Türk Psikoloji Yazıları
2006, 9 (18) 63-80

Test ve Ölçek Geliştimede Yeni Yaklaşımlar: Madde Cevap Kuramı Kapsamında Madde İşlevsel Farklılık (Madde Yanlılık) Yöntemleri

Mediha Korkmaz*
Ege Üniversitesi

Özet

Madde cevap kuramına temellenen madde işlevsel farklılık çalışmaları, farklı alt grup/evren üyelikleri olan gruplar arası karşılaştırmalarda olduğu kadar, farklı kültürler arasında da madde/ test işlevsel farklılığının belirlenmesinde ya da ölçme eşdeğerliğinin incelenmesinde oldukça güçlü yöntemler sağlamaktadır. Bir ölçek ya da madde aynı örtük özelliği her iki grupta aynı şekilde ölçemiyorsa (ölçme eşdeğerliği yoksa) testten elde edilen sonuçlara göre gruplar arasındaki benzerlik ya da farklılıklar, ilgilenilen örtük özellik kapsamında yorumlanamaz. Geleneksel test kuramı içerisinde bu olgu ölçek/test yanlılığı ya da madde yanlılığı olarak, madde cevap kuramı kapsamında ise madde işlevsel farklılığı olarak adlandırılmıştır. Bu yazıda madde cevap kuramına dayalı olarak üç madde işlevsel farklılık yönteminden söz edilecektir. Bu yöntemler; 1- Olabilirlik oranı testi model karşılaştırma yöntemi 2- Madde-test işlevsel farklılık yöntemi, 3- Parametre karşılaştırma yöntemi' dir.

Anahtar Kelimeler: Madde işlevsel farklılığı, madde yanlılığı, madde cevap kuramı

The New Approaches in Scale Development: Methods of Differential Item Functioning (Item Bias) Based on Item Response Theory

Abstract

Studies of differential item functioning (DIF) established on item response theory (IRT), provides quite powerful methods in examining the measurement equivalence or determining the item-test differential functioning among different cultures as well as in the comparison among groups that has members from sub group/population. When a scale or an item does not measure the same latent trait or ability for both groups, the results are not interpretable with respect to the latent trait. Traditionally, this phenomenon has been called scale or item bias. More recently, item bias is called as differential item functioning (DIF) in the perspective of the item response theory. This article provides a brief primer overview of three differential item functioning methods based on the item response theory. These methods are IRT likelihood ratio test (LR) with comparison model, IRT differential functioning of items and tests (DFIT), and parameter comparison method.

Key Words: Differential item functioning (DIF), item bias, item response theory

*Yazışma Adresi: Öğretim Görevlisi Mediha Korkmaz, Ege Üniversitesi, Edebiyat Fakültesi Psikoloji Bölümü, Bornova/İzmir-35100
E-posta: mediha.korkmaz@ege.edu.tr

Yazar Notu: Bu çalışma, yazarın Doktora tezinin bir bölümünü içermektedir. Bu çalışmanın danışmanlığını yapan ve katkılarından yararlandığım Prof. Dr. Oya Somer'e çok teşekkür ederim. Ayrıca çalışma Ege Üniversitesi Araştırma Fon Saymanlığı (2001/EDE/011 nolu proje) tarafından desteklenmiştir.

Günümüzde psikoloji insan davranışlarını ve bu davranışların temelinde bulunan özellikleri anlamaya çalıştığı kadar bir taraftan da bu özellikleri aydınlatabilmek için yöntemsel olarak da gittikçe güçlenmektedir. İnsanoğlunun karmaşık psikolojik yapılarının ölçümünde ileri düzeyde niceliksel yöntemlere duyulan ihtiyaç, son yıllarda madde cevap kuramının ve dolayısıyla çeşitli modellerinin gelişmesine katkı sağlamıştır. Madde cevap modelleri, eğitim alanındaki ölçmelerde önemli bir etkiye sahip olduğu kadar psikoloji alanında da yetenek ve zeka değerlendirmelerinin yanı sıra, özellikle kişilik ve tutum ölçümlerinde de giderek sıklıkla kullanılan yöntemler haline gelmişlerdir.

Modern test kuramı olarak bildiğimiz madde cevap kuramı (item response theory) ya da örtük özellikler kuramı (latent trait theory), Rasch'ın 1960'lı yıllarda ilk çalışmalarını yayımlanması ile ortaya çıkmış gibi görünmekle birlikte, aslında Thurstone'nun 70 yıl önce tanımladığı psikolojik ölçmenin mantığına dayanmaktadır (Camilli ve Shepard, 1994; Crocker ve Algina, 1986). 1980'lerde madde cevap kuramı, ölçme uzmanları arasında klasik test kuramının yanı sıra en baskın çalışma konusu olmuş, günümüzde ise artık yeni bir ölçme aracının geliştirilmesi, soru bankalarının hazırlanması, gruplar arası karşılaştırmalarda madde yanlılığının incelenmesinde olduğu kadar kültürler arası karşılaştırmalarda ölçme eşdeğerliğinin araştırılması ve dolayısıyla testlerin geçerliğinin sınanması gibi pek çok alanda sıklıkla kullanılır hale gelmiştir.

Madde cevap kuramı, bireylerin davranışlarıyla, bu davranışların altında örtük/gizil olarak bulunduğu varsayılan yetenek, tutum, kişilik vb. özellikler arasındaki ilişkileri, olasılığa temellenen matematiksel fonksiyonlar ve modellerle açıklamaktadır (Hambleton ve Swaminathan, 1989; Hambleton, Swaminathan ve Rogers, 1991; Hulin, Drasgow ve Parsons, 1983). Bu modeller, gözlenen değişkenler ile bunların altında bulunan örtük özellik arasındaki işlevsel ilişkiyi doğrusal olmayan bir regresyon ile tanımlarlar (Chernyshenko, Stark, Chan, Drasgow ve Williams, 2001; Zickar, 1998). Madde cevap kuramının merkezi elemanı, bireyin ölçülen yetenek boyutundaki düzeyi ile maddeye doğru cevap verme olasılığı arasındaki ilişkinin grafik gösterimini sağlayan madde karakteristik eğrisidir (Item Characteristic Curve, ICC; bkz. Şekil 1-2). Madde cevap modeline bağlı olarak (1, 2 ya da 3 parametreli) bir madde karakteristik eğrisi, madde ayırt edicilik parametresi (a_i), madde güçlük parametresi (b_i) ve maddenin doğru yanıtını tahmin parametresi (c_i) içermektedir. Madde ayırt edicilik parametresi, madde cevap fonksiyonunun eğimini, diğer bir ifadeyle madde karakteristik eğrisinin dikliğini belirlemektedir. Madde güçlük parametresi ise, test ile ölçülen yetenek, tutum kişilik vb. örtük özellik boyutu (theta, q) üzerinde yer alır ve bir katılımcının doğru cevabının koşula bağlı olasılığını gösterir. c_i parametresi de düşük yetenekli kişilerin bir test maddesini şansa bağlı olarak doğru cevabı tahminleme olasılığını ifade eder. Madde cevap kuramının diğer en önemli özelliklerinden biri de, testte bulunan her bir madde için madde bilgi fonksiyonlarının ve toplam test puanı için de test bilgi fonksiyonlarını içermesidir. Test ve madde bilgi fonksiyonları, klasik test kuramında güvenirlik ve ölçmenin standart hatasını karşılayan, ölçmenin doğruluğu, hassasiyeti ve mükem-

melliği hakkında bir değerlendirme yapılmasını sağlayan fonksiyonlardır (Camilli ve Shepard, 1994; Hambleton ve ark., 1991; Hambleton, Robin ve Xing, 2000; Hulin ve ark., 1983; Somer, 1999; Zickar, 1998).

