Elementary Education Online, 7(3), 758-768, 2008.
İlköğretim Online, 7(3), 758-768, 2008. [Online]: http://ilkogretim-online.org.tr

Comparison of Validity and Reliability of Two Tests Developed by Classical Test Theory and Item Response Theory [1]

Ümit ÇELEN[2]

ABSTRACT. In constructing the test process after preliminary test try outs two tests have been developed by Classical Test and Item Response Theories. The aim of the research is to compare psychometric characteristics of the item selection of these two tests. The data has been collected from a study group selected from 6th, 7th and 8th grades of elementary school, by a verbal ability test developed by the researcher. All of the research findings support the "another way to reach the same results through different paths" idea of Item Response Theory. In the light of the findings, in order to construct a valid and reliable instrument, the important thing is not selecting the theory, it is applying the necessities on the selected theory.

Key Words: classical test theory, item response theory, test construction

SUMMARY

Purpose and significance: The aim of the research is to compare psychometric characteristics of the item selection of two tests developed by Classical Test and Item Response Theories.

Methods: The data has been collected from a study group selected from 6th, 7th and 8th grades of elementary school, by a verbal ability test developed by the researcher. In order to construct verbal ability test which is to measure thinking by words and verbal implication, the necessities of both of the theories have been aimed to be met. The preliminary test try outs have been applied to 980 students, while final tests have been applied to 481 students. When the item selection for each behavior is made according to selecting items by determined measures, 5 most appropriate items out of 10 items, 27 of the 40 items are the same for the final test.

Results: Strong correlations (0.95-0.96) have been found between the test scores and ability estimations. While the level of grade increases, the increase in the score and ability estimation of two tests are also similar. The correlation of test scores developed by Classical Test Theory and the course of Turkish Language is 0.67; while the correlation is 0.65 for Item Response Theory. When the coefficients of reliability of the scores obtained from the two tests are analyzed, it is seen that the reliability of the two tests are in parallel.

Discussion and Conclusions: All of the research findings support the "another way to reach the same results through different paths" idea of Item Response Theory. In the light of the findings, in order to construct a valid and reliable instrument, the important thing is not selecting the theory, it is applying the necessities on the selected theory.

[1] Has been prepared from the thesis titled "Comparison of Psychometric Characteristics of Two Tests Developed by Classical Test Theory and Item Response Theory".

[2] Ph.D., Ankara University, Faculty of Educational Sciences, umitcelen@yahoo.com

Klasik Test Kuramı ve Madde Tepki Kuramı Yöntemleriyle Geliştirilen İki Testin Geçerlilik ve Güvenilirliğinin Karşılaştırılması[3]

Ümit ÇELEN[4]

ÖZ. Bir temel araştırma olarak planlanan bu çalışmada bir test geliştirme sürecinde deneme uygulaması aşamasından sonra madde seçimi Klasik Test Kuramı ve Madde Tepki Kuramı doğrultusunda yapılarak geliştirilen iki testin psikometrik özelliklerinin karşılaştırılması amaçlanmıştır. Araştırmada veriler ilköğretim 6, 7 ve 8. sınıf öğrencilerinden seçilerek oluşturulan çalışma grubundan, araştırmacı tarafından geliştirilen sözel yetenek testi aracılığıyla toplanmıştır. Madde seçiminin iki kurama göre ayrı ayrı yapılmasıyla oluşan nihai testlerin uygulamasından elde edilen sonuçlar karşılaştırılmış, elde edilen bulguların, Madde Tepki Kuramı'nın "farklı yollardan aynı sonuçlara ulaşmanın bir diğer yolu" olduğu görüşünü destekler nitelikte olduğu görülmüştür. Bu bulgular ışığında geçerli ve güvenilir bir araç geliştirmek için seçilecek kuramın değil, seçilen kuramın gereklerini yerine getirmenin önemli olduğu vurgulanmıştır.

