", ayırdetme parametresi "a_i" indisleri ile temsil edilmektedir. Modelde, ölçülmekte olan temel yetenek ya da özellik boyutu "θ" ile gösterilmekte ve ortalaması 0, standart sapması 1 olan bir dağılım üzerinde ölçeklenmektedir. Güçlük parametresi (b_i) maddenin örtük özellik üzerindeki konumunu, yerleşimini belirtmektedir. Güçlük parametresinin değeri, *i* maddesini .50 oranında doğru cevaplayan deneklerin bulundukları θ düzeyine karşılık gelmektedir. Tahmin edilen b_i değerleri genellikle -3.0 ile + 3.0 arasında yer almaktadır. Örneğin b_i = 2.0 olduğunda, ortalamanın iki standart sapma üzerindeki deneklerin maddeyi doğru cevaplama olasılıkları .50 olmaktadır ve maddenin güç (ya da popüler olmayan) bir madde olduğu anlaşılmaktadır. Az sayıda kişi tarafından doğru olarak yanıtlanan (kişilik ölçeklerinde, anahtarlanan yönde cevaplanma olasılığı düşük olan) maddeler, nispeten yüksek b_i değerlerine sahip olmakta ve θ üzerinde yüksek uçta yer almaktadırlar. Yani bu maddeleri "doğru" ya da "evet" diye cevaplayabilmeleri için kişilerin daha yüksek örtük özellik düzeyine sahip olmaları gerekmektedir. Ayırdetme parametresi (a_i) ise maddenin, farklı θ düzeylerinde bulunan kişileri birbirinden ayırdedebilme yeterliğini göstermektedir. Maddenin düşük bir a_i değerine sahip olması, farklı yetenek düzeyindeki kişilerin o maddeye doğru cevap verme olasılıklarının farklılaşmadığına işaret etmektedir. Yani maddeye verilen tepkiler, kişileri ilgilenilen özellik açısından farklılaştırmaya katkıda bulunamamaktadır. Ayırdetme parametresinin değerleri yaklaşık .30 ile 2.0 arasında yer almakta ve a_i değeri 1.0 civarında olan maddelerin ayırdetme düzeylerinin iyi olduğu kabul edilmektedir (Hulin, Drasgow ve Parsons, 1983).

Madde cevap modellerinin, parametre tahminleri ve bilgi eğrileri ile maddeler hakkında ayrıntılı bilgi sağlamasının yanı sıra, kişilerin tepki örüntülerinin geçerliğinin incelenmesi (Meijer, 2003), ölçek maddelerinin ölçülen kişilerin yetenek düzeylerine uygun olarak seçilebileceği adaptif testler geliştirilmesi (Weiss, 1983), ölçmenin standart hatasının farklı θ düzeylerinde bulunan kişiler için ayrı ayrı tahminlenebilmesi, gibi birçok yararlı uygulamaları vardır.

Madde cevap modelleri başlangıçta çoğunlukla yetenek ölçeklerinin geliştirilmesi ve özelliklerinin incelenmesinde kullanılmakla birlikte günümüzde ölçek geliştirme problemlerinin çok daha yoğun olduğu kişilik ölçümünde de yaygın olarak kullanılmaya başlanmıştır (Örn., Ellis ve Mead, 2000; Ferrando, 2001; Orlando ve Rand, 2002; Reise, 1999; Smith, 2002; Steinberg 2001; Steinberg ve Thissen, 1995; Waller, Thompson ve Wenk, 2000). Çalışmamızda da, karşılaştırma grupları arasında madde – ölçek eşdeğerliğini incelemek amacıyla örnek olarak bir kişilik ölçümü üzerinde çalışılmıştır.

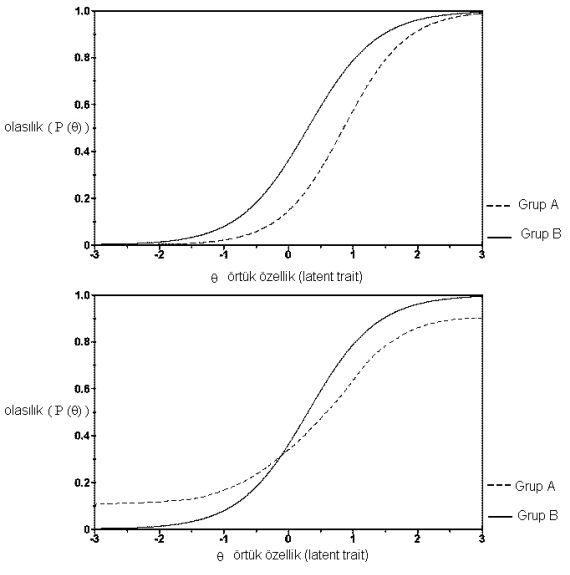
Madde Fonksiyonlarının Farklılaşması (Diffrential Item Functioning - DIF)

Farklı denek gruplarına uygulanan test maddelerinin, ortak bir örtük değişken metriği üzerinde eşleştirilerek karşılaştırılmasında, madde cevap modellerinden yararlanılmaktadır. Madde karakteristik eğrilerinin (ICC), test ile ölçülen özellikte grupların ortak bir metrik üzerinde eşleştirilmesi ile bu eğrile-

rin karşılaştırılması mümkün olmakta ve maddenin farklı gruplar için farklı bir fonksiyona sahip olup olmadığı ya da maddenin gruplar için yanlılığı incelenebilmektedir. Madde düzeyinde ele alındığında, eğer maddenin karakteristik eğrisi iki grup arasında farklılaşma göstermiyorsa ölçek eşdeğerliğinin sağlandığı belirtilmektedir (Camilli ve Shepard, 1994). Madde cevap kuramı terimleri ile tanımlandığında eğer test ile ölçülen özellikte aynı düzeyde olan kişilerin, maddeye doğru cevap verme (kişilik maddeleri için maddeyi anahtarlanan yönde işaretleme) olasılıkları farklılaşıyorsa, madde fonksiyon farklılığından ya da madde yanlılığından söz edilmektedir. Bazı durumlarda, örneğin test maddelerinin bazılarında görülen bu farklılıklar gruplar için hep aynı yönde değilse, toplam test puanlarında bu etkiler birbirini dengeleyebilmekte ve toplam puan karşılaştırmalarında etkileri silinebilmektedir. Bu durum "telafi edici madde-test farklılığı (compensatory DIF-DTF) olarak ifade edilmekte (Raju, 1990; Raju, van der Linden ve Fleer, 1995) ve böyle bir durum olduğunda ölçek eşdeğerliğinin sağlanması için yanlılık gösteren maddelerin testten çıkarılması yerine teste tutulabilmesi de önerilebilmektedir (Roznowski ve Reith, 1999; Smith, 2002). Hangi yaklaşım kullanılırsa kullanılsın, maddelerin DIF özellikleri incelenmeden yapılacak karşılaştırmalar hatalı sonuçlara yol açabilmektedir. Bu incelemeler sonucu DIF gösteren maddeler fark edildiğinde, maddenin neden böyle bir farklılık gösterdiğinin anlaşılmaya çalışılması incelenen yapı hakkında bize önemli bilgiler sağlayabilmekte, yeni maddelerin yazılmasına ışık tutabilmekte, araştırmacının amacına göre maddeleri çıkarmak, revize etmek ya da telafi edici DTF ile ölçülen yapıyı daraltmamak amacıyla maddeleri testte koruma yoluna gitmek mümkün olabilmektedir.