Madde ve Test İşlevsel Farklılık Tanımları

Madde ve test yanlılığı araştırmaları Alfred Binet ile 1910'lu yıllarda, Binet'in düşük sosyo-ekonomik tabakadan gelen çocukları test etmesiyle başlamış, daha sonraki yıllarda William Stern de Almanya'da sınıf farklılıklarını incelemiş ve insan haklarının gündeme gelmesiyle birlikte, Amerika Birleşik Devletleri'nde bir işe, okula bireyleri yerleştirmede zenci ve beyaz ırkların arasındaki eşitsizlikler, kültürel veya etnik grup yanlılıkları olarak gündeme gelmiş ve bu incelemeler test yanlılığı olarak ele alınmıştır (Camilli ve Shepard, 1994). Test yanlılığı, eğitim kurumlarına öğrencileri seçmede ve kabul etmede, özel eğitim kurumlarına öğrencileri yerleştirmede, eğitimde programların değerlendirilmesinde ve başarı standartlarını oluşturmada, işin gereklerine uygun personel seçimi ile kariyer planlamasında sıklıkla incelenmiştir. Yanlılık gösteren bir test, bu tür durumlarda kullanıldığı zaman bazı insanların lehinde işlev göstererek büyük oranda seçilmelerini sağlarken, bazı insanların da aleyhinde işlev göstererek seçilme oranlarını azaltacaktır. Bireylere eşit seçilme fırsatını sağlayamayan bir test, insan haklarına uygunsuzluğu nedeniyle kaygı yaratıcı olmasının yanı sıra toplumun bu tür ölçüm araçlarının kullanımına ilişkin güvenirlik ve geçerlik algılarını da zedelemiş olur. Psikolojik ölçme araçlarına yapılacak bu gibi olumsuz atıfları engellemenin yolu, test yapımcıları, yayıncıları ve uygulayıcılarının testin bir gruba karşı üstünlük sağlamadığına ilişkin kanıtları sunmalarından geçer (Hambleton ve ark., 1991; McAllister, 1993).

Herhangi bir psikolojik ölçme işlemi yaparken, test ya da ölçekten elde edilen puanların, test ile ölçülen psikolojik özellikten başka varyans kaynaklarından etkilenmemesi olanaksızdır. Azınlık-çoğunluk, zenci-beyaz grupları gibi etnik özellikler, kırsal-kentsel köken gibi farklı yaşam bölgelerinde bulunma, konuşulan dildeki farklılıklar, kadın-erkek olma gibi cinsiyet koşulları gerçekte ölçümlerin alındığı örneklemlerin sistematik özellikleridir. Eğer ölçme aracının, bu sözü edilen karşılaştırma gruplarından herhangi birine avantaj sağlayıp sağlamadığı ispatlanmamışsa ve katılımcıların ait oldukları alt grup/örneklem özelliklerinden dolayı test sonuçlarında, gruplar arası karşılaştırmalarda farklılıklar oluşuyorsa, ölçme sürecine karışan sistematik hataların olabileceği, daha açık bir ifadeyle testin yanlı olabileceği ileri sürülebilir. Dolayısıyla, test yanlılığı tesadüfi ölçme hatalarından kaynaklanmayan, belirli bir örneklemin gözlenen puanlarına etki eden, geçerli olmayan, sistematik hatalar olarak tanımlanmaktadır (Camilli ve Shepard, 1994).

Literatürde test ve madde yanlılığının araştırılmasında dış ölçüt yöntemi (external methods) ve iç ölçüt yöntemi (internal methods) olmak üzere iki temel istatistiksel yaklaşımdan söz edilmektedir. Dış ölçüte dayalı yanlılık incelemeleri geleneksel yöntemler (tek-grup geçerliği, ayırt edici geçerlik hipotezi, regresyon modeli vb.) içerisinde, test yanlılığı olarak adlandırılır ve testin tek tek maddelerinden ziyade toplam test puanı düzeyindeki incelemeleri kapsar (Hulin ve ark., 1983). Dış ölçüt test

yanlılık yöntemlerinden sıklıkla regresyon modeli kullanılarak, genellikle azınlık ve çoğunluk gruplarının psikolojik özellik puanlarından gerçek puanlarının yordandığı regresyon eğrilerine dayanarak değerlendirme yapılır. Regresyon modelinde alt grupların test puanlarından yordanan kriter üzerindeki regresyon eğrisi her iki alt grupta da özdeşse, yani tek bir regresyon eğrisi bulunuyorsa, bu durumda testin yansız olduğu sonucuna varılır (Camilli ve Shepard, 1994; Crocker ve Algina, 1986).

İç ölçüt yanlılık incelemeleri de literatürde madde yanlılığı olarak bilinmektedir. Testin içsel yanlılık incelemeleri, bir dış ölçüt bulunamadığı durumda tüm test ile testin her bir maddesinin arasındaki yapı geçerliği ilişkilerini ve madde analizi incelemelerini kapsar. İç ölçüt madde yanlılık incelemelerinin en temel amacı, herhangi bir dış ölçüt alınmaksızın, aynı evrenden seçilen farklı alt evrenlerde maddelerin aynı tarzda işlevinin olup olmadığını incelemektir (Hulin ve ark., 1983). Bu bağlamda, test amacıyla ilişkili olmayan ölçme koşulları ya da test maddesinin bir takım özellikleri nedeniyle, farklı alt gruplardaki katılımcıların ölçülen özellikte aynı düzeyde olmalarına rağmen maddeyi yanıtlama olasılıkları farklılaşıyorsa, bu maddenin yanlılık gösterdiği söylenir. İç ölçüte göre madde yanlılığını incelemenin diğer bir amacı da, dış ölçüte göre yapılan gerçek grup farklılıkları ile ölçmedeki yanlılık arasındaki ayrımı yapmaktır. Klasik test kuramı kapsamında yapılan madde–toplam puan korelasyonları ya da test maddesi ile toplam test puanlarının karşılaştırıldığı varyans analizi gibi madde analiz yöntemleri veya grupların ortalamaları arasındaki farklılıklar, kesin olarak madde yanlılığının bir kanıtını sağlamazlar (Camilli ve Shepard, 1994). Bu tür yöntemlere göre gerçekleştirilen madde yanlılık incelemeleri kusurludur; çünkü madde güçlük indeksleri ve testin toplam puanları örneklemin yetenek, yani psikolojik özellik dağılımından etkilenmektedir ve grup farklılıklarını oluşturan puan farklılıklarının bilgi ya da deneyim gibi bazı tesadüfi hatalardan kaynaklanması da olasıdır (Maller, 2001). Dolayısıyla, alt grupların test maddelerindeki (veya tek bir maddedeki) ortalamalarının farklılık göstermesi, doğrudan karşılaştırma grupları arasında bir yanlılık delili olarak yorumlanmamalıdır.

Madde cevap kuramı, test maddelerini bir iç ölçüt olarak değerlendirir ve karşılaştırma grupları arasında madde parametrelerinin özellikleri hakkında detaylı bilgiler sağlar. Geçmiş yıllarda madde yanlılığı kavramsallaştırmasının yerini günümüzde madde işlevsel farklılık (Differential Item Functioning-DIF) terimi almıştır. Bu terim özellikle hem madde yanlılık hem de madde-test etkisi (item-test impact) teriminden ayırtedici olması açısından tercih edilmektedir. Burada madde işlevsel farklılık ve madde etkisi arasındaki ayrımı yapmak son derece önemlidir. Madde ya da test etkisi, hem test hem de madde düzeyinde inceleme altına alınan alt grupların etnik köken, cinsiyet vb. değişkenlerde, psikolojik özellik dağılımları arasındaki "performans farklılıklarını" açıklamaktadır (Raju, Laffitte ve Byrne, 2002). Madde işlevsel farklılık terimi ise; etnik grup, cinsiyet vb. iki farklı alt grubun ya da örneklemin ölçme yapılan psikolojik özellik üzerinde eşleştirildikten sonra "madde işlevinde" (madde ile ölçülen özellik arasındaki ilişkide) ortaya çıkan farklılıkları açıklamaktadır (Camilli ve Shepard, 1994; Ra-

ju ve ark., 2002). Daha önce de belirtildiği gibi, madde yanlılığı, bir test maddesindeki azınlık ve çoğunluk gruplarının performans ortalamalarının farklılaşması olarak tanımlanması, psikometrisler tarafından eksik bir tanımlama olarak görülmüştür. Zira bu tür bir performans farklılığı sadece madde yanlılığından değil, gerçek bir farklılıktan da kaynaklanabilir. Psikometrisler tarafından kabul gören madde işlevsel farklılığı tanımı; farklı gruplarından gelen katılımcıların aynı yetenek ya da psikolojik özellik düzeyinde (eşleştirme yapıldıktan sonra), maddeyi doğru yanıtlama/onaylama olasılıklarının farklılık göstermesidir (Hambleton ve ark., 1991). Literatürde madde işlevsel farklılığı için "madde performans farklılığı (differential item performance)" ya da "beklenilmeyen madde işlevsel farklılığı (unexpected diffrential item functioning)" gibi kullanımlarına da rastlanmaktadır (Thissen, Steinberg ve Wainer, 1988).

