Türk Psikoloji Dergisi
1999, 14(44), 63 - 75

Çok Kategorili (Polytomous) Maddelerde Klasik ve Modern Test Kuramlarının Madde Analizleri, Güvenirlik ve Bilgi Kavramları Açısından Karşılaştırılması

Oya Somer *
Ege Üniversitesi

Özet

Ölçme araçlarından elde ettiğimiz veriler, her zaman bir miktar ölçme hatasını içermektedir. Bu nedenle bir kişinin test edilmesi ile elde ettiğimiz puan da, o kişiden elde edebileceğimiz olası puan dağılımından sadece bir gözlemdir. Güvenirlik ve standart hata, kişilerin gerçek puanlarının içinde yer alabileceği güven aralığına dair olasılıksal tahminler yapmamıza imkan sağlayan kavramlardır. Modern test geliştirme modellerinde, standart hata değerleri, klasik kuramdan farklı olarak, ölçülen özellik üzerinde farklı konumda olan bireyler için farklı değerler sağlamaktadır. Test bilgi fonksiyonları da ölçülen boyut üzerinde testin hangi bölgelerindeki kişiler için daha dakik bilgi sağladığını anlamamıza yardımcı olmaktadır. Bu çalışmada, çok kategorili maddeleri olan bir kişilik ölçeği, madde cevap kuramıyla ve klasik kuramla analiz edilmiş ve modern kuramdaki kavramlardan yararlanılarak, ölçülmesi hedeflenen gruba ilişkin daha dakik ölçümler sağlayan testlerin oluşturulmasının olanaklı olduğu sonucuna varılmıştır.

Anahtar Sözcükler: *Çok kategorili maddeler, klasik test kuramı, modern test kuramı, standart hata, güvenirlik, test bilgi fonksiyonu.*

Abstract

Data obtained from psychological measures always include some error variance. The score obtained from one administration is only one of the observable values from its possible score distribution. Reliability and standard error of measurement enable test users to estimate the confidence interval which contains the person's true score. In modern test theory, the values of standard error of measurement provide different values for individuals who are at different regions on the measured latent trait. Furthermore, item and test information functions in modern test developing models offer an alternative way of constructing more reliable psychological tests. In this study, a personality scale which has polytomous items, analyzed with both the Classical Test Theory and Item-Response Theory (IRT). It can be concluded that IRT concepts, such as item and test information functions, seem to be more efficient than the classical item parameters in making more precise distinctions between different regions of the latent trait.

Key-words: *Polytomous items, classical test theory, modern test theory, standart error of measurement, reliability, item and test information functions.*

* Yazışma Adresi : Doç. Dr. Oya Somer, Ege Üniversitesi, Edebiyat Fakültesi Psikoloji Bölümü 35100 Bornova-İZMİR
E.posta : osomer@edebiyat.ege.edu.tr

Günümüze değin test kuramcıları ölçülecek bir özelliği bir boyut üzerinde ölçeklemenin çeşitli yönleri üzerinde durmuşlar ve farklı modeller önermişlerdir.

Klasik kuramcılar madde güçlüğü ve ayırdetme parametreleri ile madde karakteristiklerini tanımlamışlar ve eşit aralıklı ölçekler geliştirme yolunda çaba harcamışlardır. Gulliksen (1950), test kuramcılarının karşılaştığı en büyük sorunu kullanılan örneklemden farklı yetenek dağılımlarının, farklı parametre değerleri elde edilmesine yol açması olarak ifade etmiştir. Hambleton ve Swaminathan (1989), klasik test kuramlarının iki temel sorununa değinmektedir. Birinci kuramsal problem, gerçek puan, denekler ve test puanlarının birbirine bağımlı olması ve ikinci olarak da güvenirliğin üzerinde çalışılan denek popülasyonuna bağımlı olmasıdır. Bu sorunların üstesinden gelmek için gösterilen çabalar, bugün Madde Cevap Kuramı ya da Örtük Özellikler Kuramı adı altında anılabilen, modern test geliştirme modellerine yol açmıştır. Klasik test kuramında, parametre tahminleri üzerinde çalışılan örnekleme bağımlı iken, madde cevap modellerinde, denek parametresi ve madde parametreleri birbirinden bağımsızdır. Klasik test istatistikleri test puanlarını temel alırken, modern kuramlar madde puanlarını temel almaktadır (Mellenbergh, 1996).

Madde cevap kuramı, kişilerin gözlenen tepkilerinin psikolojik kuramlarda içerilen kuramsal yapılara, olasılıksal bir yolla bağlanmasını sağlamaktadır. Diğer bir deyişle, kişilerin ve maddenin belirli özelliklerine göre, kişinin bir maddeye göstereceği belirli bir tepkinin olasılığını veren matematiksel fonksiyonları içermektedir. Madde cevap kuramı, kişinin belirli bir test maddesine belirli bir tepkiyi gösterme olasılığının, kişinin test maddelerinin altında yatan örtük özellik üzerindeki konumu ile bağlantılı olduğunu varsaymakta ve bu doğrudan gözlenemeyen özellikle, gözlenen tepkiler arasında bağlantı kurmaktadır. Madde cevap modellerinin üzerine inşa edildiği temel kavram Madde Karakteristik Eğrisi'dir (Item Characteristic Curve veya Item Trace Line). Madde karakteristik eğrisi yoluyla, ölçülmekte olan özellik üzerinde birbirinden farklı noktalarda 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.

Bir test uygulaması yapıldığında, araştırmacının gözleyebildiği test puanları içinde her zaman bir miktar hata bulunmaktadır. Burada önem taşıyan sorun, gözlenen puanların gerçek puanlar ile ne kadar ilişkili olduğudur. Gerçek ve gözlenen puanlar arasındaki korelasyona Güvenirlik İndeksi denilmektedir ve klasik kuramda gerçek puanların standart sapmasının, gözlenen puanların standart sapmasına oranı olarak ifade edilmektedir. Güvenirlik katsayısı test puanlarının tutarlılığına ve mükemmeliyetine ilişkin bilgi sağlamakta ve puanlardaki bireysel farklılıkların ne kadarının ölçülmek istenen gerçek farklılıklardan, ne kadarının şans hatalarından kaynaklandığını göstermektedir (Anastasi, 1988).

Klasik kuramda gerçek puan kavramı, testin paralel formlarının ya da çok sayıda tekrarının ortalaması olarak kabul edilmektedir. Ancak gerçek yaşamda bu verilere ulaşılması mümkün olmadığından, güvenirlik katsayılarına bazı tahminleme yöntemleri ile ulaşılabilmektedir.

Klasik kuramda, deneysel verilerden hareketle ölçme sonuçlarının güvenirliğinin tahminlenmesinde, iç tutarlılık, test-tekrar test ve paralel formlar gibi yöntemlerden yararlanılmaktadır. Güvenirliği tahmin etmenin tek ve en iyi bir yöntemi yoktur. Farklı yöntemler bize gözlediğimiz test puanlarındaki varyasyonun ne kadarının, hangi tür hata kaynaklarına atfedilebileceğine ilişkin farklı bilgiler sağlamaktadır.