Anahtar Sözcükler: klasik test kuramı, madde tepki kuramı, test geliştirme

GİRİŞ

Ölçme, bir betimleme işidir. Geniş anlamda ölçme, belli bir nesnenin ya da nesnelerin belli bir özelliğe sahip olup olmadığının, sahipse sahip oluş derecesinin gözlenip, gözlem sonuçlarının sembollerle ve özellikle sayı sembolleriyle ifade edilmesidir (Tekin, 1996). Değerlendirme ise, ölçme sonuçlarını bir ölçüte vurarak, ölçülen nitelik hakkında bir değer yargısına varma sürecidir (Turgut, 1990). Ölçmelerde gözlenen bir özelliğin gerçek değeri bulunmak istense de ölçmeye karışan çeşitli hatalar yüzünden gerçek değer doğrudan elde edilemez; gözlenen ölçme sonuçları yardımıyla kestirilmeye çalışılır. Kestirmeyi yapabilmek için bazı istatistiksel kuramlar geliştirilmiştir (Baykul, 2000). Ölçme tarihi incelendiğinde iki temel kurama rastlanmaktadır. Bunlardan kronolojik olarak daha önce olanı "Klasik Test Kuramı"dır. Bu kuramın bazı sınırlılıklarına alternatif olarak XX. yüzyılın ortalarında adını duyuran yeni bir kuram ortaya çıkmıştır. Bu kuram "Örtük Özellikler Kuramı" veya "Madde Tepki Kuramı" olarak adlandırılmaktadır (Crocker ve Algina, 1986). Psikolojik ölçme tarihinin başlangıcında itibaren test geliştirme, analiz ve psikolojik ölçeklerin puanlanmasında daha yaygın olarak kullanılan kuram Klasik Test Kuramı'dır. Günümüzde de Klasik Test Kuramı yaygın olarak kullanılıyor olmakla birlikte, Madde Tepki Kuramı giderek daha popüler ve tercih edilir olmaya başlamıştır (Hambleton, 1994; Reise, Ainsworth ve Haviland, 2005).

Madde Tepki Kuramı'na göre, bireylerin belli bir alandaki doğrudan gözlenemeyen yetenekleri ya da özellikleri ya da bu alanı yoklayan sorulardan oluşan test maddelerine verdikleri yanıtlar arasında bir ilişki vardır ve bu ilişki matematiksel olarak ifade edilebilir. Bu kurama göre geliştirilen testlerden elde edilen yetenek ölçüleri, bireye uygulanan testlerden bağımsız olarak elde edilebilmektedir. Diğer bir deyişle, veri seti ile seçilen model arasında makul bir uyum sağlandığında, Madde Tepki Kuramı modelleri değişmez madde parametreleri ve yetenek kestirimleri elde etmemizi sağlar. Bunu yanı sıra yeteneği bilinen bir yanıtlayıcının, maddeyi uygulamadan da parametreleri bilinen herhangi bir soruyu doğru yanıtlandırma olasılığı tahmin edilebilir. Klasik Test Kuramı'nda madde güçlük ve ayırıcılık gücü indeksleri testin geliştirildiği gruptan kestirilmekte ve grup değiştiğinde değerler de değişebilmektedir. Madde Tepki Kuramı gruptan bağımsız (sample free) ve değişmez parametre kestirimleri yapma iddiasındadır. Değişmezlik hem aynı özelliği ölçmeye yönelik

[3] "Klasik Test Kuramı ve Madde Tepki Kuramına Dayalı Olarak Geliştirilen İki Testin Psikometrik Özelliklerinin Karşılaştırılması" adlı doktora tezi çalışmasından hazırlanmıştır.

[4] Dr. Ümit ÇELEN, Ankara Üniversitesi Eğitim Bilimleri Fakültesi, umitcelen@yahoo.com

olarak hazırlanmış olan farklı maddelere verilen tepkilere dayalı olarak kestirilen yetenek parametrelerinin değişmezliği (test free) hem de aynı testin farklı bireylere uygulanmasıyla elde edilen madde parametrelerinin değişmezliği olarak ele alınabilir. Bu durum, bir testin Madde Tepki Kuramı'na göre bir kez ölçeklendikten sonra maddelerin özelliği değişmediğinden, pek çok kez kullanılmasına olanak sağlar. Ancak bu değişmezliğin sağlanması, madde parametrelerinin elde edilmesinde yapılan deneme uygulamasının ve bu uygulamanın yapıldığı grubun bazı şartları sağlamasına bağlıdır (Hambleton, 1990; Hambleton ve diğerleri 1991; Hambleton ve Swaminathan, 1985; Kelecioğlu, 2001). Klasik Test Kuramı'nda ise bireylerin aldıkları puanlar o testin güçlük düzeyine göre değişmektedir (Lord ve Novic, 1968).