Geleneksel tek boyutlu madde cevap kuramında (son yıllarda çok boyutlu özellikler ile de çalışılmaktadır), testin gözlenemeyen tek bir örtük değişkeni, (θ'yı) ölçtüğü varsayılır. Kurama göre, bir maddeye verilen doğru yanıtın olasılığı, belli bir yetenek düzeyinde θ'nın bir fonksiyonu olarak madde karakteristik eğrisiyle tanımlandığına göre; madde her iki grupta aynı karakteristik eğriye sahip olduğunda, bu maddenin iki grupta aynı ya da benzer bir işlevi vardır. Ancak, madde her iki grupta aynı madde karakteristik eğrisine sahip değilse, bu durumda maddenin gruplar arasında işlevsel farklılık gösterdiği kabul edilir (Thissen ve ark., 1988; Kim, Cohen ve Kim, 1994; Robie, Zickar ve Schmit, 2001). İkili puanlanan (dikotomik) madde cevap modelleri bağlamında Pine (1977), doğru cevabın olasılığının, *aynı yetenek düzeyinde*, ancak farklı grup üyelikleri olan katılımcılara değişiklik gösterdiğinde maddenin farklı bir biçimde fonksiyon gösterdiğini belirtmiştir (akt., Kim ve ark., 1994, Teresi, 2000). Genellikle Likert tipi puanlanan kişilik ve tutum ölçeklerinde çoklu kategorili maddeler söz konusu olduğunda da, bir maddenin kategori cevap eğrileri iki alt grupta çakışmıyorsa, diğer bir ifadeyle maddeyi anahtarlanan yönde işaretleme olasılıkları farklılaşıyorsa, madde işlevsel farklılığının varlığından söz edilir (Lim ve Drasgow, 1990; Reise, 1999; Reise, Smith ve Furr, 2001; Somer, 2004)

Buraya kadar sözü edilen madde işlevsel farklılık tanımlarından da anlaşılacağı üzere, madde işlevsel farklılık incelemelerinde madde parametrelerinin iki alt grupta tahmin edilmesi gerekmektedir. Madde cevap kuramında bu gruplar, referans ve fokal grup olarak adlandırılmaktadır. Madde cevap kuramı dışındaki yanlılık incelemelerinde bu gruplardan, çoğunluk ve azınlık grupları olarak bahsedilmektedir. Madde cevap kuramında referans grubu genellikle çoğunluk grubunu ve fokal grupta azınlık grubunu temsil etmekte olup referans grup, fokal grupta tahminlenen madde parametrelerinin karşılaştırılacağı ana grup olarak ele alınır (Cohen, Kim ve Baker, 1993; Kim ve Cohen, 1991, 1998).

Bu açıklamaların ışığında, madde işlevsel farklılık tanımında dikkat edilmesi gereken en önemli nokta, maddeyi doğru yanıtlama olasılığının iki alt grupta farklı olması anlamını taşımadığıdır. Madde işlevsel farklılığı, örtük özellik üzerinde yani θ'nın herhangi bir noktasında bulunan fokal grubun bir üyesinin, ***aynı θ düzeyinde*** bulunan referans gruptaki bir ka-

tılımcıdan doğru yanıtlama olasılığının farklı olması anlamına gelmektedir (Thissen ve ark., 1988).

Madde cevap modellerinin önemli özelliklerinden biri de, başlangıcından günümüze kadar sosyal bilimlerde davranış özelliklerinin, farklı alt evrenlerde olduğu kadar farklı kültürler arasında da ölçme eşdeğerliğinin incelemesine olanak tanımasıdır (Collins, Raju ve Edwards, 2000; Glöckner-Rist ve Hoijtink, 2003). Dolayısıyla, kültürler arası karşılaştırmalar söz konusu olduğunda, iki kültürün aynı madde üzerindeki madde karakteristik eğrileri örnekleme (kültür) hatalarından farklılaşıyorsa, ölçme ya da metrik eşdeğerliğinin olmadığı sonucuna varılır ve söz konusu maddenin kültürler arasında işlevsel farklılık gösterdiği ifade edilir (Huang, Church ve Katigbak, 1997). Benzer olarak Raju ve arkadaşları (2002), bir maddenin işlevsel farklılık göstermemesi için, madde parametre değerlerinin iki farklı evrende aynı olması gerektiğini belirtmişlerdir.

Yansız bir ölçme işlemi gerçekleştirmek tüm test/ölçek geliştirme süreçlerinin en önemli hedeflerden biridir. Her ölçme aracı belirli bir amaç ile ölçmeyi hedeflediği özellik/ özellikler üzerine temellenir. Madde işlevsel farklılığı, farklı evren veya grup üyelerinin yetenek, yeterlik, tutum, kişilik gibi özelliklerini ölçmek üzere kullanılan bir testin geçerliğini ciddi anlamda tehdit eder. Bazı test maddeleri bir örneklem grubundaki katılımcılar için diğer örneklem grubundaki katılımcılara nazaran farklı bir biçimde işlev gösterebilir ya da bir grubun üyeleri için diğer grubun aksine farklı bir şeyi ölçebilir. Bu tür maddeleri içeren testler grup karşılaştırmalarında geçerliği azaltacaktır; çünkü bu testlerden elde edilen puan farklılıkları, testin ölçmeyi amaçladığı özellikten daha ziyade, başka özelliklerdeki değişkenliği de ortaya çıkarabilir. Dolayısıyla bir test, kişilerin bağlı bulunduğu herhangi bir demografik grup üyeliğinin etkisi olmaksızın ölçülmesi amaçlanan yeteneği, özelliği doğru biçimde ölçmelidir.

Madde Cevap Kuramı Kapsamında Madde ve Test İşlevsel Farklılık Yöntemleri

Madde cevap kuramı kapsamında, özellikle madde düzeyinde çeşitli işlevsel farklılık yöntemleri bulunmaktadır. Bunlar; temel alan indeksleri (simple area indices), olasılık fark indeksleri (probability difference indices), madde güçlük parametreleri farkı (*b* parameter difference), küçük örneklemlerin madde karakteristik eğrisi yöntemi (ICC method for small samples), madde güçlük farkı testi (test of *b* difference), madde eğilim modeli (item drift model), Lord'un χ^2 testi (Lord's Chi-Square), madde işlevsel farklılığı için görgül örnekleme dağılımları (emprical sampling distributions for DIF indices) ve model karşılaştırma ölçümleri (model comparison measures)'dir (Camilli ve Shepard, 1994). Literatürde yapılan araştırmalar dikkate alındığında ise, en sık kullanılan üç farklı yöntemden söz edilmektedir (Cohen ve ark., 1993; Kim ve Cohen, 1995, 1998; Raju, Van der Linden ve Fleer, 1995; Teresi, 2000). Bu yöntemler;

i. iki alt grupta tahminlenen madde parametrelerinin karşılaştırılması,

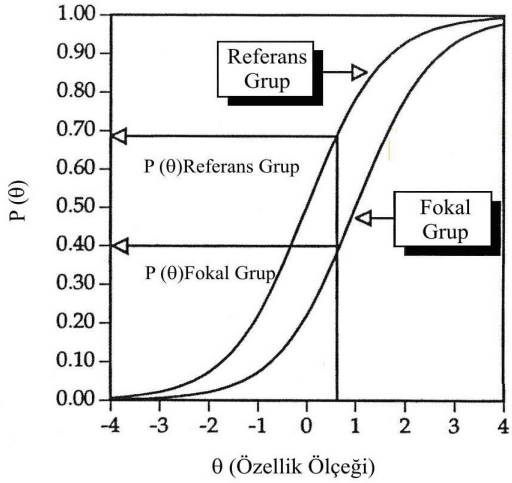
ii. iki alt grubun madde karakteristik eğrileri arasında kalan alan ölçümlerinin karşılaştırılması,

iii. iki alt grupta madde cevaplarının model-veri uyumunu değerlendirerek aynı zamanda iki gruptan gelen madde cevapları arasındaki olasılık fonksiyonlarının karşılaştırılmasıdır.

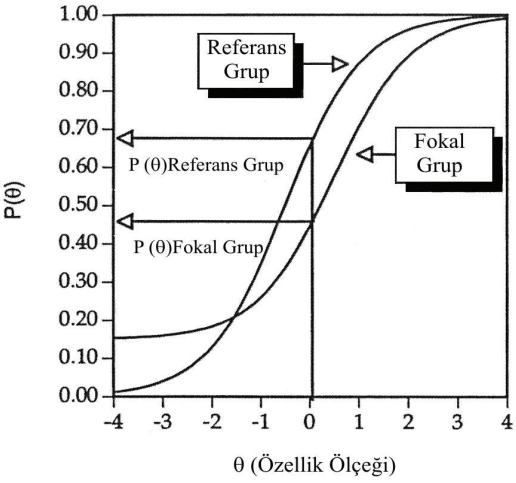
Madde işlevsel farklılıklarının incelenmesini öneren yaklaşımların temel amacı, gruplara ait madde karakteristik eğrileri arasındaki farklılığın anlamlılığını test etmektir. Madde karakteristik eğrileri referans ve fokal grupların cevaplarını karşılaştırmayı sağladığına göre; iki grubun madde karakteristik eğrileri arasındaki fark, örtük özellik üzerinde aynı θ düzeyinde bulunan referans grup ve fokal grup katılımcılarının maddeye aynı derecede doğru yanıt verme olasılıklarının olup olmadığını göstermektedir. Daha teknik bir ifadeyle söylemek gerekirse, madde işlevsel farklılığı, iki grupta doğru cevabın koşulsal (conditional) olasılığının (P(θ)) farklılık gösterdiği anlamını taşır (Camilli ve Shepard, 1994). Buradaki "koşul" "*P*" değerleri karşılaştırılırken iki örneklemdeki katılımcının aynı yetenek, tutum vb. θ düzeyinde olmasını temsil etmektedir.