DIF incelemelerinde bir diğer önemli konu, madde karakteristik eğrilerinin gruplar arasındaki farklılıklarının yönü ve şeklidir. Bu farklılıklar, düzgün formlu (uniform) ve düzgün olmayan (nonuniform) şekillerde ortaya çıkabilmektedir (Bkz. Şekil 1). Düzgün formlu DIF'te şekilde görüldüğü gibi, eğrinin a_i parametresi gruplar arasında farklılaşmamakta (yani madde karakteristik eğrisinin şekli gruplar arasında farklılık göstermemekte) fakat b_i parametresi farklılaşmaktadır. Bu durumda maddenin θ ölçeği üzerindeki yeri farklılık göstermektedir. Örneğin maddeyi "evet" olarak cevaplama olasılığı, ölçülen özellik bakımından aynı düzeyde olan kız ve erkeklerden, kızlar için daha kolay olabilmektedir. Düzgün olmayan (nonuniform) DIF'te ise, maddenin hem ayırdetme parametresi, hem de güçlük parametresi gruplar arasında farklılık göstermektedir. Örneğin θ üzerinde belli bir düzeye kadar maddeyi "evet" diye cevaplama olasılığı kızlar için daha kolayken, belirli bir düzeyden sonra kızlar için daha güç, erkekler için daha kolay olabilmektedir. Yani maddenin örtük özellik ile gösterdiği ilişki gruplar arasında etkileşimli olarak farklılaşmaktadır.

DIF incelemelerinde alan karşılaştırmaları, parametre karşılaştırmaları ya da olabilirlik (likelihood) oranlarına dayalı model karşılaştırmaları gibi farklı modeller ve bu modellerin incelenebileceği bilgisayar programları bulunmaktadır (farklı modellerin ve bilgisayar programlarının ayrıntılı karşılaştırması için Bkz. Camilli ve Shepard, 1994; Kim ve Cohen, 1995; Smith, 2002; Thissen, Steinberg ve Wainer,



Şekil 1. Düzgün ve düzgün olmayan şekilli madde fonksiyon farklılaşması örnekleri

1993). Bu çalışmada farklı gruplardan elde edilen madde parametreleri ortak bir metrik üzerinde eşitlenerek karşılaştırılmıştır (PARSCALE 4.1, Muraki ve Bock, 2003).

Yöntem

Örneklem

Araştırma örneklemini, 16-26 yaşları arasında bulunan, 1807 üniversite öğrencisi oluşturmaktadır. Örneklemin yaş ortalaması 19.7, standart sapması 2.2'dir. Örneklemin %52'si (937) kız, %48'i (870) erkek öğrencilerden oluşmaktadır.

Veri Toplama Araçları

Araştırmada kullanılan ölçek, Beş Faktör Kişilik Envanterinin (Somer, Korkmaz ve Tatar, 2002) Yumuşak Başlılık (Agreeableness) boyutunun madde havuzu arasından seçilen 16 maddeden oluşmuştur:

1. Kolayca kızmam.
2. Telefonu birinin yüzüne kapatmışlığım vardır.
3. Hayal kırıklıklarımın acısını başkalarından çıkarırım.
4. Asla öfkelenmem.
5. Başkaları hakkında çabuk hüküm veririm.
6. Sakinliğimi korurum.
7. Tatmin edilmesi zor biriyim.
8. Kin tutarım.
9. Ailem ve arkadaşlarımla sık sık tartışırım.
10. Öküz altında buzağı arayan biriyim.
11. Dik kafalı ve inatçıyımdır.
12. Göze göz dişe diş taraftarıyım.
13. İstenenin tersini yaparım.
14. Sivri dilliyim.
15. Sabit fikirlerim vardır.
16. İnsanlara acı konuşurum.

Ölçekte düşük puanlar yumuşak başlılığa, yüksek puanlar antagonistik eğilimlere işaret etmektedir. Orijinal ölçek maddeleri 5'li Likert tipinde maddeler olmakla birlikte, çalışmamızda DIF incelemelerinde, madde karakteristik eğrilerinin grafikleri incelenirken grup karşılaştırmalarında kolaylık ve anlaşılabilirlik sağlanması amacıyla madde tepkileri, iki kategorili (dikotomik) hale getirilmiştir. Ayrıca Hulin, Drasgow ve Parsons (1983), dikotomik modellerin politomik modellere göre madde - cevap kuramının tek boyutluluk varsayımını ihlal etmeye karşı çok daha dayanıklı olduğunu, çok boyutluluk durumunda bile dikotomik madde formunda madde parametrelerinin oldukça tutarlı kaldığını ifade etmektedirler. Yukarıda açıklanan nedenlerle çalışmamızda, madde formatının dikotomize edilmesi tercih edilmiştir.

Çalışma ölçeğinde Beş Faktör Kişilik Envanterinin, Yumuşak Başlılık ölçeğindeki 45 maddenin tümü kullanılmamış, maddeler madde-cevap kuramının tek boyutluluk ve normallik varsayımını karşılayacak şekilde seçilerek çalışma için bir alt boyut oluşturulmuştur. Alt boyutun oluşturulmasında, madde analizi ve faktör analizlerinden yararlanılmıştır. Hambleton ve Swaminathan (1989); Hulin, Drasgow ve Parsons (1983)' ın belirttiği gibi psikolojik özelliklerin ölçülmesinde tek boyutluluk varsayımının pratikte karşılanması tam olarak mümkün olmamaktadır. Genellikle, kişilik, test alma becerileri ve ölçülen temel özellik dışında pek çok diğer faktörün test performansını etkilediği düşünülmektedir. Ackerman (1989), Traub (1983) test görevleri ile ilgili bu bilişsel faktörlerin sayısının kişiden kişiye değişebileceği gibi, maddeye cevap vermeyi etkileyen faktörlerin de maddeden maddeye değişiklik gösterebileceğini belirtmektedirler. Lord ve Novick (1968); Hambleton ve Swaminathan (1989); Hulin, Drasgow ve Parsons (1983), pratikte bu varsayımın karşılanması için, test performansını etkileyen "baskın" bir özellik ya da faktörün bulunmasının yeterli olduğunu ifade etmektedirler. Reckase (1979) de, ilk faktörün toplam varyansın %20'ini açıklamasının madde parametrelerinin tutarlı olarak tahminlenebilmesi için yeterli olduğunu belirtmektedir (akt. Collins, Raju, Edwards, 2000). Hambleton, Robin ve Xing (2000), faktör analizinin tek boyutluluk varsayımını kontrol etmekte mükemmel bir çözüm olmadığını, çünkü değişkenler arasında doğrusal ilişkiler

olduğunun varsayıldığını ancak yine de bu analizin test verilerin boyutsal yapısı hakkında oldukça iyi yaklaşımlar sağlayabildiğini ifade etmektedirler. Çalışmamızda 16 maddeden oluşan alt ölçeğin faktör analizleri sonucunda, ilk faktör hem kız hem de erkek öğrencilerde toplam varyansın yaklaşık %20'ini oluşturmuş, tüm maddeler ilk faktörden .30'un üzerinde yük almışlardır. Bu sonuçlar madde performanslarını etkileyen baskın bir faktörün varlığına ilişkin bir ipucu olarak değerlendirilmiştir.