Klasik Test Kuramı'nda ölçme hataları tüm grup için hesaplanırken, Madde Tepki Kuramı'nda her birey için ayrı ayrı kestirilebilmektedir. Yine güvenilirlik katsayısı Klasik Test Kuramı'nda cevaplayıcı grubun puan dağılımı için tek bir değer olarak hesaplanırken, Madde Tepki Kuramı'nda her bir madde ve yetenek düzeyi için güvenilirlik madde ve test bilgi fonksiyonu şeklinde hesaplanabilmektedir. Klasik Test Kuramı'nda elde edilen tek katsayı, güvenilirliğin farklı yetenek düzeyleri için değişmediği anlamına gelmekte iken, tekrarlı ölçmelerle hesaplanan güvenilirlik katsayılarına bakıldığında, bunların ölçülen özelliğe üst düzeyde sahip bireyler için daha yüksek olduğu görülmektedir. Bu durum farklı yetenek düzeyindeki bireyler için ölçme aracının aynı düzeyde güvenilirliğe sahip olamayacağını göstermektedir (Nartgün, 2002).

Madde Tepki Kuramı'nın test geliştirme, soru bankası oluşturma, bireye uyarlanmış test geliştirme, madde yanlılığının belirlenmesi, seçenekleri ağırlıklandırma ve test eşitleme konularında karşılaşılan sorunlara çözüm getirdiği iddia edilmektedir (Hambelton ve Swaminathan, 1985).

Her iki kuram da test maddelerine verilen tepkilerin doğruluğu veya yanlışlığı üzerine odaklanmaktadır. Normallik varsayımı her iki kuram için de söz konusudur. Test puanlarının normalliği sağlandığında, iki kurama göre elde edilen ayırıcılık güçleri ve madde güçlük indeksleri arasında geçiş sağlamak mümkün olabilmektedir (Lord ve Novick, 1968; Crocker ve Algina, 1986). Belirli bir özelliği ölçmek için geliştirilmiş testin sadece o özelliği ölçmesi ve testteki bir maddeye verilen yanıtın diğerini etkilememesi gerekliliği de her iki kuram için karşılanması gereken varsayımlar olarak karşımıza çıkmaktadır (Linn, 1998; Çıkrıkçı-Demirtaşlı, 1998). Bununla birlikte, maddelere verilen tepkilerin değerlendirilmesi ve test geliştirme bakımından aralarında bazı farklılıklar vardır (Harvey ve Hammer, 1999).

Madde Tepki Kuramı, madde düzeyindeki bilgileri, Klasik Test Kuramı'na göre daha fazla temel almaktadır. Buna göre, her bir madde için kestirilen madde parametreleri ile hangi maddelerin hangi yetenek düzeylerinde daha ayırıcı olduğu, ölçme aracının hangi yetenek düzeyindeki bireyler için daha tutarlı ölçme sonuçları verdiği, farklı güçlük düzeyindeki maddelerin doğru yanıtlandırılabilmesi için gerekli olan yetenek düzeyleri Madde Tepki Kuramı'nda açık bir biçimde kestirilebilmektedir (Nartgün, 2002).

Klasik Test Kuramı'na göre test geliştirirken deneme uygulaması sonrasında ayırıcılık gücü yüksek olan ve madde varyansını en yüksek yapabilmek için orta güçlükteki maddeler nihai teste alınır. Madde Tepki Kuramı'nda ise madde parametrelerinin yanı sıra asıl ölçüt modele uyumluluktur. Bunun yanı sıra her bir maddenin teste olan katkısını her yetenek düzeyi için gösteren madde bilgi fonksiyonu madde seçiminde kriter olarak alınabilmektedir (Linden ve Hambleton, 1997).

Yukarıda benzer ve farklı yönleri irdelenen iki test kuramından birinin diğerine üstün olup olmadığı yapılan birçok çalışmaya rağmen halen tartışma konusu olmaya devam etmektedir. İki kuramın karşılaştırılmasına yönelik olarak yapılan bu çalışmada da aşağıdaki cümlede ifade edilen probleme yanıt aranmaya çalışılmıştır.

Deneme uygulaması aşamasından sonra madde seçimi Klasik Test Kuramı ve Madde Tepki Kuramı'na göre ayrı ayrı yapılarak geliştirilecek iki testin nihai uygulamasıyla elde edilecek test puanları ve yetenek kestirimlerinden hangisinin geçerliliği ve güvenilirliği yüksek olacaktır.

