

Psikolojik Ölçmenin Yeni Kuralları ve Türkiye'deki Durumu

Mediha Korkmaz

Ege Üniversitesi, Psikoloji Bölümü

Mediha.korkmaz@ege.edu.tr

Psikolojinin bir bilimdalı olarak doğuşuyla birlikte o yıllarda acilen ölçülmesi gereken bireysel farklılıklardan kaynaklanan zeka, yetenek, kişilik gibi temel bazı psikolojik yapılar, özellikler vardı. Günümüzde de hala bu özellikler varlığını artarak sürdürmektedirler. 19. yüzyıldaki psikolojik yapılar, günümüzde bir "özellikler fetişizmi" haline mi geldi? Nasıl ölçülecek? Elbette ki bir psikolog ya da bir psikometrist tarafından psikolojik yapıların ölçülmesine ilişkin verilecek en temel yanıt, yapıları kapsayan "bir psikolojik test, bir ölçek, bir envanter" ile olmasıdır. Psikolojik özelliklerin ölçümü üzerine adeta bir yarışma varmış gibi ölçme araçlarının çok fazla sayılarda oluşturulduğunu yurtdışındaki uygulamalardan izleyebilmekteyiz. Oysa, bir test geliştiricisi, psikolojik bir yapının/ özelliğin ölçümüne ilişkin bir ölçme aracının geliştirilmesinin, araştırmacının oturduğu yerden yazdığı maddelerden kolayca oluşamayacağını, aksine test geliştirme sürecinin son derece zahmetli ve zaman alıcı olduğunu çok iyi bilir. Ancak bu zahmetli sürecin ardından geliştirilen bir ölçme aracı, test kullanıcıların yani uygulamacıların hizmetine sunulur. Pek tabidir ki süreç ölçme aracının uygulamaya sunulması ile de bitmez; belki de asıl sorunların en önemli parçası bu aşamada başlamaktadır. Çünkü bir ölçme aracının en temel özelliği olan "geçerlik" bir kanıt toplama sürecidir ve bu çalışmaların sonu neredeyse yoktur. Ayrıca zaman içerisinde içinde yaşanan toplumun değişen özelliklerine karşı, daha önce geliştirilmiş olan ölçme aracının hala ölçmeyi amaçladığı psikolojik özelliği ölçüp ölçmediği sorularının periyodik olarak yanıtlanması ve ölçme aracının güncellenmesi de gerekmektedir.

"Psikolojik ölçme" adı altında yayınlanan pek çok ders kitabında başlıca psikolojide ölçme ve değerlendirme konularından söz edilmektedir. Özellikle yurtdışındaki en genel adıyla yayınlanan "Psikometrik Teori (Psychometric Theory)" ya da "Psikolojik Ölçme (Psychological Testing)" ders kitaplarında "klasik test kuramı" kapsamındaki yöntemler (güvenirlilik ve çeşitli güvenirlik yöntemleri, geçerlik ve çeşitleri, norm ve standart puanlar, bazı psikolojik özelliklerin ölçümüne ilişkin geliştirilmiş testler) yer almaktadır. İster yurtdışında isterse ülkemizde psikoloji lisans eğitimi alan her birey, yaklaşık olarak psikolojideki ölçme sorunları ve bu sorunların çözümüne ilişkin ölçme alanındaki bu klasik test kuramı yöntemleriyle tanışmaktadır.

Oysa, klasik test kuramının yanı sıra son yıllarda "madde cevap kuramı", "maksimum olasılık faktör analizi", yapısal eşitlik modelleri", "genellenebilirlik kuramı" gibi yeni ölçme modelleri psikometri yazınında yer almaktadır. Ancak buradaki ciddi sorun, gerek yurtdışında gerekse Türkiye'de halihazırda bu yeni psikometrik yöntemlerin lisans eğitimi kapsamında ele alınmamasıdır. Özellikle de ülkemizde psikoloji lisansüstü programlarının yeni yöntemler açısından eksikliği göze çarpmaktadır.

Bu yazının amacı, ölçme araçlarının geliştirme aşamalarında yer alan ve yukarıda bazı sorun alanlarına işaret edilen noktaların hepsine ayrıntılı kuramsal yanıtlar vermek değildir. Psikometri yazınında yer alan yeni ölçme yöntemlerinden madde cevap kuramı hakkında en temel bazı avantajlara değinmektir.