Oluşturulan alt boyutta yüksek puanlar, antagonizm, dikbaşlılık gibi kişilik özelliklerine, düşük puanlar ise yumuşak başlılık, sakinlik, uzlaşılabilirlik gibi özelliklere işaret etmektedir. 16 maddelik alt boyutun Kuder-Richardson iç tutarlılık güvenirlik katsayıları, kızlar için ".73" erkekler için ".71 " olarak bulunmuştur ve toplam puanlar her iki grup için de normal dağılım göstermektedir.

Analizler

Bu çalışmada, madde parametrelerinin tahminleri madde-cevap kuramının iki parametreli lojistik modeline göre yapılmıştır. Madde parametrelerinin tahminlenmesi, a_i ve b_i parametrelerinin kız ve erkeklerde karşılaştırılabilmesi için ortak bir metrik oluşturulması ve madde fonksiyon farklılıklarına ilişkin X^2 değerlerinin elde edilmesi PARSCALE 4.1 (Muraki ve Bock, 2003) istatistik programı ile gerçekleştirilmiştir.

Bulgular

Analizlerde ilk olarak ölçek maddelerinin iki parametreli modele uyumu incelenmiş ve elde edilen sonuçlar tüm maddelerin, her iki grupta ayrı ayrı iki parametreli modele uyum sağladığını göstermiştir (Tablo 1).

Maddelerin modele uyumlu olduğu görüldükten sonra, kız ve erkek öğrenci örneklemleri için ayrı ayrı ayırdetme ve güçlük parametreleri tahminlenmiş ve kız öğrencilerin parametre değerleri referans alınarak, erkek öğrencilerin parametre tahminleri aynı metrik üzerinde ölçeklenmiştir. Tablo 2'de kız ve erkekler için maddelerin tahminlenen güçlük ve ayırt etme parametre değerleri verilmiştir.

Tablo 1

Maddelerin Kız ve Erkek Grupları İçin İki Parametreli Modele Uyum Değerleri

Maddeler	Kız			Erkek		
	χ^2	sd	*p*	χ^2	sd	*p*
1	5.47153	4	.24	1.78108	4	.78
2	2.72788	4	.60	3.13183	4	.54
3	2.57756	4	.63	3.41711	4	.49
4	1.88140	3	.60	0.55929	3	.90
5	4.52994	4	.34	2.55441	4	.64
6	2.96549	4	.57	4.26098	5	.51
7	7.27870	4	.12	5.96063	4	.20
8	3.37879	4	.50	10.37484	4	.03
9	5.68365	4	.22	1.40010	4	.85
10	3.22196	4	.52	2.07475	4	.73
11	6.07709	4	.19	9.67951	4	.05
12	6.69675	4	.15	5.34916	4	.25
13	5.50343	4	.24	7.31334	4	.12
14	9.20778	4	.06	5.20164	4	.27
15	5.39578	4	.25	7.48378	4	.11
16	5.09428	4	.28	3.91950	4	.42
Toplam	77.69201	63	.10	74.46194	64	.17

Tablo 2

Kız ve Erkek Grupları İçin Maddelerin Tahminlenen Parametre Değerleri

	Kız		Erkek	
Maddeler	Eğim (a)	Güçlük (b)	Eğim (a)	Güçlük (b)
	Std. Hata	Std. Hata	Std. Hata	Std. Hata
1	0.846	0.150	0.892	0.425
	0.106	0.094	0.112	0.102
2	0.568	0.295	0.893	0.658
	0.089	0.131	0.112	0.112
3	1.069	1.086	1.329	1.075
	0.124	0.121	0.152	0.109
4	0.786	-2.124	0.559	-1.862
	0.123	0.291	0.105	0.345
5	0.885	-0.135	0.969	0.103
	0.106	0.087	0.117	0.086
6	0.883	1.192	0.722	1.800
	0.114	0.153	0.113	0.263
7	0.993	-0.333	0.692	-0.838
	0.113	0.085	0.103	0.159
8	1.020	0.677	1.091	0.460
	0.115	0.099	0.126	0.087
9	0.867	1.229	0.928	0.631
	0.112	0.155	0.115	0.107
10	0.883	0.571	0.972	0.958
	0.110	0.104	0.119	0.125
11	1.130	-0.698	1.365	-0.039
	0.127	0.092	0.144	0.069
12	0.764	-0.066	0.756	-0.946
	0.099	0.098	0.108	0.157
13	1.215	1.257	1.047	1.468
	0.139	0.124	0.141	0.174
14	1.204	0.006	1.239	-0.045
	0.128	0.070	0.133	0.074
15	0.716	-0.619	0.642	-0.808
	0.098	0.124	0.105	0.165
16	1.213	0.853	1.105	0.300
	0.130	0.096	0.127	0.082

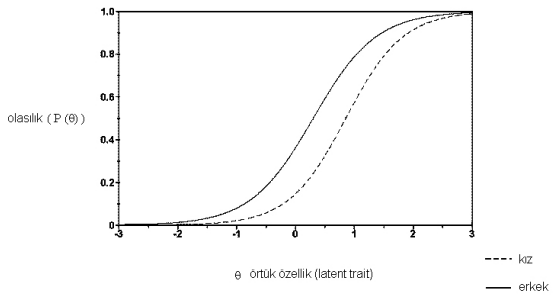
Eşit metrik üzerine getirildiğinde maddelerin gruplar için farklı fonksiyon gösterip göstermediği, madde karakteristik eğrilerinin parametrelerinin gruplar arası karşılaştırılması ile mümkün olmakta ve farklılıkların anlamlılığı X^2 istatistiği ile değerlendirilmektedir (bu çalışmada hata payı .01 olarak alınmıştır). Madde parametrelerinin gruplar arası karşılaştırması sonucu elde edilen X^2 değerleri ve istatistiksel anlamlılık düzeyleri ayırdetme parametreleri için Tablo 3'de, güçlük parametreleri için ise Tablo 4'te verilmiştir. Görüldüğü gibi, 5 madde güçlük parametresi açısından gruplar arasında farklılık göstermekte ancak ayırdetme parametreleri açısından hiçbir maddede gruplar arasında farklılık göstermemektedir. Yani tüm maddelerin madde karakteristik eğrilerinin şekilleri, gruplar arasında ben-

Tablo 3

Maddelerin Ayırdetme Parametrelerinin Kız ve Erkek Grupları İçin Karşılaştırma Sonuçları

Maddeler	Kontrast	Std. Hata	χ^2	sd	*p*
1	1.055	0.187	0.087	1	.76
2	1.573	0.315	3.298	1	.07
3	1.243	0.202	1.446	1	.23
4	0.711	0.173	2.776	1	.09
5	1.095	0.186	0.261	1	.62
6	0.818	0.167	1.192	1	.27
7	0.697	0.131	5.351	1	.02
8	1.070	0.173	0.162	1	.69
9	1.070	0.192	0.134	1	.71
10	1.101	0.193	0.274	1	.61
11	1.208	0.187	1.238	1	.27
12	0.990	0.191	0.003	1	.91
13	0.862	0.152	0.824	1	.37
14	1.029	0.156	0.035	1	.83
15	0.896	0.191	0.295	1	.59
16	0.911	0.143	0.385	1	.54
Toplam			17.762	16	.34