YÖNTEM

Çalışma Grubu: Bir temel araştırma olarak planlanan çalışmada veriler Amasya ili Merzifon ilçesinde 2005-2006 öğretim yılında ilköğretim okulları 6, 7 ve 8. sınıflarında öğrenim gören öğrencilerden elde edilmiştir. Geliştirilen testlerin deneme uygulamasında 980, nihai uygulamalarda ise 481 öğrenci yer almıştır

Kullanılan İstatistiksel Teknikler: Test puanları dağılımlarının normalliğinin test edilmesinde Kolmogrov Simirnov (KS) testi kullanılmıştır. Tek boyutluluğun test edilmesinde kullanılan faktör analizi, tetrakorik korelâsyon matrisine dayalı temel bileşenler analizidir. Bağımsız gruplardan elde edilen test puanları ve madde istatistik ve parametreleri ortalamalarını karşılaştırmak için t-testi ve tek faktörlü varyans analizi tekniklerinden yararlanılmıştır. Bu işlemler yapılırken t testleri için normallik varsayımı ve varyans analizi için normallik varsayımı yanı sıra varyansların homojenliği varsayımı test edilmiştir. İki değişken arasındaki ilişkinin incelenmesinde Pearson momentler çarpımı korelâsyon tekniği kullanılmıştır.

Madde Tepki Kuramı'na göre kestirilen yetenek ölçüleri ve Klasik Test Kuramı'na göre hesaplanan test puanlarının aynı ölçek üzerinde incelenmesine olanak sağlamak amacıyla her iki değişken de ortalaması 50, standart sapması 10 olan T puanlarına dönüştürülmüşlerdir. Test puanı ve yetenek kestirimlerinde sınıf düzeylerine göre meydana gelen farklılığın iki kurama göre değişip değişmediğini test etmek için "karışık ölçümler için iki faktörlü varyans analizi" tekniği kullanılmak istenmiş ancak "ölçüm setlerinin ikili kombinasyonları için grupların kovaryanslarının eşit olması" varsayımı (Büyüköztürk, 2007) sağlanamadığı için bu test uygulanamamıştır. Bu durumda ortak etkinin bir göstergesi olarak fark puanları karşılaştırılmıştır.

Test puanı ve yetenek kestirimi puanlarının geçerlilik katsayılarını karşılaştırmak amacıyla İki bağımlı korelâsyon katsayısı arasındaki farkın test edilmesi için Hotelling t testi (Mendeş ve Karabayır, 2003; Bruning ve Kintz, 1993) kullanılmıştır. Bu test değerleri herhangi bir paket program aracılığıyla değil, aşağıda sunulan formül kullanılarak hesaplanmıştır.

$$t = \frac{(r_{yx1} - r_{yx2})\sqrt{(N-3)(1+r_{x1x2})}}{\sqrt{2(1+2r_{yx1}r_{yx2}r_{x1x2} - r_{yx1}^2 - r_{yx2}^2 - r_{x1x2}^2)}}$$

Formülde; r_{yx1}: Y ile X1 arasındaki Pearson korelâsyon katsayısını, r_{yx2}: Y ile X2 arasındaki Pearson korelâsyon katsayısını, r_{x1x2} ise X1 ile X2 arasındaki Pearson korelâsyon katsayısını göstermektedir.

Araştırma verilerinin analizinde, Klasik Test Kuramı'na göre madde analizi için Iteman 3.50, tetrakorik korelasyon matrisine dayalı temel bileşenler faktör analizi için Statistica 5.0, Madde Tepki Kuramı'na göre madde ve test parametrelerinin ve yetenek kestirimlerinin belirlenmesinde Bilog 3.10, ve diğer betimsel istatistikler ve hipotez testleri için SPSS 12.0 paket programlarından yararlanılmıştır.

Ölçme Aracı: Araştırmada kullanılmak üzere geliştirilen test, ilköğretim 6, 7 ve 8. sınıf öğrencilerinin sözcüklerle düşünme ve sözel akıl yürütme yeteneklerini ölçmeyi amaçlayan "Sözel Yetenek Testi"dir. Testin geliştirilmesinde her iki kurama göre test geliştirme ilkeleri yerine getirilmeye çalışılmıştır. Testte ölçülecek davranışlar aşağıdaki gibi belirlenmiştir:

1. Eş-zıt anlamlı kelime bulma:
   a) Verilen bir kelimenin eş anlamlısı olan kelimeyi verilen seçenekler arasından bulup işaretleme.
   b) Verilen bir kelimenin zıt anlamlısı olan kelimeyi verilen seçenekler arasından bulup işaretleme.
2. Benzer ilişkiyi bulma: Verilen iki kelime arasındaki ilişkiyi belirleyerek, benzer bir ilişki taşıyan kelime çiftini seçenekler arasından bulup işaretleme.
3. Farklı olan kelimeyi bulma: Verilen seçenekler içerisinde diğerlerinden farklı olan kelimeyi bulup işaretleme.