Test-tekrar test yöntemi, değişen koşulların, aradan geçen zamanın uygulayıcının ya da testi alanın değişmesi ile test sonuçlarına yansıyabilecek hata varyansı hakkında bilgi sağlarken, paralel formlar yöntemi, temelde testin kapsamından kaynaklanabilecek hata varyansı ile ilgilidir. Bir dizi test maddesi ile ölçmek istediğimiz bilgi ve beceriler bu formun alternatifi olan bir dizi madde için değişiklik gösteriyorsa bu durum, bir örneklem olarak düşünebileceğimiz test maddelerindeki performanstan hareket-

le genelleme yapılmak istenen alanın iyi temsil edilemediğine işaret etmektedir. İki-yarım yöntemi ise, testin iki yarısından elde edilen formların aynı tür bir özelliği ölçüp ölçmediğine ilişkin bilgi sağlayan bir iç tutarlılık yöntemidir. Ancak testi her zaman farklı şekillerde iki yarıya bölmek mümkün olduğundan, farklı sonuçlar elde edilebilmektedir. Test maddelerine verilen tepkilerin iç tutarlılığının çok daha hassas bir göstergesi ise, genel formülü ile Cronbach Alpha, yalnızca iki değerli (dichotomous) olarak puanlanabilen maddeler için ise Kuder-Richardson iç tutarlılık katsayısıdır (Cronbach, 1990). İç tutarlılığın yüksek olması ve ölçülecek boyutun homojen maddelerden oluşması psikolojik testler için hayati diyebileceğimiz bir öneme sahiptir. Homojen testler, tüm maddeleri aynı özelliği ölçen testlerdir ve biz eğer kişileri "belirli bir özellik boyutu" üzerinde karşılaştırmak istiyorsak, iç tutarlılığı yüksek olan homojen testler oluşturmak kaçınılmazdır.

Test maddelerinin her biri, aslında ilgili özelliğin bir yansımasını taşıyan ayrı ayrı değişkenler olarak düşünülebilir. Bu anlamda maddelere verilen tepkilerin toplanması ile elde edilen toplam test puanı da bir "bileşik puan"dır. Dolayısıyla, gözlediğimiz toplam puanların varyansının içinde tek tek değişkenlerin varyanslarının yanısıra, değişkenler arasındaki kovaryanslar da yer almaktadır (Hayduk, 1987). İç tutarlılık katsayıları, test puanlarının toplam varyansı içinde, bu kovaryans terimlerinin oranını araştıran güvenirlik tahminleridir. Maddelerin birbirleriyle ilişkisiz olması durumunda, kovaryans terimleri en uç noktada '0' olmakta ve toplam test varyansı yalnızca tek tek maddelerin kendi varyanslarının toplamına eşit olmaktadır. Bu durumda elde edilen güvenirlik katsayısı da '0' olacaktır.

Görüldüğü gibi, klasik kuramın en çok kullanılan test istatistiklerinden olan, iç tutarlılık güvenirlik katsayıları bize yalnızca testin güvenirliği değil, aynı zamanda yapı geçerliği hakkında da bilgi sağlamaktadır. Ancak çeşitli yazarlar (örn, Nunnaly,1978; Steinberg & Thissen, 1995), test puanlarının tutarlı bir tepki örüntüsü yaratmasının, maddelerin tek bir faktörü ölçtüğü anlamına gelmeyeceğine, birbiriyle ilişkili farklı faktörlerin toplamının yüksek bir iç tutarlılık katsayısı sağlayabileceğine işaret etmektedirler. Klasik güvenirlik yöntemleri, aşağıda daha ayrıntılı tartışılacağı üzere, test maddelerinin ölçülen boyut üzerinde farklı yerlerde bulunan bireyler için ne kadar hassas bilgi sağladığına ilişkin ayrıntılı veriler sağlamakta yetersiz kalmaktadır.

Tek bir gözlenen puanın, kişinin gerçek puanının zayıf bir tahmini olabileceğini düşündüğümüzde, gereksinim duyduğumuz kavram, ölçmenin standart hatasıdır. Güvenirlik katsayısı ve test standart sapmasından hareketle elde edilen ölçmenin standart hatası, kişilerin gözlenen puanlarının gerçek puanları etrafında göstermesi beklenen dağılımı tahmin etme imkanı sağlar. Ölçmenin standart hatası, örneklemdeki ortalama bir bireyin, belirli bir güven aralığı içerisinde gözlenen puanının gerçek puan etrafındaki dağılımına ilişkin olasılıksal bir bilgi verir. Ancak Ferguson'un (1982) değindiği gibi, klasik kuramda ölçmenin standart hatası bir tür ortalama değerdir. Bu nedenle klasik kuramda ölçmenin standart hatası belirli bir puana uygulandığında, ancak bu puanlar ortalamaya yakın olduğu oranda anlamlıdır. Ferguson, klasik kuramda ölçmenin standart hatasının puan dağılımının ortalarında daha büyük olduğunu, uçlara doğru gidildikçe küçüldüğünü belirtmektedir. Oysa madde-cevap kuramında ölçmenin standart hatası kavramı, test maddelerinin belirli bir yetenek grubu için sağladığı bilginin miktarına dayandırılmakta ve yetenek ölçeği üzerindeki her noktada ayrı ayrı hesaplanabilmektedir. Klasik modelde güvenirlik ölçümlerinin hepsi, üzerinde çalışılan örnekleme bağımlı sonuçlar veren yöntemlerdir. Yine ölçmenin standart hatası da güvenirlik tahminlerinden hareketle hesaplanmaktadır. Klasik test kuramında güvenirlik bir denek populasyonu için genel (unconditional) olarak tahminlenirken, madde cevap modellerinde ölçülen özellik boyutunun farklı bölgeleri için durumsaldır (conditional). Örneğin ölçülen özelliği bir yetenek boyutu olarak düşünürsek, farklı yetenek düzeyinde bulunan kişiler için farklıdır.

Hambleton ve Swaminathan (1989), madde cevap kuramında, test bilgi (information) fonksiyon-

larının klasik kuramdaki güvenirlik ve ölçmenin standart hatası kavramlarına karşılık geldiğini belirtmektedirler. Test bilgi fonksiyonlarının, testle yapılan yetenek tahminlerinin mükemmeliyetinin bir ölçüsü olabileceğini gösteren iki neden vardır. Bunlardan birincisi, şeklinin tamamen testte içerilen maddelere dayanması; ikincisi ise, yetenek düzeylerinin her biri için ayrı ayrı ölçme hatası vermesidir.

Herhangi bir test geliştirildiğinde ölçmeye yöneldiği bir hedef grup vardır. Test maddelerinin bu hedef grubun göstermesi beklenen dağılımı ölçmeye uygun maddelerden oluşması ya da bu dağılımdan uzakta kalan maddeleri içermesi, test sonuçlarının bu grup için bize sağlayacağı bilginin kalitesini etkiler. Test maddelerinin seçimindeki bu kaygılar, teknik olarak "Bilgi" terimi ile ifade edilmektedir. Bilgi kavramı, belirli bir gözlem yoluyla elde edilen ölçümün hassasiyeti, mükemmeliyeti hakkında bir değer yargısına varmak söz konusu olduğunda önem kazanmaktadır. Klasik kuramlarda ölçmenin hassasiyeti, genel bir güvenirlik ve tüm grup için ortak olan ölçmenin standart hatası kavramları ile ifade edilmektedir. Oysa madde cevap kuramında ölçmenin hassasiyeti, maddelerin sayısı ile her maddenin belirli bir kişiye göre olan durumuna bağlıdır. Madde ve kişi birbirine yakın olduklarında, örneğin tam hedefte olduğunda, madde kişinin ölçümüne, uzak olduğu durumdan daha fazla katkıda bulunur. Fark arttıkça maddenin etkinliği azalır ve aynı mükemmeliyette ölçüm yapmak için daha fazla sayıda madde gerekir. Bir test, belirli bir özelliği ölçen ölçek üzerindeki farklı bölgelerde bulunan kişiler için farklı düzeylerde bilgi sağlamaktadır. Daha hassas ayırımların yapılabildiği bölgelerdeki kişiler için test daha fazla bilgi sağlamakta ve bir testin sağladığı bilgi maddelerin karakteristik eğrilerine dayanmaktadır (Crocker ve Algina,1986).