zerlik göstermekte fakat bazı maddelerin θ ölçeği üzerindeki konumları farklılaşmaktadır. Bu durumda gruplar arası farklılık gösteren maddelerin düzgün şekilli (uniform) olduğu kabul edilmektedir. Düzgün şekilli DIF gösteren bu maddeler 7, 9, 11, 12 ve 16 no'lu maddelerdir. Tablo 4'deki güçlük parametrelerinde görülen anlamlı düzeydeki farklılıklar, bazı maddelerde (7, 9, 12, 16. maddeler) ölçülen özelliğe aynı düzeyde sahip olan kız ve erkeklerden, kızlar için maddeyi evet olarak cevaplamanın daha zor olduğunu (kontrast kolonunda eksi değerler), bazılarında ise erkekler için (madde 11, kontrast kolonunda artı değer) daha zor olduğunu göstermektedir. Örneğin Madde 16, kızlar ve erkekler arasında farklılık gösteren bir maddedir ("İnsanlara acı konuşurum."). Bu maddenin karakteristik eğrisi Şekil 2'de verilmiştir. Şekilden her iki grubun madde karakteristik eğrileri incelendiğinde görülmektedir ki, maddeye verilen tepkiler ile bu tepkilere yol açtığı düşünülen gizil özellik arasındaki ilişki, karşılaştırma grupları açısından eşit değildir. Şekilde de görüldüğü gibi, eğrinin θ üzerindeki konumu kızlar için daha sağdadır yani b_i değeri daha yüksektir ve bu durum maddenin kızlar için daha güç bir madde olduğunu göstermektedir. Yani ölçülen özelliğe ayni düzeyde sahip olan kız ve erkekler için bu maddeyi "evet" olarak cevaplama olasılıkları farklılaşmaktadır. Aynı düzeyde antagonist eğilimlere sahip olan kız ve erkeklerden, erkek grubu bu maddeye daha kolayca evet diyebilirken, kızlar bu maddeye evet demekte zorlanmakta ve genel olarak maddeyi "evet" diye cevaplayabilmek için kızların daha yüksek θ seviyelerine çıkmaları gerekmektedir.



Şekil 2. Madde 16'nın kız ve erkek grupları için madde karakteristik eğrileri

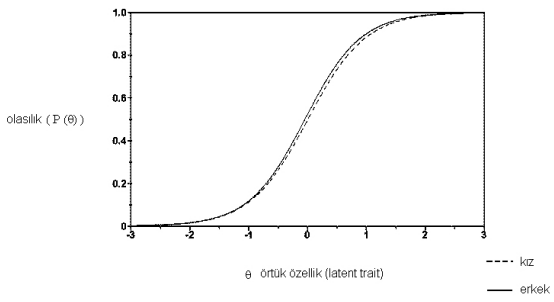
Tablo 4
Maddelerin Güçlük Parametrelerinin Kız ve Erkek Grupları İçin Karşılaştırma Sonuçları

Maddeler	Kontrast	Std. Hata	χ^2	sd	*p*
1	0.276	0.139	3.926	1	.05
2	0.363	0.172	4.423	1	.03
3	-0.010	0.163	0.004	1	.91
4	0.263	0.451	0.339	1	.57
5	0.238	0.123	3.777	1	.05
6	0.608	0.304	4.010	1	.04
7	-0.505	0.180	7.839	1	.01
8	-0.217	0.132	2.699	1	.10
9	-0.598	0.188	10.102	1	.00
10	0.387	0.162	5.689	1	.02
11	0.659	0.115	32.732	1	.00
12	-0.880	0.185	22.712	1	.00
13	0.211	0.214	0.972	1	.33
14	-0.051	0.102	0.254	1	.62
15	-0.189	0.206	0.840	1	.36
16	-0.554	0.126	19.149	1	.00
Toplam			119.467	16	.00

DIF analizlerinin ölçek eşdeğerliğinin sorgulanması dışında araştırmacıya sağlayacağı bir diğer katkı, incelenmekte olan yapının özellikleri hakkında bize ilave bilgi sağlamasıdır (Smith, 2002). Örneğin madde 16'daki ifadenin ("İnsanlara acı konuşurum.") kızlar tarafından daha zor kabullenilebilmesi muhtemelen toplumumuzdaki cinsiyet rollerinden etkilenmektedir. Kadınların daha merhametli ve hoşgörülü olmaları gerektiğine ilişkin kültürel yüklemlemeler, kızların saldırganlıklarını ifade etme biçimlerini ya da bu ifade biçimlerinin kendilerindeki varlığını kabullenmelerini erkeklerden farklılaştırmakta etkili olabilir (kültürlerarası incelemelerde de de DIF analizleri, saldırganlığın ifadesinde cinsiyet rollerinin etkisi konusunda çok yararlı bilgiler sağlayabilir).

Madde 11'de ise ("Dik kafalı ve inatçıyımdır."), kızlar ve erkekler için tersine bir DIF söz konusudur. Yani antagonizm boyutunda aynı düzeyde olan kız ve erkeklerden, kızlar kendilerini daha kolayca dik başlı ve inatçı olarak tanımlayabilirken (madde 11), acımasızlığı kabullenebilmeleri (madde 16) erkeklere göre daha zor olmaktadır. Buna karşın örneğin "Sivri dilliyim" (madde14) ifadesine "evet" cevabını verme olasılığı, aynı θ düzeyine sahip kız ve erkekler arasında farklılık göstermemektedir (Şekil 3).

DIF gösteren beş madde ölçekten çıkartılarak kalan 11 madde üzerinden analizler tekrarlanmış ve madde 8'in ("Kin tutarım.") gruplar arasında farklılık gösterdiği bulunmuştur. Bu maddenin de ölçekte çıkarılması ile kalan 10 madde üzerinden yenilenen analizler sonucunda, madde ve ölçek düzeyinde kız



Şekil 3. Madde 14'ün kız ve erkek grupları için madde karakteristik eğrileri

Tablo 5
Kalan 10 Maddenin Parametrelerinin Gruplararası Karşılaştırma Sonuçları

Kalan Maddeler	χ^2	sd	*p*
1	0.455	1	.51
2	1.042	1	.31
3	2.906	1	.08
4	0.070	1	.78
5	0.062	1	.79
6	3.581	1	.06
10	2.013	1	.15
13	0.047	1	.81
14	3.578	1	.06
15	2.342	1	.12
Toplam	16.095	10	.10

ve erkek öğrenciler için ölçme eşdeğerliğinin sağlandığı (Tablo 5) görülmüştür (aşağıda tartışma bölümünde de ele alındığı gibi DIF gösteren maddeleri ölçekten çıkarmak tek ve en uygun çözüm değildir. Bu çalışmada gerçek bir ölçek geliştirme çalışması yapılmadığı ve farklılık gösteren maddelerin çıkarılmasının etkilerini gösterebilmek amacıyla bu yola gidilmiştir).

Farklı fonksiyon gösteren maddelerin çoğunlukta olduğu bir ölçeğin toplam puanları üzerinden grup karşılaştırması yapılması yanıltıcı sonuçlar doğurabilmektedir. Örneğin bu çalışmada, kız ve erkek gruplarının ortalamaları ölçme eşdeğerliği sağlanmış toplam puanlar üzerinden karşılaştırıldığında kızlar ve erkeklerin Yumuşak Başlılık puanları arasında anlamlı bir farklılık çıkmazken (t = 0.14 sd = 1805, p > .001), madde güçlük-yerleşim parametresi (b_i) kızlar için daha düşük olan beş DIF maddesinin toplanması ile elde edilen puanlar üzerinden yapılan karşılaştırmada, kızlar anlamlı düzeyde (t = -7.45, sd = 1805, p < .001) daha az antagonist ya da daha yumuşak başlı çıkmaktadırlar.