4. Cümle tamamlama: Verilen bir cümlede boş bırakılan yere gelebilecek kelimeyi verilen seçenekler arasından bulup işaretleme.

5. Cümle kurma: Bir cümleye ait ve karışık sırada verilmiş olan kelimelerin doğru sıralanışının verildiği seçeneği bulup işaretleme.

6. Paragraf oluşturma: Karışık sırada verilmiş olan cümlelerin, anlamlı bir paragraf oluşturacak şekilde sırlanmış olduğu seçeneği bulup işaretleme.

7. Deyimlerin anlamlarını bulma:

a) Verilen cümlede kullanılmış olan deyimin anlamını verilen seçeneklerden bulup işaretleme.

b) Anlamı verilen deyimi seçenekler arasından bulup işaretleme.

8. Atasözlerinin anlamlarını bulma:

a) Verilen atasözü ile aynı anlamda kullanılan veya anlamı verilen atasözüne en yakın olan atasözünü seçeneklerden bulup işaretleme.

b) Anlamı verilen atasözünü seçenekler arasından bulup işaretleme.

Bu davranışların ölçülmesinde kullanılacak test maddelerinin yazımında içerik olarak, Türkçe Sözlük (Türk Dil Kurumu [TDK], 2005a), İlköğretim Okulları İçin Türkçe Sözlük (TDK, 2005b), İlköğretim Okulları İçin Yazım Kılavuzu (TDK, 2005c), Atasözleri ve Deyimler Sözlüğü (Kuşçu ve Kuşçu, 1997) kaynaklarından yararlanılmıştır.

Testte 4 seçenekli çoktan seçmeli madde formatı kullanılmıştır. Test kapsamında ölçülecek davranışların eşit ağırlıkta testte yer alması için her davranış için eşit sayıda madde yazılması uygun görülmüştür. Bu durumda, her davranış için 20 adet olmak üzere 160 madde yazılmıştır.

Maddeler araştırmacı tarafından bir ön elemeden geçirildikten sonra 120'ye indirilmiş ve konu alanı ve ölçme ve değerlendirme uzmanlarının görüşünü almak üzere uygun bir format haline getirilmiştir. Üç ölçme ve değerlendirme uzmanı ve iki Türkçe dersi öğretmeninden test maddelerini ölçme ilkeleri ve bilimsel doğruluk açısından incelemeleri istenmiştir. Gelen eleştiriler neticesinde maddeler her bölümde 10 madde olacak şekilde 80'e indirilmiştir.

Geliştirilen testin deneme uygulaması 20-28 Şubat 2006 tarihleri arasında 39 şubede bulunan 980, nihai formların uygulanması ise 20-24 Mart 2006 tarihleri arasında, 23 şubede bulunan 481 öğrenci ile gerçekleştirilmiştir.

Madde seçimine geçmeden önce deneme uygulaması verilerinin her iki kuramın varsayımlarını karşılayıp karşılamadığı kontrol edilmiştir. Dağılım istatistikleri ve testten elde edilen puanların histogram grafiği incelendiğinde normallik varsayımının karşılandığı görülmüştür. Ancak KS testi ile normalliğin sağlanmadığı sonucu elde edilmiştir. Normal dağılıma uygunluğun test edilmesi için uygulanan KS testinin çalışma grubu büyüdükçe dağılımın normal çıkmaması yönünde sonuçlar üretmesi nedeniyle test, grup rasgele olarak ikiye bölünmek suretiyle tekrar edilmiştir. Bu işlem sonucunda her iki grubun da normal dağılıma uygun olduğu görülmüştür (1. Grup için KS Z= 1.311, $p>0.05$; 2. Grup için KS Z= 1.180, $p>0.05$). Çarpıklık katsayının (-0.13) ±1 sınırları içinde kalmasının yanı sıra incelenen Q ve Q grafiğinde noktaların 45 derecelik çizgi üzerinde olması bulguları da dikkate alınarak test puanları dağılımının normallikten önemli ölçüde sapma göstermediği sonucuna varılmıştır.