Aşağıda Test Bilgi Fonksiyonu' nun eşitliği verilmiştir:

$$I(\theta) = \sum_{i=1}^{N} \{[P_i'(\theta)]^2 / P_i(\theta) \cdot Q_i(\theta)\}$$

(Hambleton ve Swaminathan, 1989)

Bu eşitlikte, I = Bilgi (information)

θ = Ölçülen özellik (latent trait)

$P_i(\theta)$= Belirli bir θ düzeyindeki kişilerin *i* maddesini olumlu (yetenek testlerinde 'doğru') cevaplama olasılığı.

$Q_i(\theta) = 1 - P_i(\theta)$

$P_i'(\theta$ ise belirli bir θ düzeyindeki i maddesi için madde karekteristik eğrisinin türevi (eğimi) dir.

Belirli bir θ düzeyi için madde bilgilerinin, testteki maddelerin tümü için toplanması ile Test Bilgi Fonksiyonu elde edilmektedir.

Ölçmenin standart hatası kavramı burada önem taşımaktadır. Belirli bir θ seviyesi için sağlanan bilgi, ölçmenin standart hatasının karesi ile ters orantılıdır. Test ölçümlerinin mükemmeliyeti, ölçmenin standart hatasının küçülmesi oranında artmaktadır. Maddelerin güçlüklerinin tam hedefe uygun dağılmaları durumunda, standart hataların dağılımını en küçüğe indirmek olanaklıdır. Yani bilgi belirli bir θ seviyesindeki standart hata ile ters orantılıdır. Standart hata küçüldükçe bilgi artmaktadır.

Lord (1980), modern kuramın klasik kuramdan tamamen farklı olmadığını, ancak klasik kuramın temelleri üzerinde geliştirilen yeni modellerin, klasik kuramın çözemediği bir çok sorunun üstesinden geldiğini ifade etmektedir. Mellenbergh'in (1996) de belirttiği gibi, test puanlarına dayalı klasik modeller ile madde puanlarına dayalı modern modeller arasında birçok ortak nokta vardır. Örneğin her iki modelde de gözlenen puanlar ve puanların beklenen değerleri üzerinde çalışılmaktadır. Her iki modelde de güvenirlik kavramı üzerinde düşünülmektedir. Aslında klasik kuramdaki ayırdetme indeksi olan madde toplam puan korelasyonu ile madde cevap kuramındaki ayırdetme indeksi arasında doğrudan bir ilişki vardır ve klasik kuramdaki gerçek puanı madde karakteristik eğrileri terimleri içerisinde ifade etmek mümkündür. Crocker ve Algina (1986), bu ilişkiyi aşağıdaki formülle ifade etmişlerdir.

$$T = \sum_{g} P_g(\theta)$$

Bu formüle göre gerçek puan, örtük özelliğin

doğrusal olmayan bir dönüştürmesidir. Ancak klasik test kuramında güvenirlik bir denek populasyonu için genel olarak tahminlenirken, madde cevap modellerinde ölçülen özellik boyutunun farklı bölgeleri için durumsaldır.

Somer (1996), klasik ve modern test kuramlarındaki ölçmenin standart hatası, güvenirlik ve test bilgi fonksiyonu kavramları arasındaki ilişkileri, tek parametreli Rasch modele dayalı olarak analiz edilen bir örnek üzerinde incelemiştir. Somer'in bir diğer çalışmasında ise (1998), iki kategorili olarak puanlanan (dichotomous) maddeler için uygulanabilecek iki parametreli madde cevap modeli ile yürütülen madde analizlerine örnekler bulmak mümkündür. Ayrıca modern kurama dayalı olarak yürütülen madde analizlerinin ölçek oluşturma sürecindeki katkıları konusunda yurdumuzda ve yurt dışında yayınlanmış pekçok çalışma vardır. Ancak kişilik ve tutum ölçeklerinde çoğunlukla çok kategorili (polytomous) olan maddelerle çalışılmaktadır. Madde-cevap kuramının bu tür maddelere uygun olan modelleri nispeten daha yenidir ve madde cevap modellerinin kişilik alanındaki uygulamalarına duyulan ilgi son yıllarda giderek artmaktadır (örn: Cooke, Michie, Hart & Hare, 1999; De Ayala, 1993; Kim & Pilkonis, 1999; Reise & Widaman, 1999; Zickar, 1998; Zickar & Highhouse, 1998). Bu çalışma Somer'in sözü edilen iki makalesindeki konuların bir uzantısı olarak düşünülmüş ve çok kategorili maddelere sahip olan bir kişilik alt ölçeği, klasik modelle ve madde-cevap modellerinden, iki parametreli modelle analiz edilerek, güvenirlik, test bilgi eğrileri ve standart hata kavramları üzerinde modellerin ölçek oluşturma sürecindeki katkıları tartışılmıştır.

Yöntem

Örneklem

Bu çalışmada kullanılan veriler Goldberg'in (1999) kişilik ölçekleri üzerine yaptığı bir çalışmada kullanılan 501 kişilik yetişkin Amerikan örnekleminden alınmıştır.

Veri Toplama Araçları

Çalışmada, beş faktör modeli kişilik boyutlarından Yumuşak Başlılık/ Uzlaşılabilirlik (Agreeableness) boyutunu ölçen 20 maddelik bir ölçek kullanılmıştır. Kendini değerlendirme türündeki ölçeğin maddeleri, Likert tipinde olup "tamamiyle yanlış (1)" tan, "tamamiyle doğru (5)" ya kadar uzanan beş basamaklı olarak derecelenen maddelerdir.

Analizler

Verilerin analizinde klasik test kuramına dayalı madde ve güvenirlik analizleri için SPSS istatistik programı ve iki parametreli madde cevap modeline dayalı analizler için MULTILOG (Thissen, 1991) programı kullanılmıştır.

Kullanılan ölçeğin maddeleri, Likert tipinde olduğu için analizler Samejima'nın (1997) "ağırlıklandırılmış cevaplar[1] modeli" (graded response model) temelinde gerçekleştirilmiştir. Samejima'nın ağırlıklandırılmış cevaplar modeli, sıralı çoklu kategorileri olan maddeler (ordered polytomous items) ile çalışmaya uygun olan bir madde cevap modelidir. Bütün tek boyutlu madde cevap modellerinde olduğu gibi, ağırlıklandırılmış cevaplar modeli de, bir ölçekteki maddelere gösterilen tepkilerin altında, belirleyici tek bir özelliğin bulunduğunu varsaymaktadır. u tane cevap kategorisi olan maddelerde (u=1.2....m) ağırlıklandırılmış tepki modeli, u 'ıncı kategorinin olasılığını aşağıdaki şekilde formüle eder:

$$P_u(\theta) = P^*_u(\theta) - P^*_{(u+1)}(\theta) \quad \text{(Samejima,1997)}$$

$$P^*_u(\theta) = \frac{1}{1 + \exp[-a(\theta - b_{u-1})]}$$

$$P^*_{(u+1)}(\theta) = \frac{1}{1 + \exp[-a(\theta - b_u)]}$$

Formülde: $P_u(\theta)$, x cevabının u 'ıncı kategoriye ilişkin olasılığını vermektedir. a yine madde ayırdetme parametresini, ve b_u da, u kategorisinin örtük özellik boyutu üzerindeki konumunu (location, threshold) göstermektedir. Homojen modeller-