Tartışma

Grup karşılaştırmalarında madde fonksiyon farlılıklarının incelenmesi, gruplar arasında ölçme eşdeğerliğinin sağlanmasına imkan vermesinin yanı sıra, incelenen yapının özellikleri hakkında değerli bilgiler sağlaması açısından da araştırmacıya önemli katkılarda bulunmaktadır. Ölçek eşdeğerliğinden emin olunmadan yapılacak grup karşılaştırmalarında, elde edilen farklılıkların gruplar arasındaki gerçek farklılıklardan mı yoksa ölçme yanlılığından mı kaynaklandığını bilmek mümkün olmamakta ve bu farklılıklarla ilgili olarak yapılan yorumlarda yanlış kanılara varılmasına yol açabilmektedir.

Gruplar arasında farklı fonksiyon gösteren maddeler belirlendikten sonra bu maddelerin dönüşümlü olarak (iterative - DIF gösteren her bir maddenin teker teker ölçekten çıkartılarak parametrelerin yeniden tahmin edilmesi ve DIF analizlerinin yeni parametreler üzerinden tekrarlanması) analizden çıkarılması araştırmacıya daha net bilgiler sağlayabilir. Çalışmamızda sunumu güç olduğu ve çok yer kaplayacağı düşüncesi ile bu yola gidilmemiş, DIF gösteren 4 madde birden çıkarılarak analizler tekrarlanmıştır. Ayrıca daha önce de işaret edildiği gibi telafi edici test fonksiyonlarının (compensatory TDIF) incelenmesi ile bazı maddeleri ölçekte tutma yoluna gidilmesi de tercih edilebilecek bir yoldur (ancak bu özel ve ayrıntılı bir konu olduğu için ayrı bir makale konusudur). Bazı araştırmacılar (Roznowski ve Reith, 1999; Waller, Thompson ve Wenk, 2000) maddeleri ölçekten çıkarmanın ölçülen yapının kapsamını ve varyansını daralttığını bu nedenle madde çıkarmak yerine, toplam ölçek içerisinde telafi edici özelliklere sahip olup olmadığının incelenmesinin önemine işaret etmektedirler. Gerçekten de yapının daraltılması, aşırı homojenleştirilmesi özellikle DIF gösteren maddeler çok olduğunda önemli bir sorun olarak test geliştiricinin karşısına çıkabilmektedir. Ayrıca farklı fonksiyon gösteren maddelerin revizyonuna mı gidileceği, ölçekten çıkartılacağı mı yoksa ölçekte tutulacağı mı konusunda karar verilirken yapılan çalışmanın türünün de önemli olduğu düşünülmektedir. Örneğin çalışma belirli grupların karşılaştırması amaçlı değil de test geliştirme amaçlı bir çalışma ise, hangi demografik özellikler ve gruplar için DIF yapılacak, bu nereye kadar sürdürülecek ve elimizde

ölçülen yapıdan ne kalacaktır sorusu önemli bir problematik oluşturmaktadır. Ancak araştırmada grup karşılaştırmaları yapmak amaçlandığında, ve bu farklılıkların gerçek farklılıklar olduğunu varsayabilmek söz konusu olduğunda ölçek eşdeğerliğinin sağlanmasında daha titiz davranılması gerektiği ve karşılaştırma gruplarının ölçüm eşdeğerliğinin incelenmesinin gereği açıktır. Sonuç olarak DIF gösteren maddelerin nasıl bir işleme tabi tutulacağı araştırmacının vereceği bir karardır ancak hangi yol tercih edilirse edilsin, gözlenen test puanları ile bunların altında yatan örtük özellik arasındaki ilişkinin ayrıntılı bir şekilde incelenmesi araştırmacının üzerinde çalıştığı yapıyı daha iyi tanımasına katkıda bulunacak ve yorumlarının daha güçlü olmasına yardımcı olacaktır.

Çalışmamızda ölçme eşdeğerliğinin incelenmesinde modern test kuramı kapsamında geliştirilmiş olan parametre karşılaştırmalarına dayalı bir DIF modeli kullanılmış ve örnek bir uygulaması yapılarak bulgular tartışılmıştır. Ölçme eşdeğerliğinin incelenmesinde yaygın olarak kullanılmakta olan bir diğer yaklaşım ise, klasik ölçme kuramı temelinde geliştirilmiş olan Yapısal Eşitlik Modellemelerine (Structural Equation Modeling-SEM) dayalı analizleri kapsamaktadır. Aşağıdaki bölümde bu yaklaşımın özelliklerine kısaca değinilmiş ve IRT modelleri ile gösterdiği benzerlik ve farklılıklar kısaca ele alınmıştır.

Yapısal Eşitlik Modellemeleri bağlamında ölçme eşdeğerliğinin incelenmesi temelde maddeler arasındaki kovaryans yapısının eşdeğerliğinin gruplar arasında karşılaştırılmasına dayanmakta ve bu amaçla doğrulayıcı faktör analizinden yararlanılmaktadır. Doğrulayıcı faktör analizi ile gruplar için elde edilen faktör yükleri gruplar arasında karşılaştırılmaktadır (yapısal eşitlik modellemelerinin ayrıntılı incelemesi ve ilgili bilgisayar programları için Bkz. Hayduk, 1987; Schumacker ve Lomax, 1996). Bu karşılaştırma sonucunda faktör yüklerinin yapısının gruplar arasında farklılaşmadığının bulunması, ölçek maddelerinin temsil ettikleri örtük özellikle gösterdikleri ilişkinin, gruplar arasında farklılaşmadığı anlamına gelmektedir. Faktör yüklerinin yanı sıra, çok boyutlu modellerde faktörler arasındaki covaryans yapısının da gruplararası eşdeğerliği incelenmektedir. Byrne, Shavelson ve Muthen (1989); Muthen ve Christoffersson (1981), ölçme eşdeğerliği çalışmalarında çoğunlukla yalnızca kovaryans yapısının karşılaştırıldığını, bunun ise eşdeğerliği göstermekte yetersiz olduğunu ifade etmektedirler. Byrne ve arkadaşları (1989), kovaryans yapısının yanı sıra, regresyon kesme noktasının ve ortalama yapılarının da içerilmesinin gereğine dikkat çekmektedirler. Muthen ve Christoffersson (1981), gruplar arası karşılaştırmalarda dikotomik değişkenlerle faktör ortalamalarının kullanılmasının, yaygın olarak kullanılan madde puanlarının (0-1'lerden oluşan madde puanları) toplanmasına bir alternatif oluşturduğunu vurgulamaktadırlar. Raju, Laffitte ve Byrne (2002), diğer bazı araştırmacıların da gruplar arası ölçme eşdeğerliğinin olduğunun söylenebilmesi için, yalnızca kovaryans matrislerinin yeterli olmadığı, yanı sıra regresyon doğrusunun Y eksenini kesme noktasının da karşılaştırılmasından yana olduklarını ifade etmektedirler. Ancak ve Byrne ve arkadaşları (1989) ve Raju ve arkadaşları (2002), bazı durumlarda kesme noktasının farklı bulunmasının yorumunun problematik olabildiğine ve bu konuda daha fazla araştırmaya gerek duyulduğuna dikkat çekmektedirler. Smith (2002) ise kovaryans yapısının eşdeğerliğinin sağlanması ile problemlerin bitmediğine ve Millsap'ın paradoksuna dikkat çekmektedir. Millsap (1997), kovaryans yapısının eşdeğerliği sağlanmış olsa bile, grupların varyanslarının farklılaşması durumunda bu kez regresyon eşitsizliğinin ortaya çıktığını belirtmektedir. Millsap her iki eşitliğin sağlanabilmesinin günlük hayatta elde edilen verilerde karşılanabilme olasılığının çok zayıf olduğuna işaret etmektedir.