Madde işlevsel farklılığında, madde karakteristik eğrileri, düzgün formlu (uniform) ve düzgün formlu olmayan (nonuniform) eğriler olarak adlandırılmaktadırlar (Şekil 1-2).

Düzgün formlu madde işlevsel farklılık fonksiyonu (Şekil 1), maddenin örtük özellikle ilişkisinin referans ve fokal her iki grupta da aynı yönde olduğunu, dolayısıyla grupların madde karakteristik eğrilerinin kesişmediğini, ancak maddenin gruplardan biri için daha zor iken diğeri için daha kolay olduğu ve gruplardan biri için göreceli bir avantaj sağladığını göstermektedir (Camilli ve Shepard, 1994;



Şekil 1. Düzgün Formlu (Uniform) Madde İşlevsel Farklılık Fonksiyonları (Camilli ve Shepard, 1994).



Şekil 2. Düzgün Formlu Olmayan (Nonuniform) Madde İşlevsel Farklılık Fonksiyonları (Camilli ve Shepard, 1994)

Hanson, 1998; Reise ve ark., 2001; Smith, 2002). Burada fokal ve referans grupların "madde ayırt etme" parametreleri farklılaşmamakta, ancak "madde güçlük" parametreleri farklılık göstermekte ve dolayısıyla maddenin örtük özellik (θ) ölçeği üzerindeki konumu değişmektedir. Örneğin; fokal grubun kadın, referans grubun erkekleri temsil ettiği ve birey-

lerin bir zeka testinin genel bilgi alt testinden bir maddeyi temsil ettiğini düşündüğümüz bir araştırmada, bu test maddesinin *aynı θ düzeyinde* bulunan kadın ve erkelerden, erkekler için daha kolay olduğunu, yani "doğru" yanıt verme olasılıklarının erkekler için kadınlardan yüksek olduğunu göstermektedir. Dolayısıyla bu test maddesi, erkek katılımcıların lehinde işleyen ve yanlılık gösteren bir maddedir. Düzgün formlu olmayan madde işlevsel farklılığı fonksiyonunda ise (Şekil 2), referans ve fokal grupların madde karakteristik eğrileri birbirlerinden hem farklıdırlar, hem de θ ölçeğinin bazı noktalarında kesişmektedirler. Burada madde ile ölçülen özellik arasındaki ilişki, bir grupta (örn., referans) diğer gruba (örn., fokal) nazaran daha güçlüdür, çünkü maddenin hem ayırt edicilik hem de güçlük parametreleri gruplar arasında farklılık göstermektedir (Camilli ve Shepard, 1994; Orlando ve Marshall, 2002; Smith, 2002; Van de Vijyer ve Leung, 1997). Yukarıdaki aynı örneği düzgün formlu olmayan madde işlevsel fonksiyon eğrisi açısından değerlendirdiğimizde, kadınların θ ölçeği üzerinde belli bir düzeye kadar (θ = -2.00 düzeylerinde) maddeyi "doğru" yanıtlama olasılıkları daha yüksektir ve kadınlar için daha kolay bir madde olarak görülmektedir. Fakat belirli bir düzeyden sonra kadınlar için aynı madde daha "güç" bir madde olurken, erkekler için ise daha "kolay" bir madde haline gelmiştir. Aynı zamanda Şekil 2, düşük yetenekli kadın katılımcıların şansla doğru cevabı tahmin parametrelerinde de erkelerden farklılaştıklarını da göstermektedir.

Bu yazıda, son yıllarda madde cevap kuramına dayalı olarak madde işlevsel farklılığının incelenmesinde sıklıkla kullanılan üç yöntem ve özellikleri üzerinde durulacaktır. Bunlar; (1) **olabilirlik oranı testine dayalı model karşılaştırma yöntemi,** (2) **madde-test işlevsel farklılık yöntemleri**, (3) **parametre karşılaştırma yöntemi**' dir.

Olabilirlik Oranı Testine Dayalı Model Karşılaştırma Yöntemi

Thissen, Steinberg ve Wainer (1993) tarafından geliştirilen model, iki modelin göreceli uyumunu karşılaştırmayı içerir ve madde işlevsel farklılığının istatistiksel anlamlılığını olabilirlik oranı testiyle değerlendirir (Bolt, Hare, Vitale ve Newman, 2004; Cooke ve Michie, 1999; Cooke, Michie, Hart ve Hare, 1999; Huang ve ark., 1997; Kim ve Cohen, 1995, 1998; Lambert ve ark., 2003; Maller, 2001; Orlando ve Marshall, 2002; Reise, Wideman ve Pugh, 1993; Sireci ve Berberoğlu, 2000). Yöntemdeki karşılaştırılacak modellerden ilki daraltılmış/sınırlanmış (compact model)[1] ikincisi ise genişletilmiş/sınırlanmamış (augmented model)[2] olarak adlandırılmaktadır. Geniş model[3], dar modelin[4] büyütülmüş halidir, yani geniş model dar modelin tüm parametrelerinin dışında eklenen başka parametre setini de içermektedir ve modeller birbiri içine yuvalanmıştır (nested). Bu iki modeli karşılaştırmaktaki amaç, geniş modele eklenen parametrelere gerçekten ihtiyaç olup olmadığını ve bu ilave parametrelerin modelin uyum derecesini arttırmada "sıfır"dan anlamlı bir şekilde

[1]literatürde Compact model; simpler model, baseline model, constrained model biçiminde de isimlendirilmiştir.
[2]Literatürde Augmented model; complex model, unconstrianed model olarak da bilinmektedir.
[3]Bu çalışmada augmented model yerine "geniş model",
[4]Compact model yerine" dar model" tanımı kullanılacaktır.

farklılaşıp farklılaşmadığını test etmektir. Diğer bir ifadeyle, model uyumunda anlamlı bir ilerleme/farklılaşma yoksa dar modelin referans ve fokal grupların tek bir madde karakteristik eğrisi ile sunulmasının (yani grruplar arasında işlevsel farklılaşma olmadığının varsayılması), daha karmaşık olan geniş modeldeki her iki gruba ait madde karakteristik eğrileri ile sunulmasından daha iyi olduğu kabul edilir. Buna karşın, geniş model verilere daha iyi ve anlamlı uyum gösteriyorsa, bu modele ilave edilen parametrelerin gerekli olduğu kabul edilir (Camilli ve Shepard, 1994).

İki modeli karşılaştırmak için bir modelin ne kadar iyi temsil edildiğinin ya da verilerle ne kadar iyi uyum gösterdiğinin ölçülmesi gerekir. Literatürde sıklıkla karşılaşılan "en iyi uyum istatistiği (goodness of-fit statistics)" Ki-Kare ile test edilmektedir. Madde cevap modellerinde parametre tahmini genellikle maksimum olabilirlik tahmin (Maximum Likelihood Estimation-MLE) yöntemiyle yapılır. Maksimum olabilirlik tahmini, gözlenen parametrelerden evren parametreleri yordamak istendiğinde, evren parametre değerlerini maksimize eden, yani evren değerlerine en yakın değeri sağlayan bir tahmin yöntemidir (Camilli ve Shepard, 1994; Hayduk, 1987; Meyer, 1970). Bu yordama süreciyle maddeler ve katılımcılar, gerçek performans olasılıkları ile beklenen olasılık parametreleri arasındaki ilişkinin en yakın olmasını sağlayacak şekilde θ ölçeği üzerinde yerleştirilmektedirler ve bu işlem gerçek test verileriyle madde ve yetenek tahminleri arasında en yüksek uyum sağlanıncaya kadar devam etmektedir (Somer, 1998). Maksimum olabilirlik, iki model arasındaki "en iyi uyum" un olup olmadığını istatistiksel olarak test etmede kullanılan "olabilirlik oranı test"inin kullanımını içerir. Çok daha sade olan dar modelin uyum iyiliği, göreceli olarak daha karmaşık oluşturulan geniş modelle karşılaştırılır.