[1] *Erden (1997) çalışmasında "graded response" terimini "ağırlıklandırılmış cevaplar" olarak Türkçe'leştirmiştir. Terim birliği oluşması açısından sadık kalınmıştır.*

de (Samejima, 1997), maddenin ayırdetme parametresi *a* tüm cevap kategorileri için aynı ve tektir (ancak farklı maddelerde farklı değerleri almaktadır). Model yukarıdaki formulasyonlardan anlaşılabileceği gibi, maddenin her kategorisi için ayrı bir *b* değeri vermektedir. Böylece Multilog ile yürütülen madde analizi çıktılarında, her madde için bir a değeri, (u -1) kadar *b* değeri elde edilmektedir. Burada dikkat edilmesi gereken bir nokta, her kategorinin farklı *b* değerlerine sahip olması ve bu farklı kategorilerin θ ölçeği üzerinde farklı bölgelerde işlevsel olduğuna işaret etmesidir. Örneğin 1 (tamamiyle doğru) kategorisi çok düşük θ düzeyleri için etkili iken yani bu düzeylerde seçilme olasılığı yüksek iken, bu olasılık, θ değeri ilerledikçe düşmekte, bu bölgeler için bu kategori işlevini yitirmektedir. Bu noktada 2. kategori (biraz doğru) önem kazanmakta bu kategori de belirli bir θ'daki kişilerce yüksek oranda tercih edilirken, yine θ ilerledikçe bu olasılık azalmaktadır. Yani kategori karakteristik eğrileri, yalnızca iki değerli olarak puanlanabilen maddelerde olduğu gibi monotonik olarak ilerleyen eğriler değil, idealde çan şekli gösteren eğrilerdir. Multilog ayrıca, her madde için ayrı ayrı madde bilgi fonksiyonlarının yanı sıra, toplam ölçeğin bilgi ve ölçmenin standart hatasına ilişkin fonksiyonlarını da sağlamaktadır. Bu konular, bulgular bölümünde örnek üzerinde tartışılmaktadır.

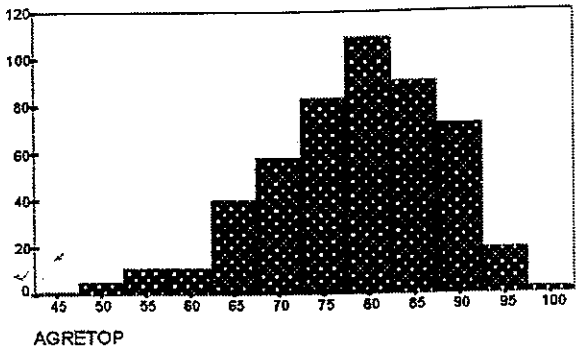
Bulgular ve Tartışma

Klasik test kuramına dayalı madde analizleri sonucunda elde edilen her madde için ortalama, standart sapma, madde-toplam puan korelasyonları, ölçek iç tutarlılık güvenirlik katsayısı (alpha) ve madde çıkarıldığında ölçek alfaları Tablo 1'de verilmiştir. Maddelerin ayırdetme gücünün göstergesi olan madde-toplam puan korelasyonları incelendiğinde 7. maddenin genelde madde analizlerinde kabul edilen yeterlik düzeyi olan .30'a çok yakın olduğu, bunun dışındaki maddelerin hepsinin ayırdedicilik değerlerinin yeterli düzeyde olduğu görülmektedir. Ölçeğin güvenirlik katsayısı (α =

Tablo 1

Klasik Madde ve Güvenirlik Analizi Sonuçları

Maddeler	Madde Ortalaması	Madde Çıktığında Ölçek Ortalaması	Madde Çıktığında Ölçek Varyansı	Madde-Ölçek Toplam Korelasyonu	Madde Çıktığında Ölçek Alpha'sı
Madde 1	3.59	74.78	84.28	.48	.84
Madde 2	3.90	74.48	86.05	.44	.84
Madde 3	4.52	73.85	87.45	.47	.84
Madde 4	4.02	74.36	85.14	.46	.84
Madde 5	3.96	74.42	85.38	.46	.84
Madde 6	4.27	74.10	86.89	.39	.84
Madde 7	3.57	74.81	87.20	.29	.84
Madde 8	4.15	74.22	85.42	.45	.84
Madde 9	3.57	74.81	84.64	.37	.84
Madde 10	3.60	74.78	83.70	.39	.84
Madde 11	3.67	74.69	81.73	.44	.84
Madde 12	4.60	73.78	85.47	.51	.84
Madde 13	3.57	74.80	82.30	.47	.84
Madde 14	4.30	74.08	84.35	.47	.84
Madde 15	4.67	73.70	86.23	.54	.84
Madde 16	3.72	74.65	82.98	.40	.84
Madde 17	3.78	74.60	83.42	.46	.84
Madde 18	3.59	74.78	83.14	.42	.84
Madde 19	3.59	74.78	82.48	.44	.84
Madde20	3.74	74.64	84.08	.41	.84



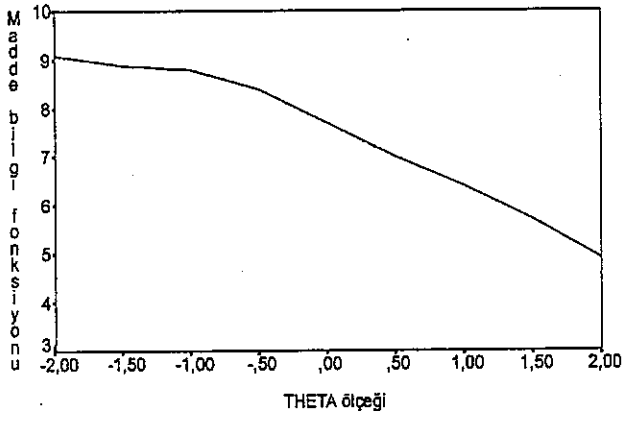
Şekil 1. Ham puanların frekans dağılımı

.84) da bize ölçeğin kişilik ölçekleri düşünüldüğünde oldukça yüksek olarak kabul edilebilecek bir iç tutarlılığa sahip olduğunu göstermektedir.

Klasik kurama dayalı olarak test geliştiren birçok araştırmacı bu noktaya gelindiğinde testin kalitesini yeterli olarak görmektedir. Çünkü testin iç tutarlılığı yüksektir ve maddeler de beklenen yönde bu tutarlılığa yeterli düzeyde katkıda bulunmaktadırlar. Ancak toplam test puanlarının frekans dağılımına baktığımızda (Şekil 1) bunun negatif kaymalı bir dağılım olduğu görülmektedir. Üzerinde çalışılan örneklemin, amaçlanan grubun temsili bir örneklemi olduğunu varsaydığımızda buradan çıkarabileceğimiz sonuç, testin toplam puanının bu özellikte düşük olan kişileri daha etkili olarak ayırdettiği, yüksek olan kişileri ise birbirinden daha az etkili olarak ayırdedebildiği ve yüksek puanlarda bir miktar yığılmaya yol açtığıdır. Bu sorun toplumda istenen, beğenilen özellikleri ölçen testlerin ortak bir sorunudur ve tutum ve kişilik ölçeklerinde sıklıkla yaşanmaktadır. Oysa biz psikolojik bir özelliği ölçtüğümüzde, özellikle de bu patolojik değil normal bir kişilik özelliği ise, istenen bu özellikte yüksek ya da düşük olsun her yöndeki kişilerin birbirinden ayırdedilebilmesi ve bu boyut üzerinde bu farklılıkları görebileceğimiz şekilde ölçeklenebilmesidir.