Madde cevap kuramı (item response theory) ya da örtük özellikler kuramı (latent trait theory), modern test kuramı olarak da adlandırılmaktadır. Böylece klasik test kuramının “eski”, modern test kuramının da “yeni” ölçme yöntemleri olduğuna vurgu yapılmış olur.

Madde cevap kuramı, bireylerin davranışları ile bu davranışların altında örtük olarak bulunduğu varsayılan özellikler arasındaki ilişkileri, olasılığa temellenen fonksiyonlarla ifade eder. Kuramın odak noktasını farklı yetenek düzeyindeki deneklerin örtük özellik üzerindeki bir maddeyi nasıl yanıtlamaları gerektiğini gösteren matematiksel modeller oluşturur. Bu modeller, gözlenen değişkenler ile bunların altında yatan örtük özellik arasındaki işlevsel ilişkiyi doğrusal olmayan bir regresyon ile tanımlarlar (Chernysenko ve ark., 2001; Zickar, 1998). Madde cevap kuramında madde (item) terimi bireylerin gözlem birimini, test/ölçek (test/scale) terimi maddeler topluluğunu ve özellik/yetenek (trait/ability) terimi bireyin örtük özelliğini tanımlamak üzere kullanılmaktadır.

Madde cevap kuramının merkezi elementi, madde karakteristik eğrisidir (item characteristics curve - ICC) ve bireyin ölçülen yetenek boyutundaki düzeyi ile maddeye doğru cevap verme olasılığı arasındaki ilişkinin grafik gösterimini verir. Kuramda madde güçlüğü ve yetenek/örtük özellik ölçüm tahminleri aynı boyut üzerinde, yani θ test ile ölçülen yetenek boyutu (θ) üzerinde yer almaktadır. Dolayısıyla madde karakteristik eğrisi, doğru cevabın koşula bağlı olasılığın (conditional), yani herhangi bir belirli θ yetenek düzeyi için $P(\theta)$ belirlenmiş olan doğru cevabın eğrisidir (Hambleton, Swaminathan ve Rogers, 1991; Hulin, Drasgow ve Parsons, 1983). Madde cevap kuramının en önemli diğer bir özelliği de, madde ve test bilgi fonksiyonlarıdır (item

and test information function). Kuramda bu fonksiyonlar, klasik test kuramındaki güvenilirlik ve ölçmenin standart hatası kavramlarını karşılamaktadır. Madde ve test bilgi fonksiyonları, gözlem yoluyla elde edilen verilerin parametre tahminlerinin yapılmasının ardından θ yetenek tahminlerinin yerleşiminin ve aynı zamanda ölçümün doğruluğu, hassasiyeti ve mükemmelliği hakkında bir değerlendirme yapılmasını sağlar. Madde cevap kuramında ölçmenin hassasiyeti, ölçekteki her maddenin belirli bir kişinin yeteneğine göre olan durumuna bağlı olarak, diğer bir ifadeyle, madde ve kişinin birbirlerine yakınlık derecesi ile açıklanır ve madde ile kişi arasındaki fark arttıkça maddenin etkinliği azalır (Somer, 1998, 1999, 2004). Bir test belirli bir özelliği ölçen θ yeteneğinin farklı noktalarında bulunan kişiler için farklı düzeylerde bilgi vermektedir. θ 'nın her bir değeri için elde edilen informasyonların θ 'ya karşı grafiğinin çizilmesi ile “madde bilgi fonksiyonu” elde edilir ve belirli bir θ düzeyi için tüm madde bilgi fonksiyonlarının toplanması ile de “test bilgi fonksiyonu” elde edilmiştir.

Niçin Madde Cevap Kuramı?