Tek boyutluluk varsayımının kontrolü amacıyla yapılan maddeler arası tetrakorik korelâsyon matrisine dayalı temel bileşenler analizine göre 80 maddeden 69'unun en yüksek yük değerini 1. faktörde aldığı, birinci faktöre ait özdeğerle ikinci faktöre ait özdeğer arasında önemli bir fark olduğu, diğer faktörler arasında ise böyle bir farklılık bulunmadığı görülmüştür. Bu sonuca göre tek boyutluluk varsayımının karşılandığına karar verilmiştir. Yerel bağımsızlık varsayımı ayrıca test edilmemiş, tek boyutluluk sağlandığı için yerel bağımsızlığın da karşılandığı kabul edilmiştir. Testin madde tepki kuramının modellerine uyumu BILOG 3 (Mislevy ve Bock, 1990) programı yardımıyla incelenmiş ve bir ve iki parametreli modellerle uyum sağlamadığı gözlenmiştir. Üç parametreli modelle olan uyum katsayısı (kay kare)=606.70, $p=0.76$ olarak hesaplanmıştır. p değerinin 0.05'ten

büyük olması modelle veri arasındaki uyumu gösterdiğinden testin üç parametreli modele uygun olduğu sonucuna varılmıştır.

Deneme uygulaması sonrasında Klasik Test Kuramı'na göre madde analizi yapılmıştır. Bu analiz sonucunda, test kapsamında ölçülen her davranış için denenen 10 maddeden güçlük indeksi 0.10-0.90 aralığında olan ve ayırıcılık gücü en yüksek 5'i nihai teste seçilmiştir. Nihai teste seçilen maddelerin madde güçlük indeksleri incelendiğinde en düşüğü 0.33, en yükseği 0.86 olmak üzere ortalamasının 0.62; çift serili korelâsyon olarak hesaplanan ayırıcılık gücü indekslerinin ise en düşüğü 0.41, en yükseği 0.72 olmak üzere ortalamasının 0.59 olduğu görülmüştür.

Madde Tepki Kuramı'na göre madde seçiminde de, testin her bölümü için en uygun 5 maddenin nihai teste alınması esas kabul edilmiş, bu nedenle bütün bölümdeki maddeler modele uyumlarının yanı sıra a, b ve c parametreleri ve ortalama bilgi değerleri bakımından incelenmiştir. Modelle uyum sağlamayan 17. madde öncelikle elenmiş, geriye kalan 79 madde yukarıda sıralanan özellikler bakımından değerlendirilmiştir. a parametresi değerinin 1.00 değerine yakınlığı esas alınmış, a değeri 0.50'nin altında ve 1.50'nin üstünde olan maddeler nihai teste alınmamıştır. b parametresi değerinin ise 0.00'a olan yakınlığı tercih nedeni olmuş, -2.00'den küçük ve 2.00'den büyük olan maddeler elenmiştir. Böylece a parametresi değerleri 0.63 ile 1.45 arasında ve ortalaması 0.928; b parametresi değerleri -1.60 ile 1.91 arasında ve ortalaması -0.071 olan 40 madde Madde Tepki Kuramı'na göre nihai teste seçilmiştir.

Sonuç olarak, iki kurama göre oluşturulan 40 maddeli nihai testlerin 27'şer maddesinin ortak olduğu görülmüştür. Ortak maddelerin dışında 13'er madde her iki teste alınmış ve 27 madde her iki teste de alınmamıştır.

Testlerin nihai uygulamasından sonra, her iki kurama için deneme uygulaması verileriyle yapılan varsayım kontrolleri tekrar edilmiş ve varsayımların sağlanmadığına ilişkin bulgulara rastlanmamıştır.

BULGULAR

İki nihai testin uygulanmasıyla 481 öğrenciden elde edilen yanıtların doğru yanıta 1 yanlış ve boş yanıtlara 0 olacak şekilde puanlanması sonucunda elde edilen test puanları ve Madde Tepki Kuramı'na göre geliştirilen testten elde edilen yetenek kestirimlerinin betimsel istatistikleri Tablo 1'de sunulmuştur.