Eğer Tablo 1'de sunduğumuz yüksek kaliteli madde analizi ve güvenirlik sonuçlarına rağmen ölçeğin toplam puanı ile ulaşılabilen dağılımdan rahatsızlık duyulur ise, elimizdeki klasik analizlere dayalı bilgilerle daha fazla ne yapılabilecektir? Hangi maddeler bu probleme yol açmaktadır? Hangi maddeler kişilik boyunun hangi bölgelerinde bulunan kişileri daha etkin olarak ölçmemize katkıda bulunmaktadır? Hangi maddeleri değiştirirsek normale daha yakın bir dağılım elde edebiliriz? Belki bir yol maddelerin ortalamalarına bakmaktır. Ancak Tablo1'den görülmektedir ki maddelerin ortalamaları 3.5 ile 4.5 arasında değişmektedir ve ayırdetme güçleri de oldukça birbirine yakındır. Bu veriler de bize hangi tür maddeler üzerinde revizyona gitmemiz gerektiği konusunda fazla bir bilgi sağlamamaktadır. Bunun nedeni klasik test kuramına dayalı olarak elde edilen madde parametrelerinin bize, maddenin, populasyonun geneli üzerindeki etkisi hakkında toplam bir bilgi sağlaması, ancak ölçülen boyut üzerinde *farklı bölgelerde* bulunan kişilerin ölçümlerine katkısı hakkında durumsal (conditional) bir ilave bilgi sağlamamasıdır.

Kanımızca bu noktada, modern test modelleri ile yürütülecek madde analizleri, özellikle de madde-bilgi ve test bilgi fonksiyonu kavramları test geliştirmekte olan araştırmacıya yararlı ipuçları sağlayacaktır. Tablo 2'de iki-parametreli ağırlıklandırılmış cevaplar modeliyle yürütülen madde analizi sonuçları, Şekil 2'de ise test bilgi eğrisi verilmiştir. Tablo 2'de, analizler bölümünde açıklandığı gibi, her madde için madde ayırdetme parametresi değeri (a), her cevap kategorisi için ayrı ayrı (b) değerleri ve madde düzeyinde bilgi fonksiyonlarının yanısıra son satırda toplam test puanları için,



Şekil 2. 20 Maddelik ölçeğin test bilgi eğrisi

Tablo 2

Ağırlıklandırılmış Cevaplar Modeli Madde Analizi Sonuçları

Madde 1		A	B(1)	B(2)	B(3)	B(4)				
	tahmin	1.15	-4.07	-1.99	-.46	2.19				
	θ ölçeği	-2.0	-1.5	-1.0	-.5	.0	.5	1.0	1.5	2.0
	I (θ) bilgi (inf.)	.39	.39	.39	.38	.36	.33	.32	.34	.35
Madde 2		A	B(1)	B(2)	B(3)	B(4)				
	tahmin	1.18	-4.84	-2.67	-1.41	1.74				
	θ ölçeği	-2.0	-1.5	-1.0	-.5	.0	.5	1.0	1.5	2.0
	I (θ) bilgi (inf.)	.42	.40	.36	.32	.28	.29	.32	.35	.34
Madde 3		A	B(1)	B(2)	B(3)	B(4)				
	tahmin	1.63	-6.47	-3.66	-2.73	-.17				
	θ ölçeği	-2.0	-1.5	-1.0	-.5	.0	.5	1.0	1.5	2.0
	I (θ) bilgi (inf.)	.55	.47	.52	.65	.66	.50	.30	.15	.07
Madde 4		A	B(1)	B(2)	B(3)	B(4)				
	tahmin	1.17	-4.91	-2.49	-1.60	1.12				
	θ ölçeği	-2.0	-1.5	-1.0	-.5	.0	.5	1.0	1.5	2.0
	I (θ) bilgi (inf.)	.41	.39	.36	.33	.32	.34	.36	.33	.27
Madde 5		A	B(1)	B(2)	B(3)	B(4)				
	tahmin	1.38	-4.12	-2.46	-1.22	1.19				
	θ ölçeği	-2.0	-1.5	-1.0	-.5	.0	.5	1.0	1.5	2.0
	I (θ)bilgi (inf.)	.57	.56	.52	.46	.43	.46	.49	.46	.35
Madde 6		A	B(1)	B(2)	B(3)	B(4)				
	tahmin	1.26	-4.62	-3.40	-1.96	.37				
	θ ölçeği	-2.0	-1.5	-1.0	-.5	.0	.5	1.0	1.5	2.0
	I (θ)bilgi (inf.)	.45	.42	.40	.40	.41	.40	.34	.25	.16
Madde 7		A	B(1)	B(2)	B(3)	B(4)				
	tahmin	.59	-6.25	-2.92	-1.29	4.59				
	θ ölçeği	-2.0	-1.5	-1.0	-.5	.0	.5	1.0	1.5	2.0
	I (θ) bilgi (inf.)	.10	.10	.10	.09	.09	.08	.08	.08	.08
Madde 8		A	B(1)	B(2)	B(3)	B(4)				
	tahmin	1.30	-4.11	-2.68	-1.79	.70				
	θ ölçeği	-2.0	-1.5	-1.0	-.5	.0	.5	1.0	1.5	2.0
	I (θ) bilgi (inf.)	.50	.46	.41	.39	.41	.44	.41	.33	.22
Madde 9		A	B(1)	B(2)	B(3)	B(4)				
	tahmin	.87	-4.03	-1.80	-.86	2.33				
	θ ölçeği	-2.0	-1.5	-1.0	-.5	.0	.5	1.0	1.5	2.0
	I (θ) bilgi (inf.)	.23	.23	.23	.22	.20	.20	.19	.20	.20

Tablo 2'nin devamı

Madde 10		A	B(1)	B(2)	B(3)	B(4)				
	tahmin	.86	-3.54	-1.86	-.85	1.92				
	θ ölçeği	-2.0	-1.5	-1.0	-.5	.0	.5	1.0	1.5	2.0
	I (θ) bilgi (inf.)	.23	.23	.23	.22	.21	.21	.21	.20	.20
Madde 11										
	tahmin	1.07	-3.62	-1.24	-.56	.87				
	θ ölçeği	-2.0	-1.5	-1.0	-.5	.0	.5	1.0	1.5	2.0
	I (θ) bilgi (inf.)	.32	.33	.35	.35	.35	.34	.31	.26	.20
Madde 12		A	B(1)	B(2)	B(3)	B(4)				
	tahmin	1.63	-4.57	-2.75	-1.94	-.77				
	θ ölçeği	-2.0	-1.5	-1.0	-.5	.0	.5	1.0	1.5	2.0
	I (θ) bilgi (inf.)	.82	.79	.76	.66	.46	.26	.13	.06	.03
Madde 13		A	B(1)	B(2)	B(3)	B(4)				
	tahmin	.98	-3.87	-1.53	-.42	1.55				
	θ ölçeği	-2.0	-1.5	-1.0	-.5	.0	.5	1.0	1.5	2.0
	I (θ) bilgi (inf.)	.28	.29	.30	.30	.29	.28	.28	.27	.24
Madde 14		A	B(1)	B(2)	B(3)	B(4)				
	tahmin	1.16	-4.66	-2.73	-1.73	-.08				
	θ ölçeği	-2.0	-1.5	-1.0	-.5	.0	.5	1.0	1.5	2.0
	I (θ) bilgi (inf.)	.42	.41	.40	.39	.37	.31	.24	.16	.10
Madde 15		A	B(1)	B(2)	B(3)	B(4)				
	tahmin	1.95	-5.61	-3.01	-2.00	-.82				
	θ ölçeği	-2.0	-1.5	-1.0	-.5	.0	.5	1.0	1.5	2.0
	I (θ) bilgi (inf.)	1.11	1.07	1.05	.89	.53	.25	.10	.04	.02
Madde 16		A	B(1)	B(2)	B(3)	B(4)				
	tahmin	.91	-3.50	-1.88	-.72	1.02				
	θ ölçeği	-2.0	-1.5	-1.0	-.5	.0	.5	1.0	1.5	2.0
	I (θ) bilgi (inf.)	.26	.26	.26	.26	.25	.25	.23	.21	.18
Madde 17		A	B(1)	B(2)	B(3)	B(4)				
	tahmin	1.06	-4.78	-2.00	-.73	1.20				
	θ ölçeği	-2.0	-1.5	-1.0	-.5	.0	.5	1.0	1.5	2.0
	I (θ) bilgi (inf.)	.33	.34	.35	.34	.33	.33	.32	.29	.24
Madde 18		A	B(1)	B(2)	B(3)	B(4)				
	tahmin	.92	-4.73	-1.50	-.50	1.52				
	θ ölçeği	-2.0	-1.5	-1.0	-.5	.0	.5	1.0	1.5	2.0
	I (θ) bilgi (inf.)	.23	.25	.26	.26	.26	.25	.24	.23	.21