Psikologların çoğu araştırmalarında test ve ölçekleri değerlendirmek üzere klasik test kuramı kapsamında ele alınan “güvenirlilik, madde-toplam puan korelasyonu, Spearman-Brown düzeltme formülasyonu” gibi aşına olduğumuz tekniklerini sıklıkla kullanmaktadırlar. Klasik test kuramının varsayımlarını karşılamak daha kolaydır ve kuramın uygulanmasında avantaj sağlayarak kullanılma sıklığını artırmıştır. Bu durumu Hambleton, Robin ve Xing (2000), kuramın yeterince güçlü olmadığı ve bazı yetersizliklerinin bulunduğu şeklinde yorumlarken; Budgell, Raju ve Quartetti (1995), kuramın özellikle örnekleme bağlı varsayımlarından dolayı grup içi karşılaştırmalar için uygun olduğunu, buna karşın gruplararası

karşılaştırmalarda yetersiz kaldığını belirtmektedirler. Madde cevap kuramı modelleri çok daha güçlü varsayımların karşılanmasını (tekboyutluluk, yerel bağımsızlık vb.) gerektirmektedir; eğer varsayımlar karşılanmaz ise, sözkonusu modelleri değerlendirmede kullanmak hatalı sonuçlar ortaya çıkarır. Madde cevap kuramı, klasik test kuramına temellenen psikometrik değerlendirmedeki yetersiz kalan bazı önemli sorunları aşmak üzere geliştirilmiştir.

Klasik test kuramının aşamadığı sorunları madde cevap kuramı nasıl aşıyor? Embretson (1999) her iki kuramın temel problemler üzerinden karşılaştırmasını eski ve yeni kurallar üzerinden şu şekilde yapmaktadır:

Eski Kural 1: *Ölçmenin standart hatası belirli bir örnekleme için tüm puanlara uygulanır.*

Yeni Kural 1: *Ölçmenin standart hatası farklı cevap örüntüleriyle bireyler arasında değişir ancak farklı popülasyonlara genellenebilir.*

Klasik test kuramında güvenilirlik katsayısına bağlı olarak hesaplanan ve kişilerin gerçek puanlarının güven aralıklarını tahminlemeye yarayan ölçmenin standart hatasının örnekleme için tüm denek puanları için eşit, aynı olduğu varsayımı vardır. Oysa, bir örnekleme için güvenirlilik katsayıları değişim gösterir dolayısıyla ölçmenin standart hatası da değişir. Örneklem birebir eşdeğer olmadığı sürece de bir örnekleme için puan aralıkları diğer örnekleme göre değişiklik gösterecektir. Buradaki asıl sorun ise, ölçmenin standart hatasının örnekleme için tüm bireyler için aynı olmasından kaynaklanmaktadır; yani düşük yetenekteki bir birey için olduğu kadar yüksek yetenekteki bir birey için de aynı puan aralıklarının kullanılmasıdır. Sonuçta, klasik kuramda bir kişinin belirli bir aralıkta ölçmenin standart hatasına uygun olarak gerçek puan aralıklarını saptamanın doğruluğu sorun olmaktadır.

Madde cevap kuramında ise, tahminlenen yeteneğin/özelliğin doğruluğunun bir ölçütü olarak madde ve test bilgi fonksiyonları kullanılmaktadır ve tahminlenen farklı yetenek (theta) değerlerinin doğruluğuna ilişkin klasik kurama kıyasla daha hassas puan aralıklarına ulaşılmasına olanak verirler. Diğer bir deyişle, her bir maddenin, madde bilgi fonksiyonunun olması kuramı örneklemeden bağımsız bir özelliğe taşır, popülasyonun gerçek puan ve hata varyanslarına bağlı değildir. Madde ve test bilgi fonksiyonları bir testi geliştirirken ölçümün doğruluğunu arttırmak için ne tür maddelerin eklenmesi ya da ne tip maddelerin testten uzaklaştırılmasını saptamada oldukça faydalı bilgiler sağlarlar. Samejima (1977) klasik test kuramında güvenirliliğin cansız bir kavram olduğunu çünkü bir gruptan diğerine farklılık gösterdiğini ve genellenebilirliğinin oldukça sınırlı olduğunu ifade etmektedir (akt. Hulin, Drasgow ve Parsons, 1983).