Tablo 1. Nihai Test Puanları ve Yetenek Kestirimlerinin Betimsel İstatistikleri (n=481)

İstatistik	Klasik Test Kuramı'na Göre Geliştirilen Test Puanları	Madde Tepki Kuramı'na Göre Geliştirilen Test Puanları	Madde Tepki Kuramı'na Göre Geliştirilen Testten Elde Edilen Yetenek Kestirimleri
$\bar{x}$	24.57	24.79	0.00
Ortanca	25.00	25.00	0.06
Tepe Değeri	32.00	32.00	-0.13
Standart Sapma	8.11	7.80	1.00
Standart Hata	0.37	0.36	0.05
Çarpıklık	-0.24	-0.27	-0.16
Basıklık	-0.74	-0.66	-0.33
Ranj	37.00	36.00	5.21
En Düşük	3.00	4.00	-2.72
En Yüksek	40.00	40.00	2.49

Tablo 1'de görüleceği gibi Klasik Test Kuramı'na göre geliştirilen testten elde edilen puanların aritmetik ortalaması 24.57; Madde Tepki Kuramı'na göre geliştirilen testten elde edilen puanların aritmetik ortalaması 24.79'dur. Bu testten elde edilen yetenek kestirimlerinin ise aritmetik ortalaması 0.00'dır. Her iki test puanının da ortancası 25.00, tepe değeri 32.00'dir. Yetenek kestirimlerinin ise ortancası 0.06, tepe değeri 0.13'tür. Klasik Test Kuramı'na göre geliştirilen testten elde edilen puanların ranjı 1 puan daha geniştir ve standart sapması 8.11 olarak hesaplanmıştır. Madde Tepki Kuramı'na göre geliştirilen testten elde edilen puanların standart sapması 7.80'dir. Madde Tepki Kuramı'na göre geliştirilen testten elde edilen yetenek kestirimleri ise -2.72 ile 2.49 arasında değerler almıştır. Çarpıklık ve basıklık katsayıları incelendiğinde her iki testten edilen puan ve Madde Tepki Kuramı'na göre geliştirilen testten elde edilen yetenek kestirimi değerleri dağılımının bir miktar çarpık ve basık olduğu görülmektedir. Ölçmenin standart hatası Klasik Test Kuramı'na göre geliştirilen test için 2.69; Madde Tepki Kuramı'na göre geliştirilen test için 2.67 olarak hesaplanmıştır.

Nihai testlerden elde edilen test puanı ve yetenek kestirimlerinin sınıf düzeylerine göre değişimi ve test puanı ve yetenek kestirimi arasındaki ilişkileri gösteren bulgular Tablo 2'de sunulmuştur.

Tablo 2. Nihai Test Puanı ve Yetenek Kestirimi Ortalamalarının Sınıf Düzeylerine Göre Karşılaştırılması

		Test Puanı*		**Yetenek Kestirimi***		**Fark** (Test Puanı-Yetenek Kestirimi)		**r**
Sınıf	**n**	$\bar{x}$	**S**	$\bar{x}$	**S**	$\bar{x}$	**S**	
6	179	48.28	10.27	48.50	9.76	-0.22	0.08	0.954**
7	152	49.65	9.42	49.23	10.00	0.42	3.01	0.954**
8	150	52.41	9.84	52.57	9.88	-0.16	2.78	0.960**
Genel	**481**	**50.00**	**10.00**	**50.00**	**10.00**	**0.00**	**2.98**	**0.956****
Analiz Sonucu		F=7.31, p<0.001		F=7.61, p<0.001		F=2.26, p>0.05		

* Her iki dağılım da T dağılımına dönüştürülmüştür.
** p<0.001

Tablo 2'de görülebileceği gibi, Klasik Test Kuramı'na göre geliştirilen testin puan ortalaması ile Madde Tepki Kuramı'na göre geliştirilen testin yetenek kestirimleri arasında hesaplanan korelâsyon katsayıları, 6. ve 7. sınıf düzeyinde 0.954; 8. sınıf düzeyinde 0.960 ve grubun tümünde ise 0.956 olarak bulunmuştur. Klasik Test Kuramı'na göre geliştirilen testin test puanlarında sınıf düzeyi arttıkça meydana gelen artışın, Madde Tepki Kuramı'na göre geliştirilen testten elde edilen yetenek puanlarında da benzer şekilde olduğu görülmektedir. Her iki karşılaştırma için yapılan varyans analizi sonucunda sınıf düzeylerine göre test puanı ortalamaları arasındaki fark istatistiksel olarak manidardır. Test puanı ve yetenek kestirimleri arasındaki farkların sınıf düzeylerine göre değişimi incelendiğinde istatistiksel olarak manidar bir fark bulunmadığı görülmektedir. Bu durum sınıf ve test türü ortak etkisinin olmadığını, diğer bir deyişle sınıf düzeylerine göre gözlenen farkların testlere göre değişmediğini göstermektedir.