Tablo 2'nin devamı

Madde 19		A	B(1)	B(2)	B(3)	B(4)				
	tahmin	.92	-4.34	-1.45	-.58	1.51				
	q ölçeği	-2.0	-1.5	-1.0	-.5	.0	.5	1.0	1.5	2.0
	I (q) bilgi (inf.)	.24	.25	.26	.26	.25	.25	.24	.23	.21
Madde 20		A	B(1)	B(2)	B(3)	B(4)				
	tahmin	.92	-4.60	-2.18	-.70	1.36				
	q ölçeği	-2.0	-1.5	-1.0	-.5	.0	.5	1.0	1.5	2.0
	I (q) bilgi (inf.)	.26	.26	.26	.26	.25	.25	.24	.23	.20
Toplam test										
bilgi fonksiyonu	θ ölçeği	-2.0	-1.5	-1.0	-.5	.0	.5	1.0	1.5	2.0
(20 madde)	I (θ)bilgi (inf.)	9.1	8.9	8.8	8.4	7.7	7.0	6.4	5.7	4.9
	standart hata (θ)	.33	.33	.34	.34	.36	.38	.40	.42	.45
Toplam test										
bilgi fonksiyonu	θ ölçeği	-2.0	-1.5	-1.0	-.5	.0	.5	1.0	1.5	2.0
(10 madde)	I (θ) bilgi (inf.)	4.3	4.4	4.4	4.3	4.2	4.1	4.1	3.9	3.6
	standart hata (θ)	.48	.48	.48	.48	.49	.49	.49	.51	.53

bilgi ve standart hata değerleri verilmiştir.

Tablo 2'nin son sütunu ve Şekil 2 incelendiğinde, test puanlarının "Yumuşak Başlılık" boyutunda yüksek olan kişiler için, (θ ölçeği üzerinde, - 2.0'den başlayarak 2.0'ye kadar test bilgi fonksiyonu, 9.1, 8.9, 8.8, 8.4, 7.7, 7.0, 6.4, 5.7, 4.9 değerlerini aldığı görülmektedir. Değerlerde, negatif uçtan pozitif uca doğru gidildikçe azalma görülmekte ve bu durum test puanlarının yüksek puanlı kişiler için düşük olan kişilere göre daha az bilgi sağladığı ve bu bölgelerde ölçmenin standart hatalarının daha yüksek olduğu görülmektedir. Bu noktada yararlı olabilecek analizler madde düzeyindeki bilgi fonksiyonlarıdır (bkz. Tablo 2). Tablo incelendiğinde 7. maddenin, klasik analizlerdeki sonuçlara (α = .29) paralel olarak oldukça düşük bir ayırdetme parametresine (a = .59) sahip olduğu görülmektedir (genellikle 1.0 ve üzerindeki değerlere sahip maddeler yeterli düzeyde ayırdedici maddeler olarak kabul görmektedir (Hulin, Drasgow & Parsons, 1983). Bu maddenin bilgi fonksiyonu incelendiğinde her θ düzeyi için (-2.0 ve 2.0 arasında) çok düşük bilgi düzeylerine ulaştığı (.10 ile .08 arasında) görülmektedir. Maddeler için analiz sonuçlarını genel olarak incelediğimizde, örneğin 1. maddenin ayırdetme gücünün yüksek olduğu (1.15) ve tüm θ düzeyleri için (-2.0 ve 2.0 arasında) yaklaşık olarak eşit düzeylerde bilgi sağladığı (.39 ile .34 arasında) görülmektedir. Yani bu madde Yumuşak Başlılık boyutunda her düzeydeki kişiler için yaklaşık aynı düzeyde bir ayırdetmeye katkıda bulunmaktadır. Bu maddenin klasik analizlerdeki madde-toplam korelasyonu da (r = .47) yüksektir. Ancak 15. madde incelendiğinde, bu maddenin ayırdetme parametresinin hem klasik analizlerde (.53) hem de madde cevap modeli analizinde yüksek (a = 1.95) olmasına rağmen, bu ayırdediciliğin daha çok yumuşak başlılık özelliğinde ortalamaya yakın ve daha aşağıya doğru olan kişiler için çok iyi olduğu (.66 ile .55 arasında) görülmekte, ortalamadan pozitif yöne doğru uzaklaştıkça (.50, .30, .15 ve giderek .07'ye) düşmektedir. Yani bu bölgelerdeki kişileri ayırdetmekte bu maddenin katkısı çok düşüktür. Bu durum, maddenin kalitesinin düşüklüğüne işaret etmemektedir. Madde belirli θ seviyeleri için oldukça yüksek düzeyde ayırdedici

bir maddedir ancak bu özelliği ile düşük puanlı kişileri ayırdetmeye katkıda bulunmaktadır. Zaten maddenin ayırdedicilik değerinin oldukça yüksek olması bu maddenin θ üzerinde belirli bölşelerde çok iyi ayırdetme işlevine rağmen diğer bölgeler için etkili olamamasını da beraberinde getirmektedir. Testte bu durumu dengelemek üzere, karşı uçta da yüksek düzeyde ayırdedici maddelere ihtiyaç vardır. Diğer test maddelerinin ayırdediciliğe katkıda bulunduğu bölgeleri, bilgi fonksiyonlarından incelediğimizde test maddelerinin çoğunluğunda (özellikle de yoğun olarak 3, 6, 8,12, 14'üncü maddelerde) yine bu tür bir özellik gözlemekteyiz. Yani test maddelerinin ayırdediciliği çoğunlukla düşük puanlı kişiler için yüksekken, yüksek puanlara doğru gidildikçe azalmaktadır. Bu da toplam test puanlarının bilgi ve standart hata değerlerinde kendini göstermektedir.

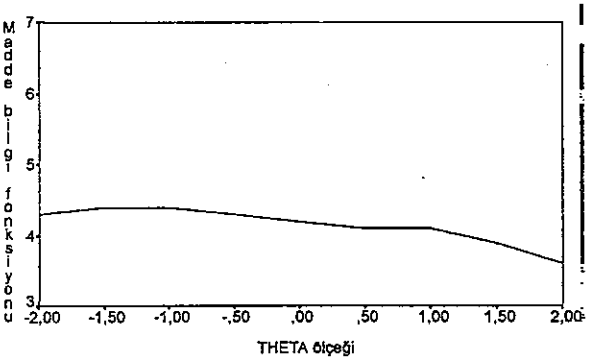
Bu çalışmada halihazırda geliştirilmiş 20 maddelik bir ölçek üzerinde çalıştığımızdan, üst özellik düzeyindeki kişiler için daha çok bilgi sağlayan maddelerin değiştirilmesini deneme ve test ve madde bilgi fonksiyonundan yararlanarak dağılım üzerinde elde edilecek değişiklik ve avantajları ampirik verilerle gösterme imkanımız kısıtlıdır. Ancak elimizdeki verilerden hareketle, yine de örtük özellik boyutunun tüm bölgelerinde yakın düzeyde bilgi sağlayan maddelerin seçimiyle oluşturulacak bir kısa form üzerinde test bilgi eğrisinin şeklinde oluşturulabilecek değişikliği gözlemek mümkündür.