Yapısal Eşitlik Modellemelerine dayalı analizler IRT temelli DIF analizleri ile karşılaştırıldığında bir çok benzer yönleri olmakla birlikte bazı farklılıkları da olduğu görülmektedir. Reise, Widaman ve Pugh (1993), IRT modellerindeki "α" parametresinin madde tepkilerinin, örtük özellikle gösterdiği ilişki-

nin bir ifadesi olduğunu, doğrulayıcı faktör analizlerinde de faktör yüklerinin (λ) aynı fonksiyonu gösterdiklerini ifade etmektedirler. "α" katsayısında olduğu gibi "λ" değerleri de arttıkça madde ile örtük özellik arasındaki ilişki kuvvetlenmektedir. Her iki model de iki farklı örneklemden olup, örtük özellik üzerinde eşit konumda bulunan kişilerin gerçek puanlarının benzerlik derecesini incelemektedir. Örtük özellik ile kişilerin maddeye doğru cevap vermeleri arasındaki ilişki doğrulayıcı faktör analizinde doğrusal iken, IRT modellerinde doğrusal değildir (Raju ve ark., 2002, Reise ve ark. 1993). Ancak Mellenbergh (1994), bazı IRT modellerininin, doğrusal regresyon modellerinin özel hali olduğunu ifade etmekte ve DIF analizlerini de Genellenmiş Doğrusal Madde Cevap Kuramı çerçevesinde incelemektedir.

Raju ve arkadaşları (2002), maddelerin dikotomik olarak puanlanması durumunda lojistik bir regresyon modelinin ölçülen yapı ve değişken arasındaki ilişkiyi temsil etmeye daha uygun olduğunu ve bu bağlamda, dikotomik verilerle IRT modellerini kullanmanın daha uygun olabileceğini ifade etmektedirler. Doğrulayıcı faktör analizleri madde düzeyinde ele alındığında eğer maddeler dikotomik ise, yaygın olarak kullanılmakta olan bazı bilgisayar paket programları uygun olmayan sonuçlar verebilmektedir (Reise, Waller ve Comrey, 2000;. Hoijtink, Rooks ve Wilmink, 1999). Ancak Reise ve arkadaşları Mplus gibi bilgisayar programlarının bu sorunu çözdüklerini belirtmektedirler. Glöckner-Rist ve Hoijtink de (2003), Mplus programı ile yaptıkları örnek uygulamada, dikotomik maddelerin analizinde her iki modelin içiçe kullanımının mümkün olduğunu, madde cevap kuramına dayalı modellerin faktör analizi temel alınarak modellenebildiğini göstermekte ve her iki modelin güçlü ve zayıf yönlerine açıklık getirmektedirler (Örn., IRT modellerinin bireysel yorumlarda daha güçlü iken yapısal eşitlik modellemelerinin açık ve örtük değişkeler arasındaki ilişkileri modellemede daha güçlü olduğunu ifade etmektedirler). Takane ve Leeuw (1987) da, çalışmalarında dikotomik değişkenlerde IRT (iki parametreli normal ogiv modelinde) ve faktör analizlerinin marjinal olabilirliklerinin eşdeğerliğini göstermektedirler.

Tek boyutluluk varsayımının faktör analizi yoluyla incelenmesi madde-cevap kuramına dayalı modellerde model parametrelerinin yordanması ile eş zamanlı olarak test edilememesine karşın, bu varsayım yapısal eşitlik modellerinde eşzamanlı olarak test edilebilmektedir. Yine Raju ve arkadaşları (2002), tek boyutlu modellerde IRT temelli DIF analizlerinden rahatça yararlanmak mümkün iken, çok boyutluluk durumunda yapısal eşitlik modellemelerinin daha kullanışlı olduğuna işaret etmektedirler. (Rasch model temelinde çok boyutlu IRT uygulamaları ile doğrulayıcı faktör analizinin ilişkisi için Bkz. Hoijtink, Rooks ve Wilmink, 1999 ve çok boyutlu IRT uygulamaları ve faktör modellerini entegre bir biçimde ele alan bir çalışma için Bkz. Glöckner-Rist ve Hoijtink, 2003).

Sonuç olarak her iki model de, ölçme eşdeğerliğinin sağlanmış olması ile, grupların performans dağılımlarının aynı şey olmadığını, tam tersi bu gruplar arası farklılığın incelenebilmesi için ölçme eşdeğerliğinin sağlanmış olması gereğini göstermektedirler. Ölçme eşdeğerliğinin olmadığı bulgusu ortaya çıktığında, her iki yöntem de bunun kaynaklarının araştırılmasında önemli ipuçları sağlamaktadır (IRT ve Yapısal Eşitlik Modellemelerine dayalı analizlerin ayrıntılı karşılaştırmaları için bakınız; Raju ve ark., 2002; Reise ve ark. 1993; Reise ve Widaman, 1999; Hoijtink, Rooks ve Wilmink, 1999). Lambert ve ark. (2003); Reise, Keith ve Widaman, (1999) ile Glöckner-Rist ve Hoijtink'in (2003) çalışmaları ise, bu iki yöntemin ölçeklerin psikometrik özelliklerini incelemede birarada kullanılmasının sağlayacağı zenginliklere iyi birer örnek oluşturmaktadır.

Kaynaklar

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and non-compensatory multidimensional items. *Applied Psychological Measurement, 13* (2), 113-117.

Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychological Bulletin, 105* (3), 456-466.

Camilli, G., & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items.* California: Sage Pub. Inc.

Collins, C. C., Raju, S. N., & Edwards, J. E. (2000). Assessing differential functioning in a satisfaction scale. *Journal of Applied Psychology*, *85* (3), 451-461.

Ellis, B. B., & Mead A. D. (2000). Assessment of measurement equivalence of a Spanish translation of the 16PF Questioanaire. *Educational and Psychological Measurement, 60* (5), 787-807.

Ferrando, P. J. (2001). The measurement of neuroticism using MMQ, MPI, EPI and EPQ items: a psychometric analysis based on item response theory. *Personality and Individual Differences*, *30*, 641-656.

Geisinger, K. F. (1994). Cross-cultural normative assesment: translation and adaptation issues influencing the normative interpretation of assesment instruments. *Psychological Assesment, 6* (4), 304-312.

Hambleton, R. K., Frederic, R., & Xing, D. (2000). Item Response Models for the analysis of educational and psychological test data Personality Research. In (Eds.) H. E. A. Tinsley & S. D. Brown, *Handbook of Applied Multivariate Statistics and Mathematical Modeling.* (pp. 553-581). San Diago: Academic Press.

Hambleton, R. K., & Swaminathan, H. (1989). *Item Response Theory, Principles and Applications.* Kluwer Nijhoff Publishing, Boston.

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

Hayduk, L. A. (1987). *Structural Equation Modeling with LISREL, Essentials and advances.* The John Hopkins Press Ltd., London..