Eski Kural 2: *Çok maddeli uzun testler, az maddeli kısa testlerden daha güvenilirdir.*

Yeni Kural 2: *Az maddeli kısa testler, çok maddeli uzun testlerden daha güvenilir olabilir.*

Klasik test kuramında, ölçme aracındaki madde sayısı arttıkça testin güvenirlilik katsayısının da artacağı ve madde sayısının az olmaması tercih edilir. Buna karşın madde cevap kuramında az sayıda madde içeren testler oldukça güvenilir olabilir. Klasik kuramda Spearman-Brown düzeltme formülüyle ölçme aracının madde sayısı artırıldığında, hata varyansı azalır gerçek varyans miktarı arttığı için testin güvenirlilik derecesi de artmaktadır (Embretson, 1999). Madde cevap kuramında ise, ölçmenin standart hatası her bir örtük özellik aralığında tanımlandığı için hata varyanslarının miktarı azalır, güvenirlilik katsayısı yükselir. Bu durum özellikle deneklerin yetenek düzeylerine uygun madde bankalarının

oluşturulmasında kullanılır (örneğin testte 25 madde olsa bile, deneklerin yetenek düzeylerine uygun oluşturulan daha az sayıdaki madde setleri kullanılır). Çünkü deneklerin yetenek düzeylerine en uygun maddeler kullanıldığında standart hatalar da küçük olacaktır, bu da testin güvenilirliğini arttıracaktır. Dolayısıyla az sayıda maddeleri olan kısa testlerde çok daha güvenilir olabilir.

Eski Kural 3: *Bir testin farklı formlarından elde edilen test puanlarının karşılaştırılması ancak testlerin paralel ve eşit olmasına bağlıdır.*

Yeni Kural 3: *Testin farklı formlarından elde edilen test puanlarının karşılaştırılması testin güçlük düzeyi bireyler arasında değişiklik gösterdiğinde en mükemmeldir: Test puanlarının karşılaştırılması için paralel formlar gerekmez.*

Klasik test kuramında bireylerin gözlenen test puanları, gerçek puan ve hata bileşenlerinin bir fonksiyonu olarak ele alınır. Kuram doğrudan test düzeyinde bilgileri sağlamak üzere geliştirilmiştir ve doğrudan ölçme yapılan araca bağlıdır. Madde cevap kuramında ise, ölçülen yetenek/özelliğ ve madde puanları aynı ölçek üzerinde (θ metriği) tanımlanmaktadır. Klasik ve madde cevap kuramındaki gerçek puan ve θ değerleri farklı metriklerde aynı yeteneği/özelliği temsil ederler. Ancak ikisinin arasında çok önemli bir ayrım vardır. θ metriğinin ortalama ve varyans değerleri bir kez saptandıktan sonra, bir bireyin θ değeri ölçüm yapılan araca bağlı olmaktan çıkar (Hulin, Drasgow ve Parsons, 1983). Böylece θ metriği ile ölçülen bir bireyin yeteneğinin farklı madde setlerini kapsayan ölçüm araçları ile karşılaştırılmasına da olanak tanımış olur. Bu özellik madde-test parametrelerinin farklı örneklem ve farklı test formlarında değişmezlik/sabitlik (invariance) olarak bilinir ve kuramın uyarlamalı ölçme (adaptive testing) uygulamalarında kullanılır. Buna karşın

klasik kuramda ölçme araçları *birebir-mutlak* paralel olmadığı sürece bu tür bir yetenek karşılaştırması yapılamaz.

Eski Kural 4: *Madde istatistikleri popülasyonun temsili olan örnekleme bağlıdır.*

Yeni Kural 4: *Madde istatistikleri temsil edici örneklemere bağlı değildir.*

Klasik test kuramında madde güçlüğü ve madde ayırt edicilik istatistikleri büyük ölçüde üzerinde çalışılan örnekleme bağlıdır ve örneklemin yetenek/özelliğ seviyesi bu test istatistiklerinin düzeyini etkilemektedir (MacDonald ve Paunonen, 2002). Çünkü kuramda denek puanları madde güçlük düzeylerinin bir fonksiyonu olarak ele alınmaktadır ve bu nedenle de bir testin güçlüğü farklı alt popülasyonlarda değişiklik göstermektedir. Dolayısıyla değişen güçlük düzeylerinde bir testin değişik formları kullanıldığında denekler arasında karşılaştırmalar yapılamaz (Hambleton, Robin ve Xing, 2000). Madde cevap kuramında ise, madde ayırt edicilik parametresi (ai), madde güçlük parametresi (bi) ve maddenin doğru yanıtını tahmin parametresi (ci) deneklerin ait olduğu alt popülasyonlara bağlı değildir ve bir örneklemden diğerine değişiklik içermezler. Çünkü madde cevap modelleri verilerin modellenmesine temellenir, ölçme aracı ile değerlendirilen yetenek ya da örtük özellik ve testteki her bir madde arasındaki ilişkiye odaklanır, hem madde parametrelerinin örneklemden bağımsız, hem de kişi parametrelerinin maddelerden bağımsız olması nedeniyle kişi ve madde parametre tahminlerini yapar. Böylece test puanlarının yorumlanması için standardizasyonu yapılmış orijinal örnekleme bağlı kalmak gerekmez.