Bu amaçla maddelerin bilgi fonksiyonlarını incelediğimizde, 1, 5, 9, 10, 13, 16, 17, 18, 19, 20. maddelerin, örtük özellik boyutunun yaklaşık tüm düzeyleri için daha homojen bilgi düzeyleri sağladığını görmekteyiz. 20 madde yerine bu 10 madde ile bir form oluşturduğumuzda elde edilen test bilgi eğrisi Şekil 3'de verilmiştir.

Madde sayısının yarıya inmesi nedeniyle beklendiği şekilde klasik güvenirlik katsayısında (α = .74) bir azalma görülmektedir. Yine madde sayısının azalmasına bağlı olarak her θ düzeyindeki test bilgi değerlerinde de azalma gözlenmektedir. Zira test bilgi düzeyi, doğrudan doğruya madde bilgilerinin toplanması ile elde edilmektedir ve dolayısıyla az ya da çok katkıda bulunsun, her madde ilavesi test bilgi düzeyini arttırmaktadır. Ancak amacımız bir kısa form oluşturup onun etkinliğinin incelenmesi değil de, yalnızca farklı düzeylerinde ayırdedici maddelerin yer alması ile test bilgi dağılımında elde edilecek değişiklikleri gözlemek olduğundan madde sayısının azalması ile ortaya çıkan test bilgi değerleri ve güvenirlikte ortaya çıkan bu düşüşler göz ardı edilmektedir.

On maddelik formdan elde edilen test bilgi fonksiyonu incelendiğinde (Şekil 3), bilgi eğrisinin -2.0 ve 2.0 arasında nispeten düzgün bir dağılım gösterdiği, Şekil 2'deki 20 maddelik ölçeğin bilgi eğrisinin ise açık bir şekilde yüksek puanlara gidildikçe alçaldığı görülmektedir. Amacımız hedef grup için her bölgede etkili olan bir bilgi fonksiyonu elde etmek olduğundan maddelerin kalitesi hakkında ayrıntılı verilere sahip olunduğunda, istenen bölgeleri daha etkin biçimde ayırdeden maddelerle testi oluşturmanın olası olduğunu görmekteyiz.

Bu analizlerden şu sonucu çıkarabiliriz. Test maddelerinin yalnızca yüksek düzeyde ayırdetme özelliğine sahip olması (klasik ve modern analizlerin her ikisinde de) ve iç tutarlılığı yükseltmeye katkıda bulunması bizim amacımıza uygun bir ölçek oluşturmamızı garanti etmemektedir. Bu durumda yapılabilecek şey, modern analizlerdeki madde bilgi fonksiyonlarını inceleyerek, düşük puanlarda ayırdedici olan maddelerin testte kapsandığı kadar, yüksek puanları ayırdedici maddelerin de bulunarak testin kapsamına alınmasıdır. Böylece hedef grubu her düzeyde ayırdedebilen,



Şekil 3. 10 maddelik ölçeğin test bilgi eğrisi

daha hassas ve ölçme hatası daha az ölçekler oluşturmak olasıdır. Lord (1980), bunu en iyi şekilde sağlayabilmek için, maddelerin kademeli olarak teste girip çıkarılarak hedefe en uygun dağılımın elde edilebildiği bir sistem önermektedir.

Sonuç

Bu makalede bir kişilik ölçeği geliştirmekte olan bir araştırmacının klasik kurama dayalı madde analizleri ile elde ettiği sonuçların yanı sıra modern test kuramlarına dayalı analizlerle sağlayabileceği yararlar üzerinde durulmuştur.

Ölçmenin dakikliği, mükemmeliyeti söz konusu olduğunda, yukarıdaki verilerin ve tartışmaların ışığı altında modern kurama dayalı madde ve güvenirlik analizlerinin, klasik kuramla sağlanan verilere ilave bazı bilgiler sağladığı sonucu çıkarılabilir. Bu nedenle bir testin geliştirilmesinde ve güvenirliğinin sağlanmasında yalnızca klasik madde analizi yöntemlerine dayanan güvenirlik hesaplamalarının yetersizlikleri göz önünde bulundurularak test maddelerinin seçimi ve yapılandırılmasında, test ve madde bilgi eğrilerinden yararlanılması, test ölçümlerinin doğruluğunun sağlanmasında ve ölçme hatalarının azaltılmasında önemli katkılar sağlayabilmektedir. Hambleton, Swaminathan ve Rogers (1991), test bilgi fonksiyonlarının aynı yetenek boyutunu ölçen birden fazla testin etkililiğini karşılaştırmadaki yararına da dikkat çekmektedirler. Test bilgilerinin birbirine oranlanması ile, belirli bir θ seviyesi için bir testin diğerine göre ne kadar daha hassas bir ölçüm sağladığına ve diğer testin seviyesine ulaşmak için, testin ne kadar uzatılması gerektiğine dair bilgi sağlanabilmektedir. Böylece bilgi eğrileri, ölçülmesi hedeflenen grubu en iyi temsil eden maddelerin seçimine olanak sağlayarak, güvenirlikleri daha yüksek, hedef grup için daha çok bilgi sağlayan ve ölçmenin standart hatasının daha düşük olduğu testlerin elde edilmesine olanak vermektedir. Bilgi kavramı test puanlarının farklı bölgelerdeki mükemmeliyetini belirtirken, klasik güvenirlik kavramı test puanlarının genel olarak mükemmeliyetinin bir göstergesidir. Bu özellikleriyle kanımızca, her iki analiz test puanlarının mükemmeliyeti hakkında birbirini tamamlayıcı bilgiler sağlamaktadırlar.

Ancak bu noktadan sonra sorulabilecek bir soru veya tartışılabilecek bir nokta akla gelmektedir. Test geliştirmede klasik kuramlara göre daha karmaşık istatistiksel yöntemleri gerektiren modern analizleri kullanarak testin ölçülen boyut üzerindeki kişiler arası ayırdediciliğinin yükseltilmesi, testin gerçek yaşam durumlarına ilişkin yapabileceği yordamalarda bir ilerleme sağlamakta mıdır? Zira modern test analizi modellerinin kullanılması araştırmacılar yönünden ciddi teknik çabaları gerektirmekte, kullanılması ve yorumlanmasındaki güçlükler nedeniyle testin maliyetini arttırmaktadır. Bu yanıtlanması oldukça zor bir sorudur ve modellerin karşılaştırılmasında kapsamlı yordama geçerliği çalışmalarını gerektirmektedir. Kanımızca bu maliyet sorununa verilebilecek bir cevap, klasik test kuramlarına dayalı analizlerlerden elde edilen sonuçlarla, istenen özellikte maddelere ve hedef gruplara uygun bir dağılıma ulaşılabiliyorsa daha ileri karmaşık analizlerin gereksiz olacağıdır. Çünkü birçok araştırmacı (örn., Baykul, 1980; Berberoğlu, 1989; Erden, 1997; Fan, 1998) klasik ve modern kuramlara dayalı analizlerle elde edilen sonuçların birbirine yakın olduğunu ve büyük ölçüde aynı maddelerin seçimine dair ipuçları verdiklerini bildirmektedir. Ancak bu çalışmada verilen örnekteki gibi hedef gruba ilişkin dağılım sorunları ile karşılaşıldığında, klasik madde analizi sonuçları yeterli gözükse de sonuçta ulaşılan ölçek amaca tam uygun olmayabilecektir. Bu noktada modern analizlerin kullanılması bir gereklilik gibi gözükmektedir.