Olabilirlik oranı testi kavramını şu şekilde açıklayabiliriz: H_0 hipotezinin dar modeli, yani yalnızca N sayıda madde parametresini içerdiğini ve H_1 hipotezinin de geniş modeli, yani N parametrelerine ilave olarak M parametrelerini de içerdiğini varsaydığımızda; dar model daha sade bir model olarak geniş modelden daha az sayıda parametreye sahip olmakta ve iki model için olabilirlik oranı şu şekilde tanımlanmaktadır (Camilli ve Shepard, 1994).

$$\text{Olabilirlik oranı (LR)} = \frac{L^*(\text{Dar Model})}{L^*(\text{Geniş Model})}$$

Burada şu soruya yanıt aranmaktadır: Örneklem verileri H_0 hipotezini desteklemekte midir? Diğer bir deyişle, ilave edilen M parametrelerine gerek var mıdır? Olabilirlik oranı doğrudan bu soru ile ilişkilidir. Olabilirlik oranı, doğal -2 defa logaritmik dönüşümü alındığında M serbestlik derecesinde χ^2'ye yaklaşık dağılım gösteren bir test istatistiğidir ve özellikle geniş örneklem büyüklüklerinde χ^2 dağılımı göstermektedir.

$$\chi^2(\mathrm{M}) \approx -2\ln(LR) = [-2\ln L^*(Dar\ Model)] - [-2\ln L^*(Geniş\ Model)]$$

Kısaca olabilirlik oranı testi model karşılaştırma yöntemi, dar ve geniş modellerin -2 defa logaritmik dönüşümünden elde edilen değerlerin birbirlerinden çıkarıldığında elde edilen ilerlemeyi ifade etmektedir. Bu ilerleme istatistiksel olarak anlamlı ise, geniş modelin gerekliliğine işaret eder, yani madde parametlerinin iki grupta farklılaştığını gösterir.

Madde işlevsel farklılığını test etmek üzere olabilirlik oranı testi kullanıldığı zaman iki modelin karşılaştırma yöntemi şu şekilde yapılır. Dar model, referans ve fokal gruplarda tüm madde parametreleri sınırlanarak, yani her iki grupta madde parametreleri eşitlenerek oluşturulur (Bolt ve ark., 2004, Meade ve Lautenschlager, 2004a). Bu model aynı zamanda madde işlevsel farkının yokluk hipotezini sınayan, diğer bir ifadeyle hiçbir maddenin işlevsel farklılık göstermeyeceğine ilişkin oluşturulan hipotezi test eder ve testin bağ (anchor)[5] maddeleri belirtilmemişse tüm test maddeleri kullanılır. Tek bir kalibrasyon çalıştırıldığında modelin logaritmik olabilirlik değeri elde edilir (Korkmaz, 2005). Daha sonraki aşamada, testteki her bir madde sırasıyla madde işlevsel farklılığının olduğu varsayılan geniş model ile test edilir. Geniş modelde, madde işlevsel farklılığı için incelenecek olan maddenin bir ya da daha fazla sayıda parametresi (a_i, b_i ya da c_i) fokal ve referans gruplarda serbest bırakılır (Kim ve Cohen, 1995). Bu testteki tüm madde parametrelerinin logaritmik olabilirlik değerlerinin, incelemeye alınan "i" maddesi dışında sınırlandırıldığını ifade eder (Thissen, 2001). Geniş modelin olabilirlik değerleri, her bir madde için ayrı ayrı yapılan kalibrasyonlar sonucunda elde edilir. Sonuçta "k" sayıda madde içeren bir test için "bir" tane dar model ve k+1 sayıda kalibrasyon yapmayı gerektiren k sayıda geniş model bulunur (Camilli ve Shepard, 1994). Bu iki modelin karşılaştırması şu şekilde formüle edebilir:

$$G^2 (s.d) = -2 (\text{logolabilirlik}_{\text{tüm maddeler eşit}} - \text{logolabilirlik}_{\text{madde1eşit değil}})$$

Madde işlevsel farklılığının anlamlılık düzeyini test etmede kullanılan serbestlik derecesi bir, iki ve üç parametreli modeller için modelde bulunan parametre sayısına eşittir. Model karşılaştırma yaklaşımıyla madde işlevsel farklılığı incelenirken bağ maddeleri test deseni kullanılır. Bu yöntemde test maddeleri "bağ" ve "incelenen" madde seti olarak ikiye ayrılır. İncelenen madde seti halihazırda incelenen maddeyi kapsarken, bağ madde seti referans ve fokal gruplar için önceden çeşitli yöntemler kullanılarak madde işlev farklılığının olmadığına karar verilen referans alınacak maddeleri içerir (Camilli ve Shepard, 1994; Korkmaz, 2005).

Madde-Test İşlevsel Farklılık Yöntemi

Raju ve arkadaşları (1995), madde cevap kuramı çerçevesinde işlevsel farklılığı hem test hem de madde düzeyinde ölçen madde-test işlevsel farklılık yöntemini önermişlerdir. Bu yöntemde üç işlevsel farklılık indeksi önerilmiştir (Collins ve ark, 2000.; Flowers, Oshima ve Raju, 1999; Maurer, Raju ve Collins, 1998; Oshima, Raju, Flowers ve Slinde, 1998; Raju ve ark., 1995).

i- Test işlevsel farklılığı (Differential Test Functioning-DTF),

ii- Telafi edici madde işlevsel farklılığı (Compensatory Differential Item Functioning-CDIF),

iii- Telafi edici olmayan madde işlevsel farklılığı (Noncompensatory Differential Item Functioning-NCDIF).

[5]Anchor: Karşılaştırma içeren durumlarda bir testin ya da test içindeki bazı maddelerin sabit tutularak referans ya da kaynak olarak kullanılması. Türkçe'de tam karşılığının olmamasına rağmen "anchor item" yerine "bağ maddesi" ifadesi kullanılmıştır.

Test İşlevsel Farklılık İndeksi

Madde test işlevsel farklılık yönteminde test uygulanan grup, referans ve fokal gruplar olmak üzere iki bağlamda düşünülmektedir. Dolayısıyla, her bir katılımcının, referans grubun bir üyesi olarak bir gerçek puanı (T_{sR}), fokal grubun bir üyesi olarak da diğer bir gerçek puanı (T_{sF}) olmak üzere iki gerçek puanı vardır. İki gerçek puan birbirine eşit ($T_{sF} = T_{sR}$) olduğunda katılımcının gerçek puanı da grup üyeliğine bağlı olmaktan çıkar. İki gerçek puan arasındaki farkın büyük olması, test işlevsel farklılığının da fazla olduğunu gösterir. Tek bir katılımcı düzeyinde test işlevsel farklılık ölçümü $(T_{sF} - T_{sR})^2$ olarak tanımlanmaktadır ve tüm katılımcılar üzerinden test işlevsel farklılığının ölçümü de aşağıdaki gibi tanımlanmaktadır (Raju ve ark., 1995; Collins ve ark., 2000).

$$DTF = \varepsilon\,(T_{sF} - T_{sR})^2$$

Bu eşitlikteki (ε), beklenilen taraf referans ya da fokal gruplardan herhangi birisi için ele alınabilir. Örneğin, fokal grup için ele aldığımızda eşitlik şöyle olur:

$$DTF = \varepsilon_F\,(T_{sF} - T_{sR})^2$$

İki gerçek puan arasındaki farkı "D" ile tanımladığımızda DTF denklemi şu halini alır.

$$DTF = \varepsilon_S D_S^2 = \int_{\Theta} D_S^2 f_F(\Theta)\,d\Theta = \sigma_D^2 + (\mu_{TF} - \mu_{TR})^2 = \sigma_D^2 + \mu_D^2,$$

Bu denklemde $f_F(\Theta)$, Θ'nın fokal gruptaki yoğunluk fonksiyonunu ve μ_{TF} ve μ_{TR} referans ve fokal gruplardaki katılımcıların beklenen doğru oranı ortalamalarını ifade etmektedir.