Klasik test kuramında bireylerin gözlenen test puanları, maddelere verilen cevapların toplamından oluşur ve standart puanlara dönüştürülür. Kuram doğrudan test düzeyinde bilgileri sağlar ve ölçme yapılan araca bağlıdır. Madde

cevap kuramında ise, kişinin örtük özellik düzeyini saptamak için madde cevaplarının nasıl toplanacağı (toplam puan) kuramın yanıtlaması gereken sorular arasında değildir. Madde cevap kuramı daha çok kişilerin davranışlarını etkileyen maddelerin özelliklerinin nasıl olduğu ve kişinin cevaplarını açıklamak için en uygun özellik düzeyinin ne olduğu ile ilgilenir (Embretson, 1999). Kısaca; klasik kuram test düzeyinde test puanlarını tanımlamaya odaklanırken, madde cevap kuramı madde düzeyinde ve madde performansının modelleşmesine ilişkin bilgileri sağlamaya odaklanır (Cohen, Bottge ve Wells, 2001; Kolen ve Brennan, 1995).

Madde Cevap Kuramının Psikolojiye Bazı Katkıları

Madde cevap kuramı modelleri, özellikle test yapılandırma sürecinde maddelerin psikometrik özellikleri ve testlerin yapı geçerliğine ilişkin ayrıntılı bilgiler sağlar. Kuram geçmişte olduğu gibi günümüzde de eğitimde ölçme ve değerlendirme araştırmacıları tarafından aktif olarak kullanılmaktadır. Literatür taramaları incelendiğinde bu durum hem ülkemizde hem de yurtdışı uygulamalarında karşımıza çıkmaktadır. Son 50 yıllık dönemde de madde cevap kuramı yurtdışı psikoloji araştırmalarında önemli bir mevki edinmiştir. Ancak ülkemizde psikoloji araştırmacıları tarafından gerektiği kadar tanınmamaktadır. Oysa, madde cevap kuramı yeni bir test-ölçek geliştirme çalışmalarında “doğrulamalı faktör analizi” kadar büyük bir etkiye sahiptir. Yukarıda ele aldığımız madde cevap kuramının özellikleri klasik test kuramının gerçekleştiremediği bazı ölçme problemlerinin üstesinden gelebilmektedir. Kuramın psikoloji araştırmalarına en önemli katkısı belki de ölçme aracındaki potansiyel yanlılık (bias) incelemelerine izin vermesidir.

Psikoloji araştırmalarında üzerinde çalışılan örneklemelerin sistematik bazı özellikleri (cinsiyet, eğitim, dil, ırk vb.) vardır ve gruplar arası karşılaştırmalar da bu sistematik özellikler üzerinden yapılabilmektedir. Madde cevap kuramı ölçme araçlarının bu sistematik özellikler açısından incelenmesine olanak sağlar. Kuramda bu özellik *madde ve test işlevsel farklılık fonksiyonu* (differential item and test functioning) olarak tanımlanır, ölçme aracını oluşturan maddelerin sistematik grup üyeliklerinden herhangi birine avantaj sağlayıp sağlamadığı sorunu incelenir (Korkmaz, 2005; 2006). Böylelikle test maddelerin karşıt örneklem içerisindedir farklı bir işlevselliğinin olmadığı ve ölçümlerin yapı geçerliğinin önemli kanıtları toplanmış olur.