Güvenirlik konusundaki katkılarının yanısıra modern test analizleri, ölçülen boyut üzerinde farklı bölgelerde yeralan farklı kişilerin maddelere nasıl tepki vereceklerine ilişkin olasılıksal bir bilgi sağlamakla, maddelerin ölçülen boyutla ilişkisini daha net bir biçimde anlamamıza yardımcı olmakta ve bu yolla "yapı geçerliği" daha yüksek ölçekler oluşturmaya da imkan vermektedir.

Kaynaklar

Anastasi, A. (1988). *Psychological testing.* (6th ed.) New York: MacMillan Publishing Company.

Baykul, Y. (1980). Örtük özellikler ve klasik test kuramları üzerine bir karşılaştırma. *Doğa,* 33-40.

Berberoğlu, G. (1989). Erişi testlerine madde seçiminde klasik test kuramı ve Rash modelinin karşılaştırılması. Eğitim ve Bilim, 13 (74), 64-69.

Crocker, L., & Algina, J., (1986). *Introduction to classical and modern test theory.* New York: CBS College Publishing.

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.

Cronbach, L. J. (1990). *Essentials of psychological testing.* New York: Harper Collins Pub. Inc.

De Ayala, R. J. (1993). An introduction to polytomous item response theory models. *Measurement and Evaluation in Counseling and Development, 23,* 172-189.

Erden-Başarır, D. (1997). Örtük özellikler ve klasik test teorisi yaklaşımına dayalı olarak geliştirilen Likert tutum ölçeğinin psikometrik özelliklerinin karşılaştırılması. *Yayınlanmamış Doktora Tezi.* Ankara: Hacettepe Üniversitesi.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58* (3), 357-381.

Ferguson, G. A. (1982). *Statistical analysis in psychology and education.* Tokyo: McGraw-Hill Kosaido Printing Co.Ltd.

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several Five-Factor models. In I., Mervielde, F. Deary, F. De Fruyt, & F. Ostendorf. *Personality Psychology in Europe.* Vol. 7, (pp. 7-28). Tilburg Netherlands: Tilburg University Press.

Gulliksen, H. (1950). *Theory of mental tests.* New York: John Wiley & Sons, Inc.

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

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

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

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

Kim, Y., & Pilkonis, P. A (1999). Selecting the most informative items in the IIP scales for personality disorders: An application of item response theory. *Journal of Personality Disorders, 13*(2), 157-174.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* New Jersey: Lawrance Erlbaum.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods. 1* (3), 293-299.

Nunnally, J. C. (1978). *Psychometric theory.* New York: McGraw-Hill, Inc.

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

Somer, O. (1996). Klasik ve modern test kuramlarında standart hata, güvenirlik ve informasyon kavramlarının karşılaştırılması. *Psikoloji Seminer, 11,* 96-105.

Somer, O. (1998). Kişilik testlerinde klasik ve modern test kuramları ile madde analizi. *Türk Psikoloji Dergisi, 13*(41), 1-17.

Steinberg, L., & Thissen, D. (1995). Item response theory in personality research. In P. E. Shrout, & T. Fiske, *Personality research, methods, and theory.* (pp.161-181). New-Jersey: Lawrence Earlbaum.

Samejima F. (1997). Graded response model. In, Van der Linden, W. J., Hambleton, R. K. *Handbook of modern item response theory.* New York: Springer-Ferlag.

Thissen D. (1991). *MULTILOG user's guide.* Version 6.0. Chicago: Scientific Software Inc.

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

Zickar, M. J., & Highhouse, S. (1998). Looking closer at the effects of framing on risky choice: An item response theory analysis. *Organizational Behavior and Human Decision Processes, 75*(1), 75-91.

Türk Psikoloji Dergisi
1999, 14(44), 77-78

Summary

Comparison of Reliability and Information Concepts of Classical and Modern TestTheory with Polytomous Item Tests

Oya Somer *
Ege Üniversitesi

Data obtained from psychological measures always include some error variance. Hence, the score obtained from one administration is only one of the observable values from its possible score distribution. Reliability and standard error of measurement enable test users to estimate the confidence interval which contain the person's true score. In modern test theory, the values of standart error of measurement provide different values for individuals who are at different regions on the measured latent trait. Furthermore, the item and test information functions in modern test developing models offer an alternative way of constructing more reliable psychological tests. In this study, a personality scale which has polytomous items was analyzed with both the Classical Test Theory and Item-Response Theory (IRT).

Method

Participants and Measures

The analysis reported here was based on the data, obtained from 501 American adults (Goldberg, 1999). Participants described themselves on an Agreeableness scale (one of the Big-Five dimensions) developed by Goldberg. The scale consist of 20 items with a 5-point rating scale, ranging from extremely accurate to extremely inaccurate.

Analyses

Item analysis was performed by a computer program (MULTILOG) which is appropriate for IRT based analysis of polytomous items. Two parameters logistic model was used for estimating item parameters. Samejima's (1997) Graded Response Model was used for the item analysis which is appropriate for ordered polytomous items. The two parameter Graded Response Model used in this study is simply generalization of the two parameter logistic model for dichotomous items and was specified by Samejima as:

$$P_u(\theta) = P^*_u(\theta) - P^*_{(u+1)}(\theta) \quad \text{(Samejima,1997)}$$

$$P^*_u(\theta) = \frac{1}{1 + \exp[-a(\theta - b_{u-1})]}$$

$$P^*_{(u+1)}(\theta) = \frac{1}{1 + \exp[-a(\theta - b_u)]}$$

Results and Discussion

The results of the classical item analysis are presented in Table 1, and the results of IRT analysis are presented in Table 2. As can be seen in Table 1, item-total correlations (item discrimination parame-

* Address for correspondence : Assoc. Prof. Oya Somer, Ege Üniversitesi, Edebiyat Fakültesi Psikoloji Bölümü 35100 Bornova-İZMİR,TURKEY
E.mail : osomer@edebiyat.ege.edu.tr

ter in classical theory) of all the items, except item 7, were high enough for constructing a homogenous scale, and the internal consistency reliability coefficient of the scale was considerably high (a = .84). These results suggested that items were adequately discriminating between high and low scorers on the Agreeableness dimension. However, it is not clear in which region of the latent trait the items were working more effectively. To differentiate more precisely in certain regions of latent trait, IRT concepts such as item and test information functions need to be used. As seen in Table 2, the results related to these concepts can help one make a much more extensive and interesting interpretations about the items and the scale in general. When the information functions of the items in Table 2 are examined it can be seen that for most of the items information values were high on the negative pole of the latent trait. However, these values decreased on the positive pole at the latent trait, resulting in a negatively skewed frequency distribution of the total scores (See Figure 1). Thus, the information concept enables us to evaluate the discrimination power of the scale much more easily and to adapt it to our target population. Because the IRT analyses supply different error of measurement along the continuum of the latent trait (in contrast, classical analysis gives only one average error of measurement value) by increasing information on certain regions it is now possible to construct more reliable scales with less error. So it can be concluded that when one needs to differentiate more precisely on certain regions of the latent trait, IRT concepts such as item and test information functions can be more efficient than classical item parameters.

Conclusion

The application of the IRT measurement models affords several advantages over the Classical Test Theory based psychometric procedures. Especially item and test information functions of IRT supply superior opportunities for constructing precise tests over the classical reliability concept. In addition, sample free and model based nature of the IRT analyses solve problem of invariant parameter estimation which had been the basic question of psychometricians for long years.

In spite of its strengths, however, the IRT based test construction has some disadvantages. One of them is the strong assumptions required for the nature of the data compared to classical procedures. These features constrain the applicability of IRT in some fields especially related to personality measurement. A second broad limitation is the highly technical and expensive nature of the IRT analyses. Over the past decade a great deal of softwares became available for interested researchers but the IRT still requires highly specialized training.