Telafi Edici Madde İşlevsel Farklılık İndeksi

Telafi edici madde işlevsel farklılık indeksi, madde düzeyinde bir indekstir ve bir maddenin test işlevsel farklılığına olan net katkısını ifade eder. Bu indeksin en önemli özelliği, testte inceleme altına alınan bir maddenin işlevsel farklılığının, referans ve fokal gruplarda yönünün nasıl olduğuna ilişkin bilgi vermesidir. Bu durum, iki maddenin işlevsel farklılığının benzer düzeylerde olduğu halde birbirine zıt yönlerde (pozitif veya negatif) olabileceği, yani maddelerden birinin fokal grubun lehine destekleyici olurken, diğerinin ise referans grubun lehine destekleyici işlev gösterebileceğini ifade etmektedir (Facteau ve Craig, 2001; Meade ve Lautenschlager, 2004b). Test maddelerinin CDIF indeks değerlerinin pozitif ya da negatif olması ve CDIF indeks değerlerinin toplanabilme özeliğinden dolayı test düzeyinde telafi edici etki, gruplar arasında işlevsel farksızlıkla sonuçlanabilir. Bu durumda madde düzeyinde işlevsel farklılık olmasına rağmen, test düzeyinde işlevsel farklılık görülmeyebilir ve araştırmanın amaçlarına dayalı olarak işlevsel farklılık gösteren maddeleri testtin nihai formundan çıkarmaya gerek kalmayabilir. Telafi edici madde işlevsel farklılık indeksinin, test işlevsel farklılığına katkısı şu şekilde formüle edilmektedir (Raju ve ark., 1995).

$$DTF = \sigma_D^2 + \mu_D^2 = \sum_{i=1}^{n} CDIF_i = \sum_{i=1}^{n}\left[COV(d_i, D) + \mu_{d_i}\mu_D\right],$$

Bu eşitlikte; "n" testteki madde sayısını, "d_i" **i,** maddesinin referans grup madde gerçek puanından çıkarılan fokal grubun madde gerçek puanını, "D" referans grup gerçek toplam puandan çıkarılan fokal grup gerçek toplam puanını ve her iki puan arasındaki kovaryans

değerlerini ifade etmektedir. Bu eşitlik aynı zamanda CDIF indeksinin toplanabilir ($DTF = \sum_{i=1}^{n} CDIF_i$) ve testteki her bir maddenin DTF indeksine katkısı olduğunu göstermektedir. CDIF indeksinin toplanabilme özelliği bir anlamda CDIF gösteren bir maddenin testten ihracının, o maddenin DTF indeksine katkısını tamamen yok etmediğini gösterir (Raju, 2004b).

CDIF indeksi madde işlevsel farklılığın tanımlanmasında ve değerlendirilmesinde dolaylı, ancak önemli bir rol oynamaktadır, indeksin önemini Raju (2004b) şu şekilde sıralamaktadır.

a) CDIF indeksi, bir maddenin testteki diğer maddeler bağlamında DTF indeksine ne kadar katkı sağladığını gösterir. Bu bilgi testteki hangi maddelerin çıkarılacağına ilişkin karar vermede son derece faydalıdır.

b) CDIF indeksi, test düzeyindeki işlevsel farklılığın telafi edici yapısını gösterir. Yani pozitif ve negatif değerdeki CDIF indeksleri, test düzeyinde işlevsel farklılık ile ilişkili olarak nasıl birbirlerinin etkilerini yok ettiklerini açıklamaktadırlar.

c) CDIF indeksi, halihazırda bilinen çoğu DIF indeksinin temelinde bulunan nadiren tartışılan ve ihmal edilen varsayımlara odaklanmayı sağlar. Bu varsayım, testteki diğer maddelerin DIF'ten bağımsız olduğudur.

Telafi Edici Olmayan Madde İşlevsel Farklılık İndeksi

Telafi edici olmayan madde işlevsel farklılık indeksi (NCDIF), inceleme altında bulunan fokal ve referans grupların gerçek puan farklarını yansıtan madde düzeyinde bir indekstir. NCDIF indeksi, fokal grup üyeleri için hesaplanan beklenen puanları, referans grubun üyelerininmiş gibi varsayarak hesaplar. Bu doğrultuda, ilk olarak fokal grubun üyesi olan bir kişinin, o grubun madde parametre tahminleri kullanılarak beklenen puanı hesaplanır. Daha sonra ikinci olarak, fokal gruptaki bu kişinin puanları bu kez referans grubun madde parametre tahminleri kullanılarak beklenen puanları hesaplanır (Meade ve Lautenschlager; 2004b). Sonuçta bu puanlar arasında istatistiksel düzeyde anlamlı olmayan bir fark ($d_i = 0$) varsa madde işlevsel farklılığı olmadığına karar verilir (Collins ve ark., 2000; Facteau ve Craig, 2001; Raju, 2004b). Telafi edici olmayan madde işlevsel farklılık indeksi, testteki diğer maddelerden gelen işlevsel farklılık hakkında bilgi vermez. Testteki tüm maddelerin incelenen madde dışında madde işlevsel farklılık göstermeyeceği varsayımı bulunduğunda, $d_i = 0$ eşitliği doğru olmalıdır. Telafi edici olmayan madde işlevsel farklılık indeksi şu eşitlik ile açıklanmaktadır (Raju ve ark., 1995):

$$NCDIF_i = \sigma^2_{d_i} + \mu^2_{d_i}$$

NCDIF indeksi, telafi edici değildir, çünkü hiçbir zaman negatif değer alamaz ve dolayısıyla testteki diğer maddelerin etkilerini yok edici özellik gösteremez (Raju ve ark., 1995). Raju ve arkadaşları (1995), madde-test işlevsel farklılık (DFIT) yönteminde, hem madde hem de ölçek/test düzeyinde işlevsel farklılığın istatistiksel anlamlılık düzeyinin değerlendirilmesinde, istatistiksel anlamlılık testleri ve kesme puanı ölçütleri önermişlerdir. Kesme puanı özellikle geniş örneklem büyüklüklerinde test işlevsel farklılığı ve madde işlevsel farklılığı sonuçlarının çok fazla istatistiksel anlamlılık verebileceği durumlarda uygulamacılara faydalı bilgiler sağlar.

Madde ve test işlevsel farklılığı (DFIT) kapsamında elde edilen DTF, CDIF ve NCDIF indekslerinin her biri pratik uygulamalarda kullanılma amacına bağlı olarak araştırmacılara oldukça önemli bilgiler sunmaktadırlar. Psikolojik bir ölçme aracından sağlanan toplam test puanları "seçme ve yerleştirme" amaçları doğrultusunda kullanıldığında, test işlev farklılığı (DTF) indeksi daha yararlı olur. Ölçme aracındaki bazı maddeler fokal grubun lehinde ve diğer bazı maddeler de referans grubun lehinde anlamlı işlevsel farklılıklar gösteriyorlarsa, telafi edici madde işlevsel farklılık (CDIF) indeks değerlerinin incelenmesi önem taşır. Daha önce de belirtildiği gibi, anlamlı madde işlevsel farklılık gösteren tüm maddeler testin nihai formundan her zaman çıkarılmazlar ve özellikle de DTF indeksi bu tür maddeler üzerinde bir denkleştirme etkisine sahip olduğu için madde cevap kuramı kapsamında kullanılan diğer madde işlevsel farklılık indekslerine göre bir üstünlüğe sahiptir (Raju ve ark., 1995). Testteki maddelerin karşılaştırma yapılan gruplardan herhangi biri için olumsuz etkilerinin varlığını, yani işlevsel farklılık oluşturduğunu görmek için, telafi edici olmayan madde işlevsel farklılık (NCDIF) indeksini kullanmak faydalı olur.

Özetle; DFIT'te, referans ve fokal grupların madde parametre tahminleri ve örtük özellik (θ) puanları kullanılarak, aynı zamanda referans ve fokal grupların parametreleri birleştirme katsayıları ile ortak bir metrik üzerinde eşitlendikten sonra, hesaplanan madde karakteristik eğrilerinin karşılaştırılmasıyla, madde ve test işlevsel farklılığı saptanmaktadır (Stark, Chernyshenko, Chan, Lee ve Drasgow, 2001; Korkmaz, 2005).

Madde Parametrelerini Karşılaştırma Yöntemi

Madde parametrelerini karşılaştırma yönteminde madde işlevsel farklılık incelemeleri, madde ayırt edicilik ile madde güçlük/yerleşim parametrelerinin referans ve fokal gruplar arasındaki karşılaştırmalara (fark/contrast) temellenmektedir. Muraki ve Bock (1996) tarafından hazırlanan PARSCALE programı, karşılaştırma yapılacak gruplarda madde parametrelerinin kolay bir biçimde sınırlandırılmasına ya da serbest bırakılmasına ve böylece parametrelerin gösterdikleri farklılıkların ayrı ayrı test edilmesine olanak tanımaktadır. Madde düzeyinde, madde işlevsel farklılık incelemelerinde her bir madde için öncelikle referans ve fokal gruplarda madde ayırt etme ve madde güçlük/yerleşim işlevsel farklılık istatistik değeri şu şekilde hesaplanır (Morales, Reise ve Hays, 2000; Reise ve ark., 2001):

Madde işlev farklılığı = $\hat{a}_{i(referans)} - \hat{a}_{i(fokal)}$

Madde işlev farklılığı = $\hat{a}_{i(referans)} - \hat{a}_{i(fokal)}$

Bu eşitliklerde, ölçme aracında bulunan ve inceleme altına alınan bir maddenin referans gruptaki ayırt etme parametre tahmin değerinden, fokal gruplardaki parametre tahmin değerinin çıkarılmasıyla, madde işlevsel farklılığının doğrudan bir ölçümü elde edilir. Daha sonraki aşamada elde edilen bu değerin standardize hale getirilmesi gerekir. Ayırt etme parametresi için "standardize madde işlevsel farklılığı" aşağıdaki eşitlik ile sağlanır (Morales ve ark., 2000; Reise ve ark., 2001).