Bu özellik ülkemiz psikolojisi açısından aslında son derece önemlidir. Halihazırda ülkemizde kullanılan psikolojik yapılara ilişkin pek çok ölçme aracının aslı yurtdışında geliştirilmiş olup, ülkemiz için Türkçe çeviri ve adaptasyon çalışmaları yapılmıştır. Bu durumda karşılaştığımız sorun daha da büyümektedir. Çünkü testteki maddelerin grup üyeliklerine karşı bir avantaj gösterip göstermediğinin yanı sıra “kültürlerarası” farklılıkların da diğer bir sistematik özellik olarak karşımıza çıktığını görmekteyiz. Bir kültür için geliştirilmiş bir ölçme aracı, acaba diğer bir kültürde ne kadar geçerliğe sahiptir? Kaldı ki bu ölçme araçları diagnostik ve karar verme amaçlı kullanıldıklarında ne kadar doğru bir sürecin gerçekleştirildiğinden emin olabilir miyiz? Klasik test kuramı bu sorulara yanıt vermekte maalesef yetersiz kalmaktadır ve Türkiye’deki psikoloji araştırmalarına madde cevap modellerinin dahil edilmesi dünya literatürünü daha iyi izlememizi sağlayacaktır.

Başlangıçta madde cevap modelleri diktomatik madde formatlarının ölçülmesine ilişkin geliştirilmiş olmakla birlikte, günümüzde daha karmaşık madde formatlarına

yönelmiştir. İki cevap seçeneğinden daha fazla sayıda seçeneği olan çoklu kategorili maddeler (polytomous) için pek çok model geliştirilmiştir (örneğin; Masters (1982) "Kısmi Puanlama Modeli" (Partial Credit Model), Samejima (1969) "Ağırlıklandırılmış Cevap Modeli" (Graded Response Model), Andrich (1978) "Dereceli Ölçek Modeli" (Rating Scale Model) ve Bock (1972) "Sınıflamalı Cevaplar Modeli" (Nominal Response Model)). Böylece psikolojinin tutum, kişilik, örgüt ve organizasyon araştırma konuları gibi pek çok alanda uygulanabilir hale gelmiştir.

Dünya literatüründen bazı örnekler vermek gerekirse; Ellis ve Mead (2000) 16 Faktör Kişilik Envanteri'nin İspanyolca versiyonunun ölçme eşdeğerliğinin incelenmesinde, Cooke, Michie, Hart, Hare (1999, 2004), Psikopati kontrol listesinin (Psychopathy Checklist - revised - PCL-R) geliştirilmesinde hemde kültürler arası karşılaştırmalarının yapılmasında, Huang, Church ve Katigbak (1997) NEO Kişilik Envanterinin kültürel farklılıklarının incelenmesinde, Smith ve Reise (1998) Stres Tepki Ölçeğinin (Stress Reaction Scale) madde işlevsel farklılık incelemelerinde, Childs, Dahlstrom, Kemp ve Panter (2000), MMPI-2 Depresyon Ölçeğinin madde parametre tahminleri ve madde işlevsel farklılığının incelenmesinde, Orlando ve Marshall (2002) Post-Travmatik Stres Bozukluğu Kontrol Listesinin İspanyolca versiyonunun madde işlevsel farklılığının incelenmesinde, Maller, (2001), WISC-III standardizasyon örnekleminin cinsiyete göre madde işlevsel farklılığının incelenmesinde değişik madde cevap modelleri ve uygulamaları kullanılmıştır.

Madde Cevap Kuramının Pratik Uygulamalarda Kullanımı Zor mu?

Madde cevap modellerinin pratik uygulamalarda kolayca yer edinmemesinin en önemli nedeni, hemen hemen her bir modelin kendine özgü bir matematiksel işlem

ve bilgisayar programı gerektirmesidir. Ayrıca programların DOS ortamındaki yazılımları da kullanılma sıklığını etkilemiştir. Ancak son yıllarda madde cevap modellerine uygun olarak geliştirilen bilgisayar paket programlarının sayısında önemli bir artış vardır ve programların kullanımına ilişkin ortalama bir bilgisayar kullanıcı bilgisi yeterlidir (elbetteki sadece programın kullanımı değil aynı zamanda kuramsal düzeyde de bilgiye sahip olmak her türlü paket programın kullanımı için bir gerekliliktir). Bu teknolojik gelişme araştırmacıları teorik yapılanmadan ziyade çeşitli alanlardaki (eğitim, psikoloji, tıp, sosyoloji, biyoloji vb.) pratik uygulamalara yönlendirmiştir.

Madde cevap modellerine uygun olarak geliştirilen ve sık kullanılan önemli bazı bilgisayar paket programları BILOG-MG3, MULTİLOG, PARSCALE, TESTFACT'dir. Ayrıca ilgilenen araştırmacılar "www.assess.com" adresinden çeşitli programların özelliklerini görebilirler.