Hoijtink, H., Rooks, G., & Wilmink, F. W. (1999). Confirmatory Factor analysis of items with a dichotomous response format using the multidimensional Rasch Model. *Psychologicl Methods, 4* (3), 300-314.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item Response Theory: Application to Psychological Measurement.* Illinois: Dow Jones-Irwin.

Kim, S-H., & Cohen, A. S. (1995). A comparison of Lord's Chi-Square, Raju's Area Measures, and Likelihood Ratio Tests on detection of Differential Item functioning. *Applied Measurement in Education, 8* (4), 291-312.

Lambert M. C., Schmitt, N., Vaughan, m. E. S., An, J. S., Fairclough, M., & Nutter, C. A. (2003). Is it prudent to administer all items for Each Child Behavior Checklist cross-informant syndrome? Evaluating the psychometric Properties of the Youth Self-Report Dimensions with confirmatory factor analysis and item response theory. *Psychological Assessment, 15* (4), 550-568.

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods, 8* (1), 72-87.

Mellenbergh, G. J. (1994). Generalized lineer Item Response Theory. *Psychological Bulletin, 115* (2), 300-307.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the Single-factor case. *Psychological Methods, 2* (3), 248-260.

Muraki, E., & Bock, R. D. (2002). *PARSCALE: Parameter scaling of rating data* (Version 4. 1) (Software Manual). Chicago: Scientific Software Inc.

Muthen, B., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika, 4674,* 407-419.

Orlando, M., & Rand, G. N. M. (2002). Differential item functioning in a spanish translation of the PTSC Checklist: Detection and evaluation of impact. P*sychological Assesment, 1* (1), 50-59.

Raju, N. S. (1990). Determining the significance of estimated signed and and unsigned areas between two item response functions. *Applied Psychological Measurement, 14* (2), 197-207.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalance: a comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87* (3), 517-529.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19* (4), 353-368.

Reise, S. P. (1999). Personality measurement issues viewed through the eyes of IRT. In (Eds.) S. E Embretson., S. L. Hershberger *The New Rules of Measurement: What Every Psychologist Should Know* (pp. 219-241). Mahwah, New Jersey: L. Earlbaum Associates, Publishers.

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods, 8* (2), 164-184.

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12* (3), 287-297.

Reise, S. P., & Widaman, K. F. (1999). Assessing the fit of measurement models at the individual level: Am response theory and covariance structure approach. *Psychological Methods, 4* (1), 3-21.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and Item Response Theory, two approaches for exploring measurement invariance. *Psychological Bulletin, 114* (3), 552-566.

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containin differentially functioning items: do biased items result in poor measurement. *Educational and Psychological Measurement, 59*, 248-270.

Schumacker, R. E., & Lomax, R. G. (1996). *A beginner's guide to Structural Equation Modeling.* Lawrencw Erlbaum Associates Inc., New Jersey.

Somer, O. (1999). Çok Kategorili (Polytomous) Maddelerde, Klasik ve Modern Test Kuramları İle Madde Analizleri, Güvenirlik ve Bilgi Kavramlarının Karşılaştırılması. *Türk Psikoloji Dergisi, 14* (44), 63-78.

Somer, O. (1998). Kişilik Testlerinde Klasik ve Modern Test Kuramları İle Madde Analizi. *Türk Psikoloji Dergisi, 13* (41), 1-17.

Somer, O., Korkmaz, M., & Tatar, A. (2002). Beş Faktör Kişilik Envanteri'nin Geliştirilmesi I: Ölçek ve Alt Ölçeklerin Oluşturulması. *Türk Psikoloji Dergisi, 17* (49), 21-33.

Smith, L. L. (12002). On the usefulness of item bias analysis to personality psychology. *Society for Personality and Social Psychology, 28* (7), 754-763.

Steinberg, L. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology, 81* (2), 332-343.

Steinberg, L., & Thissen, D. (1995). Item Response Theory in Personality Research. In (Eds.) P. E. Shrout & T. Fiske, *Personality Research, Methods, and Theory.* (pp. 161-181). Hillsdale, N. J. : Earlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of Item Response Theory. In (Eds.) P. H. Holland & H. Wainer, *Differential Item Functioning*, (pp. 67-113). Hillsdale, NJ: Earlbaum.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R.K. Hambleton (Ed.), *Applications of Item Response Theory.* Vancouver. BC: Educational Ressearch Institude of British Columbia.

Van de Vijver, Fons J. R., & Leung, K. (2000). Methodological issues in psychological research. *Journal of Cross-Cultural Psychology, 31* (1), 19-33.

Waller, N. G., Thompson, J. S., & Wenk, E. (2000). Using IRT to seperate measurement bias from true group differences on homogeneous and heterogeneous Scales: An illustration with the MMPI. *Psychological Methods, 5* (1), 125-146.

Weiss, D. J. (1983). *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing.* New York: Academic Press.

Türk Psikoloji Dergisi
2004, 19 (53), 83 - 86

Summary

Measurement Equivalence in Group Comparisons: Differential Item and Test Functioning

Oya Somer*
Ege Üniversitesi

Both in within and cross-cultural settings, measurement equivalence is one of the most important methodological problems in comparisons of group differences. Models based on Item Response Theory are being widely used in recent years in holding measurement equivalence. These models (generally take place under the title of Differential Item-Test Functioning-DIF, DTF) refer to the methods analyzing the relations between observed scores and the latent attribute measured by the test, across comparison groups. The existence of DIF-DTF is evidenced when these relations are different across comparison groups. DIF means, indicating the probability of endorsing an item show differences for members of the reference group and the focal group that are having the same latent trait level. In this study, DIF analyses of 16 items of a personality scale (agreeableness-antagonism) were performed on a student sample. The aim of the study was to illustrate application of IRT-DIF analysis to a personality scale and to encourage the researchers for using these models more widely in their studies.

Method

Participants and Instrument

Participants were 1807 undergraduates, 870 men and 937 women. Mean age of the sample was 19,7 and standard deviation was 2,2.

All of the participants completed Agreeableness-Antagonism sub dimension of the Five Factor Personality Inventory (Somer, Korkmaz & Tatar, 2002). The original five-point Likert type item format was dichotomized because parameter estimates were more robust with dichotomous format against possible violations of unidimensionality assumption of IRT. IRT estimates of two parameter logistic model and DIF analysis (women were the reference group and the men the focal group) were performed using Parscale 4.1 (Muraki & Bock, 2003).

Results

Item parameter estimates (for all of the 16 items) are fitted to the two parameter logistic IRT model ($p > .01$). The results of fit statistics are presented for men and women in Table 1. The item parameter estimates of 16 item Agreeableness scale for men and women are presented in Table 2. The results of DIF analysis for discrimination parameters are presented in Table 3 and for location parameters in Table 4 . According to the results of DIF analysis some of the items showed uniform DIF (single degree of freedom X^2, $p < .01$) between men and women.

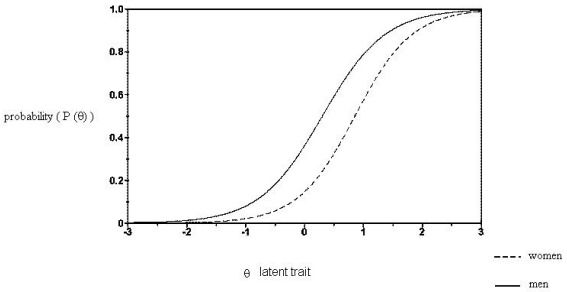


Figure 1. ICC's of item 16 for men and women.