Standardize madde işlev farklılığı (SDIF) =

$$\frac{DIF}{\sqrt{\text{var}\,\hat{a}_{i(\text{referans})} + \text{var}\,\hat{a}_{i(\text{fokal})}}}$$

Madde yerleşim parametresi için "standardize madde işlevsel farklılık" değeri de aşağıda sunulan eşitlik ile sağlanır (Morales ve ark., 2000; Reise ve ark., 2002; Smith, 2002).

Standardize madde işlev farklılığı (SDIF) =

$$\frac{DIF}{\sqrt{\text{var}\,\hat{b}_{i(\text{referans})} + \text{var}\,\hat{b}_{i(\text{fokal})}}}$$

Standardize madde işlevsel farklılık istatistiği z ya da T-puanlarına benzer ve referans ile fokal grupların madde parametre tahminleri arasındaki farkın (contrast) bir ölçümünü verir (Smith, 2002). Örneğin, referans grubun kadın ve fokal grubun erkek olduğu gruplararası bir karşılaştırmada, bir maddenin standardize madde işlevsel farklılık değeri (SDIF) pozitif (+) olduğunda, bu madde fokal (erkek) grup üyeleri için kolay olduğunu ve standardize madde işlevsel farklılık (SDIF) değeri negatif (-) olduğunda da referans (kadın) grup üyeleri için daha kolay olduğunu göstermektedir (Smith, 2002). Thissen, Steinberg ve Wainer (1993) standardize madde işlevsel farklılık değerinin karesi alındığında "1" serbestlik derecesinde χ^2 istatistiği olarak değerlendirilebileceğini önermektedirler (akt., Reise ve ark., 2001; Smith, 2002). Bu ölçüte göre, eğer bir madde 0.01 ya da 0.05 nominal alfa düzeylerinde anlamlı bir χ^2 değerine sahipse, bu durumda maddenin gruplar arasında işlevsel farklılık gösterdiğine karar verilir. Hangi nominal alfa değerinin alınacağı ise, araştırmanın amacına ya da araştırmacıya bağlıdır.

Sonuç

Psikologların çoğu araştırma ve uygulamalarında test ve ölçekleri değerlendirmek üzere klasik test kuramı tekniklerini (güvenirlik, madde-toplam puan korelasyonları, Spearman Brown düzeltme formülasyonu v.b.) sıklıkla kullanmaktadırlar. Klasik test kuramının varsayımlarını karşılamak daha kolaydır ve bu durum kuramın uygulanmasında avantaj sağlayarak, kullanılma sıklığını arttırmıştır. Klasik test kuramında, test ve madde özellikleri, üzerinde çalışılan örnekleme ve testin kendisine bağlıdır. Ayrıca, ölçmenin standart hatasının tüm örneklemdeki kişiler için sabit bir değer olması ve kişilerin gözlenen puanları ile gerçek puanları arasındaki fonksiyonel ilişkinin doğrusal olduğunun varsayılması, yine kuramın önemli eksikliklerini göstermektedir (Budgell, Raju ve Quartetti, 1995; Embretson, 1999; MacDonald ve Paunonen, 2002). Klasik test kuramı, özellikle örnekleme bağlı varsayımlarından dolayı, grup içi karşılaştırmalar yapmaya uygundur, ancak gruplar arası karşılaştırmalar söz konusu olduğunda kuram yetersiz kalmaktadır. Bu noktada madde cevap kuramı, klasik test kuramının güçsüzlüklerini karşılayan ve daha güçlü varsayımlar gerektiren modellerle son yıllarda çeşitli uygulamalarda yerini almıştır.

Madde işlevsel farklılığı ya da ölçme eşdeğerliğinin incelenmesinde başlıca iki analiz yönteminden bahsedilmektedir. Bunlar; yapısal eşitlik modeli (Structural Equating Model-SEM) ya da doğrulayıcı faktör analizi (Comfirmatory Factor Analysis-CFA) olarak bildiğimiz doğrusal yöntemler ile madde cevap kuramına dayalı modelleri içeren doğrusal olmayan yöntemlerdir (Raju ve ark., 2002). Bu ya-

zıda ölçme eşdeğerliğinin incelenmesinde madde cevap kuramına dayalı olan modeller üzerinde durulmuştur.

Madde işlevsel farklılık araştırmaları, ırksal ve etnik grup farklılıkları, cinsiyet farklılıkları, yaş, sosyo-ekonomik düzey, eğitim, kırsal-kent kökeni v.b. farklılıkların alt evrenlerde karşılaştırmalarını içermektedir. Testteki bir maddenin, katılımcının içinde yer aldığı demografik grup üyeliği söz konusu olmaksızın yetenek, kişilik, tutum v.b özellikleri doğru biçimde ölçebilmesi, testlerin güvenirliği ve geçerliği bağlamında psikometri alanındaki incelemelerin en önemli yönünü oluşturmaktadır. Madde cevap kuramı, gerek kültürler arası ölçek/test uyarlamalarında ölçme eşdeğerliğinin sınanmasında, gerekse yeni bir psikolojik ölçme aracının geliştirilmesi sürecinde testlerin gruplar arası karşılaştırmalarında maddelerin, madde güçlük, ayırdedicilik vb. psikometrik özellikleri ve testlerin yapı geçerliğine ilişkin ayrıntılı bilgiler sağlamaktadır. Madde işlevsel farklılığı, test/ölçeklerin farklı alt evrenlerde, hem test hem de madde düzeyinde karşılaştırılmasını imkan vermektedir; bu da kuramı çok daha çekici ve işlevsel hale getirmektedir.

Madde işlevsel farklılık uygulamaları son yıllarda madde cevap modellerine uygun bilgisayar programlarının geliştirilmesiyle eğitim, psikoloji, sağlık vb. pek çok alanda araştırmacılara hizmet vermeye başlamıştır. Model karşılaştırma yönteminde olabilirlik oranı testi değerlerini elde etmek üzere MULTILOG programı (Thissen, 1991), madde parametrelerini karşılaştırma yöntemi için PARSCALE programı (Muraki ve Bock, 1996) ve madde-test işlevsel farklılık yöntemi (DFIT) için de DFITP6.0 programı kullanılmaktadır (Raju, 2004a). Ayrıca çeşitli modellere uygun değişik yazılım programları da mevcuttur.

Ülkemizdeki psikologlar arasında madde cevap kuramı ile ilgili çalışmaların var olduğu görülmekle birlikte, henüz madde ve test işlevsel farklılığı konusundaki incelemeler yok denecek kadar azdır. Ayrıca, ülkemizde psikolojik ölçme araçları olarak kullanıma sunulan ölçek/test/envanter gibi pek çok materyalin büyük çoğunluğunun çeviri ve adaptasyon çalışmalarını kapsıyor olması, aslında psikologların farkında olmadıkları bir problemi göz ardı ettiklerini kısmen göstermektedir. Başka bir deyişle, bir kültürden diğer bir kültüre uyarlanan ölçme araçlarının kültürler arası ölçme eşdeğerliği incelemeleri yapılmadığında, o ölçme aracının tam olarak neyi ölçtüğünden emin olunamaz. Ülkemizde bu tür incelemelerin hali hazırda yeterince yerine getirilmediği görülmektedir. Dünya literatüründe ise, madde-test işlevsel farklılık incelemeleri, eğitim ve psikoloji alanında çok hızlı ilerlemeler gerçekleştirmiş olup, diğer disiplinlerdeki araştırmacıların da dikkatini çekerek tıp, ekonomi, eczacılık, biyoloji, sosyoloji v.b. alanların da pratik uygulamaları içerisinde kendisine yer açmıştır. Dolayısıyla, ülkemizde psikolojik testlerin geliştirilmesinde yaygın olarak kullanılan klasik test kuramının uygulama kolaylıklarının yanı sıra, psikoloji araştırmalarına madde cevap kuramına dayalı madde işlevsel farklılık incelemelerinin yer almasının yararlı olacağı düşünülmektedir.