Dünya literatürü psikoloji araştırmalarında önemli bir yer edinmiş olan madde cevap modellerinin Türk psikoloji araştırmalarında da kapsanması gerekmektedir. Çünkü test ve ölçek geliştirmede son yıllarda madde cevap modelleri ve doğrulayıcı faktör analizi yöntemleri başa baş gitmekte, araştırmalarda bu iki yöntemin karşılaştırmaları yapılmaktadır. Özellikle de lisansüstü programlarda ölçmenin yeni yöntemlerinin ele alınması, ülkemiz araştırmacılarını dünya yazınına yakınlaştıracaktır. Eğer bu gelişmeler izlenmez ise, ülkemizdeki psikoloji araştırmalarının yetersiz kalması, kuramların uygun bir şekilde test edilememesi, araştırma problemlerine ilişkin yanlış verilerin toplanması, dergilerdeki yayın kalitesinin düşmesi, istatistik, ölçme ve yöntem konularını değerlendirecek uzman sayısında azalma olması gibi potansiyel sorunlarla karşılaşmak kaçınılmaz olacaktır.

Kaynaklar

- Budgell, G. R., Raju, N. S. ve Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement*, 19 (4), 309-321.
- Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F. ve Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, 36 (4), 523-562.
- Childs, R. A., Dahlstrom, W. G., Kemp, S. M. ve Panter, A. T. (2000). Item response theory in personality assessment: A demonstration using the MMPI-2 depression scale. *Assessment*, 7 (1), 37-54.
- Cohen, A. S., Bottge, B. A. ve Wells, C. S. (2001). Using item response theory to assess effects of mathematics instruction in special populations. *Exceptional Children*, 68 (1), 23-44.
- Cooke, D. J ve 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. ve 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.
- Ellis, B. B. ve Mead, A. D. (2000). Assessment of the measurement equivalence of a Spanish translation of the 16PF questionnaire. *Educational and Psychological Measurement*, 60 (5), 787-807.
- Embretson, S. E. (1999). Issues in the measurement of cognitive abilities. S. E. Embretson ve S. L. Hershberger (Ed.), *The new rules of measurement: What every psychologist and educator should know* içinde (pp. 1-15). Lawrence Erlbaum Associates.
- Hambleton, R. K., Robin, F. ve Xing, D. (2000). Item response models for the analysis of educational and psychological test data. H. E. A. Tinsley ve S. D. Brown (Ed.), *Handbook of applied multivariate statistics and mathematical modeling* içinde (pp. 553-581). San Diego: Academic Press.
- Hambleton, R. K, Swaminathan, H. ve Rogers, H. J. (1991). *Fundamentals of item response theory*. London: Sage Publications.
- Huang, C. D., Church, A. T. ve 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. ve Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Dow Jones-Irwin.
- Kolen, M. J. ve Brennan, R. L. (1995). *Test equating: Methods and practices*. Springer-Verlag New York, Inc.
- Korkmaz, M. (2006). Test ve ölçek geliştirmede yeni yaklaşımlar: Madde cevap kuramı kapsamında madde işlevsel farklılığı (madde yanlılık) yöntemleri. *Türk Psikoloji Yazıları*, 9 (18), 63-80
- Korkmaz, M. (2005). *Madde cevap kuramına dayalı olarak çok kategorili maddelerde madde ve test yanlılığının (işlevsel farklılığın) incelenmesi*. Yayınlanmamış doktora tezi, Ege Ün. Sosyal Bilimler Enstitüsü, İzmir.
- MacDonald, P. ve 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.
- Orlando, M. ve Marshall, G. N. (2002). Differential item functioning in a Spanish Translation of the PTSD checklist: Detection and evaluation of impact. *Psychological Assessment*, 14 (1), 50-59.
- Smith, L. L. ve Reise, S. P. (1998). Gender difference on negative affectivity an IRT study of differential item functioning on the multidimensional personality questionnaire stress reaction scale. *Journal of Personality and Social Psychology*, 75 (5), 1350-1362.
- 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, güvenilirlik 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.
- Zickar, M. J. (1998). Modeling item-level data with item response theory. *Current Directions in Psychological Science*, 7 (4), 104-109.