*Address for Correspondence: Oya Somer, Ege Üniversitesi, Edebiyat Fakültesi, Psikoloji Bölümü Bornova, İzmir, Turkey.
E-mail: osomer@edebiyat.ege.edu.tr

The items showing DIF on location parameter were 7, 9, 11, 12, and 16. Item 7 (I am hard to satisfy.), item 9 (Often argue with my family and friends.), item 12 (Believe in an eye for an eye.), and item 16 (Say bitter things to people.) are located on the lower levels of latent trait for men compared to women (the higher scores were indicating antagonistic tendencies). As an example, ICC's of "item 16" for men and women were presented in Figure 1. It can be seen from the figure that, the relationships between the item responses and the latent trait are different across men and women. This means that for the men and women which are on the same position on the latent trait, the item is easier for men to respond as "yes" than it is for women. The expectations in Turkish culture that women should be more tenderhearted and tolerant may make it hard for women to accept these kind of statements addressing themselves as pitiless compared to men which are culturally more tolerated when behaved in this manner (this point related to gender roles may be similar in many cultures, and it is also a subject which can be studied by DIF efficiently). On the contrary, the ICC of item 11 (I obstinate and stubborn.) is located on the higher levels of theta for men than women which means that it is easier for women to say "yes" or accept for their self image being obstinate and stubborn than men who have the same level of antagonistic tendencies. The properties of the other DIF items and how to handle them are discussed in the article.

Excluding DIF items, the analyses were continued until finding a subscale which has no DIF items that is thought as having measurement equivalence for the comparison groups (although it is not the most convenient way to exclude DIF items for holding measurement equivalence, with the aim of an illustration of the impact of DIF items on the group mean comparisons, this method is preferred). While no significant (t = 0.14, df. = 1805, p > .001) differences were found between men and women on the agreeableness scores that have measurement equivalence, the scores of a subtest including mostly DIF items for women, gave statistically meaningful differences (t = -7.45, df.= 1805, p < .001) implying that women are more agreeable or less antagonistic than men. This results show that scales including DIF items may lead to find unrealistic group differences (measurement bias versus true group differences) in comparison studies.

Another widely used method in measurement equivalence is confirmatory factor analysis which is based on Structural Equation Modeling (SEM). The inter-item covariance structure is compared between groups in SEM based measurement equivalence studies. Finding no differences in the covariance structure between groups is interpreted as the evidence of measurement equivalence. Lately most of the authors state that besides the covariance structure, the mean structure and the intercept should also be subjected to comparison between groups, for deciding measurement equivalence. A brief comparison between IRT and SEM based models is made and discussed in the final section. Both of the models tell us that, finding meaningful differences between group means cannot be interpreted as true group differences without holding measurement equivalence. At the end of this discussion, it is concluded that both of the models have some weaknesses and strengths, and using these two approaches in conjunction will enhance the psychometric quality of the studies.

Table 1

Fit Statistics for Two Parameter Logistic Model

Items	Women			Men		
	χ^2	df	*p*	χ^2	df	*p*
1	5.47153	4	.24	1.78108	4	.78
2	2.72788	4	.60	3.13183	4	.54
3	2.57756	4	.63	3.41711	4	.49
4	1.88140	3	.60	0.55929	3	.90
5	4.52994	4	.34	2.55441	4	.64
6	2.96549	4	.57	4.26098	5	.51
7	7.27870	4	.12	5.96063	4	.20
8	3.37879	4	.50	10.37484	4	.03
9	5.68365	4	.22	1.40010	4	.85
10	3.22196	4	.52	2.07475	4	.73
11	6.07709	4	.19	9.67951	4	.05
12	6.69675	4	.15	5.34916	4	.25
13	5.50343	4	.24	7.31334	4	.12
14	9.20778	4	.06	5.20164	4	.27
15	5.39578	4	.25	7.48378	4	.11
16	5.09428	4	.28	3.91950	4	.42
Total	77.69201	63	.10	74.46194	64	.17

Table 2

Estimated IRT Parameters of Two-Parameter Model for Men and Women

Items	Women		Men	
	Slope (a) Std. error	Location (b) Std. error	Slope. (a) Std.Error	Location (b) Std. error
1	0.846	0.150	0.892	0.425
	0.106	0.094	0.112	0.102
2	0.568	0.295	0.893	0.658
	0.089	0.131	0.112	0.112
3	1.069	1.086	1.329	1.075
	0.124	0.121	0.152	0.109
4	0.786	-2.124	0.559	-1.862
	0.123	0.291	0.105	0.345
5	0.885	-0.135	0.969	0.103
	0.106	0.087	0.117	0.086
6	0.883	1.192	0.722	1.800
	0.114	0.153	0.113	0.263
7	0.993	-0.333	0.692	-0.838
	0.113	0.085	0.103	0.159
8	1.020	0.677	1.091	0.460
	0.115	0.099	0.126	0.087
9	0.867	1.229	0.928	0.631
	0.112	0.155	0.115	0.107
10	0.883	0.571	0.972	0.958
	0.110	0.104	0.119	0.125
11	1.130	-0.698	1.365	-0.039
	0.127	0.092	0.144	0.069
12	0.764	-0.066	0.756	-0.946
	0.099	0.098	0.108	0.157
13	1.215	1.257	1.047	1.468
	0.139	0.124	0.141	0.174
14	1.204	0.006	1.239	-0.045
	0.128	0.070	0.133	0.074
15	0.716	-0.619	0.642	-0.808
	0.098	0.124	0.105	0.165
16	1.213	0.853	1.105	0.300
	0.130	0.096	0.127	0.082

Table 3

DIF Results for Slope Parameters

Items	Contrast	Std. Error	χ^2	df	*p*
1	1.055	0.187	0.087	1	.76
2	1.573	0.315	3.298	1	.07
3	1.243	0.202	1.446	1	.23
4	0.711	0.173	2.776	1	.09
5	1.095	0.186	0.261	1	.62
6	0.818	0.167	1.192	1	.27
7	0.697	0.131	5.351	1	.02
8	1.070	0.173	0.162	1	.69
9	1.070	0.192	0.134	1	.71
10	1.101	0.193	0.274	1	.61
11	1.208	0.187	1.238	1	.27
12	0.990	0.191	0.003	1	.91
13	0.862	0.152	0.824	1	.37
14	1.029	0.156	0.035	1	.83
15	0.896	0.191	0.295	1	.59
16	0.911	0.143	0.385	1	.54
Total			17.762	16	.34

Table 4

DIF Results for Location Parameters

Items	Contrast	Std. Error	χ^2	df	*p*
1	0.276	0.139	3.926	1	.05
2	0.363	0.172	4.423	1	.03
3	-0.010	0.163	0.004	1	.91
4	0.263	0.451	0.339	1	.57
5	0.238	0.123	3.777	1	.05
6	0.608	0.304	4.010	1	.04
7	-0.505	0.180	7.839	1	.01
8	-0.217	0.132	2.699	1	.10
9	-0.598	0.188	10.102	1	.00
10	0.387	0.162	5.689	1	.02
11	0.659	0.115	32.732	1	.00
12	-0.880	0.185	22.712	1	.00
13	0.211	0.214	0.972	1	.33
14	-0.051	0.102	0.254	1	.62
15	-0.189	0.206	0.840	1	.36
16	-0.554	0.126	19.149	1	.00
Total			119.467	16	.00