Kaynaklar

Bolt, D.M., Hare, R.D., Vitale, J.E., & Newman, J.P. (2004). A multigroup item response theory analysis of the Psychopathy Checklist-Revised. *Psychological Assessment*, *16*,(2), 155-168.

Budgell, G.R., Raju, N.S., & Quartetti, D.A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement, 19*, (4), 309-321.

Camilli, G., & Shepart, L.A. (1994). *Methods for identifying biased test items.* London: Sage Publication

Chernyshenko, O. S., Stark, S., Chan, K.Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36*(4), 523-562.

Cohen, A.S., Kim, S., & Baker, F.B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement*, *17*, (4), 335-350.

Cooke, D.J., & Michie, C. (1999). Psychopathy across cultures: North America and Scotland compared. *Journal of Abnormal Psychology*, *108*, (1), 58-68.

Cooke, D.J, Michie, C., Hart, S.D., & Hare, R.D. (1999). Evaluating the screening version of the Hare Psychopathy Checklist-Revised (PCL: SV): An item response theory analysis. *Psychological Assessment, 11*(1), 3-13.

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

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

Embretson, S.E. (1999). Issues in the measurement of cognitive abilities. In S.E Embretson & S.L. Hershberger (Eds.), T*he new rules of measurement: What every psychologist and educator should know* (pp. 1-15). Lawrence Erlbaum Associates,Inc.

Facteau, J.D., & Craig, S.B. (2001). Are performance appraisal ratings from different rating sources comparable? *Journal of Applied Psychology*, *86*(2), 215-227.

Flowers, C.P., Oshima, T.C. & Raju, N.S. (1999). A Description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement*, *2* (4), 309-326.

Glöckner-Rist, A., & Hoijtink, H. (2003). The best of both words: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling*, *10*(4), 544-565.

Hambleton, R.K., Robin, F., & Xing, D. (2000). Item response models for the analysis of educational and psychological test data. In H.E.A.Tinsley & S.D.Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp.553-581). San Diego: Academic Press.

Hambleton, R.K., & Swaminathan, H. (1989). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff Publishing

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

Hanson, B.A. (1998). Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational And Behavioral Statistics*, *23*(3), 244-253.

Hayduk, L.A. (1987). *Structural equation modeling with LISREL*. USA: The Johns Hopkins University Press

Huang, C.D., Church, A. T., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits: Differential item functioning in the NEO personality inventory. *Journal of Cross-Cultural Psychology*, *28* (2), 192-248.

Hulin, C. H., Drasgow, F., & Parsons, C.K. (1983). *Item response theory*: Application to psychological measurement, Dow Jones-Irwin.

Kim, S.-H., & Cohen, A.S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied Psychological Measurement*, *15* (3), 269-278.

Kim, S.-H. & Cohen, A.S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education*, *8*(4), 291-312.

Kim, S.-H., & Cohen, A.S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement*, *22*(4), 345-355.

Kim, S.-H., Cohen, A.S., & Kim, H.-A. (1994). An investigation of Lord's procedure for the detection of differential item functioning. *Applied Psychological Measurement*, *18*(3), 217-228.

Korkmaz, M. (2005). *Madde cevap kuramına dayalı olarak çok kategorili maddelerde madde ve test yanlılığının (işlevsel farklılığın) incelenmesi.* Yayımlanmamış doktora tezi, Ege Üniversitesi, Sosyal Bilimler Enstitüsü, İzmir.

Lambert, M.C., Schmitt, N., Samms-Vaughan, M.E, Shin An, J., 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.

Lim, R.G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology*, *75*(2), 164-174.

MacDonald, P., & Paunonen, S.V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, *62*(6), 921-943.

Maller, S.J. (2001). Differential item functioning in the WISC-III: Item parameters for boys and girls in the national standardization sample. *Educational and Psychological Measurement*, *61*(5), 793-817.

Maurer, T.J., Raju, S.N., & Collins, W.C. (1998). Peer and subordinate performance appraisal measurement equivalence. *Journal of Applied Psychology*, *83*(5), 693-702.

McAllister, P.H. (1993). Testing, DIF and public policy. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 389-396). Lawrence Erlbaum Associates, Inc.

Meade, A.W., & Lautenschlager, G.J. (2004a). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, 7(4), 361-388.

Meade, A.W., & Lautenschlager, G.J. (2004b). Same questions, different answers: CFA and two IRT approaches to measurement invariance. *19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago.*

Meyer, P.L. (1970). *Introductory probability and statistical applications.* (2nd ed.). Addison-Wesley Publishing Company, Inc.

Morales, L.S, Reise, S.P., & Hays, R.D. (2000). Evaluating the equivalence of health care ratings by Whites and Hispanics. *Medical Care*, *38*(5), 517-527.

Muraki, E., & Bock, R.D. (1996). *PARSCALE (V4.1). Parameter scaling of rating data.* Chicago, IL: Scientific Software, Inc.

Orlando, M., & Marshall, G.N. (2002). Differential item functioning in a Spanish Translation of the PTSD checklist: Detection and evaluation of impact. *Psychological Assessment, 1* (1), 50-59.

Oshima, T.C., Raju, N.S, Flowers, C.P., & Slinde, J.A. (1998). Differential bundle functioning using the DFIT framework: Procedures for identifying possible sources of differential functioning. *Applied Measurement in Education*, *11*(4), 535-369.

Raju, N.S. (2004a). *DFITP6: A FORTRAN program for calculating DIF/DTF* [Computer Software]. Chicago: Illinois Institute of Technology.

Raju, N.S. (2004b). *Some notes on the DFIT framework.* (unplished manuscript). Chicago: Illinois Institute of Technology

Raju, N.S., Laffitte, L.J., & Byrne, B.M. (2002). Measurement equivalence: 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 S.E. Embretson & S.L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 219-241). Lawrence Erlbaum Associates,Inc.

Reise, S.P., Smith, L., & Furr, R.M. (2001). Invariance on the NEO PI-R neuroticism scale. *Multivariate Behavioral Research*, *36*(1), 83-110.

Reise, S.P., Wideman, 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.

Robie, C., Zickar, M.J., & Schmit, M.J. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. *Human Performance*, *14*(2), 187-207.

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

Smith, L.L. (2002). On the usefulness of item bias analysis to personality psychology. *Personality and Social Psychology Bulletin*, *28*(6), 754-763.

Somer, O. (1998). Kişilik testlerinde klasik ve modern test kuramları ile madde analizi. *Türk Psikoloji Dergisi*, *13*(41), 1-15.

Somer, O. (1999). Çok kategorili (polytomous) maddelerde klasik ve modern test kuramlarının madde analizleri, güvenirlik ve bilgi kavramları açısından karşılaştırılması. *Türk Psikoloji Dergisi*, *14*(44), 63-75.

Somer, O.(2004). Gruplararası karşılaştırmalarda ölçek eşdeğerliğinin incelenmesi: Madde ve test fonksiyonlarının farklılaşması. *Türk Psikoloji Dergisi*, *19*(53),69-82.

Stark, S., Chernyshenko, O.S., Chan, K., Lee, W.C., & Drasgow, F. (2001). Effects of testing situation on item responding: cause for concern. *Journal of Applied Psychology*, *86* (5), 943- 953.

Teresi, J.A. (2000). Applications of item response theory to the examination of the psychometric differential item functioning of the comprehensive assessment and referral evaluation dementia diagnostic scale among samples of Latino, African American, and white non-Latino elderly. *Research on Aging*, *22*(6), 378-414.

Thissen, D. (1991). *MULTILOG: Multiple category item analysis and test scoring using item response theory (Version 7.03).* Chicago: Scientific Software International, Inc.

Thissen, D. (2001). IRTLRDIF v.2.0: *Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioni*ng. Chapel Hill: University of North Carolina.

Thissen, D., Steinberg, L., & Wainer, T. (1988) *Use of ıtem response theory ın the study of group differences on trace lines.* Ed.H. Wainer, Braun, H.I, *Test Validity*, Lawrence Erlbaum Associates Inc.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In. P.W.Holland & H.Wainer (Eds). *Differential item functioning* (pp. 67-113).Hillsdale, NJ: Erlbaum.

Van de Vijyer, F., & Leung, K. (1997). Methods and data analysis of comparative research. In J.W. Berry & Y.H. Poortinga (Ed.*), Handbook of cross-cultural psychology, Vol.1: Theory and method* (2nd ed., pp.257-300). Needham Heights, MA: Alleyn&Bacon.

Zickar, M.J. (1998). Modeling item-level data with item response theory. *Current Directions in Psychological Science*, *7*, (4), 104-109.