6-7, 7-8 ve 6-8. sınıfların ortalamaları karşılaştırdığında ise yine benzer sonuçlara ulaşılmaktadır. Her iki test için de 6. ve 7. sınıfların ortalamaları arasında istatistiksel olarak manidar bir fark bulunmamıştır (t_1=-1.23; t_2=-0.66, p>0.05). 8. sınıfların ortalaması 6. sınıfların ortalamasından istatistiksel olarak manidar şekilde yüksektir (t_1=-3,71; t_2=-3.74, p<0.001). Yine benzer şekilde, 8. sınıfların her iki test için hesaplanan ortalamaları 7. sınıfların ortalamalarından istatistiksel olarak manidar şekilde yüksektir (t_1=-2,50, p<0.05; t_2=-2.92, p<0.01).

Öğrencilerin sözel yetenek testinden aldıkları puanların bir geçerlilik kanıtı olarak Türkçe dersi yılsonu başarı notları ve yılsonu başarı not ortalamasıyla olan ilişkileri incelenmiş ve bulgular Tablo 3'te sunulmuştur.

Her sınıf düzeyinde ve grubun genelinde bahsedilen notlarla her iki test ve yetenek puanları arasında pozitif yönlü ve istatistiksel olarak manidar korelâsyonlar bulunmuştur. Grubun tümünün Klasik Test Kuramı'na göre geliştirilen testten aldıkları puanlar ile Türkçe dersi başarı not ortalaması arasındaki korelâsyon r=0.669; Madde Tepki Kuramı'na göre geliştirilen teste ait yetenek kestirimleriyle Türkçe dersi not ortalamaları arasındaki korelâsyon ise r=0.652 olarak bulunmuştur. Klasik Test Kuramı'na göre geliştirilen testin geçerliliğinin daha yüksek olduğu görülmektedir ancak iki korelâsyon katsayısı arasındaki fark istatistiksel olarak manidar değildir. Benzer durum test puanları ve yetenek kestirimleri ile yılsonu başarı not ortalaması arasındaki ilişkilerde de söz konusudur ve burada da aradaki fark, Klasik Test Kuramı'na göre geliştirilen test lehine olmak üzere, 0.009'dur ve istatistiksel olarak manidar değildir.

Tablo 3. Nihai Test Puanları ve Yetenek Kestirimleri ile Türkçe Dersi Yıl Sonu Başarı Notları ve Yıl Sonu Genel Başarı Not Ortalamaları Arasındaki İlişkiler

Sınıf Düzeyi	Test Türü	Türkçe Dersi Notları	t	Yıl Sonu Not Ortalaması	t
6	KTK	.691*	0.85	.749*	0.73
	MTK	.677*		.738*	
7	KTK	.720*	3.57*	.680*	2.75**
	MTK	.659*		.630*	
8	KTK	.611*	-1.35	.607*	-1.40
	MTK	.637*		.634*	
Genel	KTK	.669*	1.65	.666*	0.87
	MTK	.652*		.657*	

* p<0.001, ** p<0.01

Tablo 3'te sunulan geçerlilik katsayılarına sınıf düzeylerine göre ayrı ayrı bakıldığında ise 7. sınıf düzeyinde Klasik Test Kuramı'na göre geliştirilen test lehine bulunan farkların istatistiksel olarak manidar olduğu görülmektedir. Diğer sınıf düzeylerinde gözlemlenen korelâsyonlar arasındaki farklar ise manidar bulunmamıştır.

Her iki kurama göre geliştirilen testlerin KR-20 iç tutarlılık ve Lord'un güvenilirlik katsayıları Tablo 4'te sunulmuştur. Tablo incelendiğinde bütün katsayıların 0.88'in üzerinde olduğu görülmektedir. En yüksek güvenilirlik katsayısı değeri, Klasik Test Kuramı'na göre geliştirilen test için Madde Tepki Kuramı'na göre hesaplanan katsayıdır. En düşük değer ise Madde Tepki Kuramı'na göre geliştirilen test için hesaplanan KR-20 katsayısında gözlemlenmiştir. Ancak bütün katsayılar birbirine çok yakındır.

Tablo 4. İki Kurama göre Geliştirilen Testlerin Güvenilirlik Katsayıları

	KTK'ya Göre Geliştirilen Test	MTK'ya Göre Geliştirilen Test
KR-20	**0.889**	*0.883*
Lord'un Güvenilirlik Katsayısı	*0.905*	**0.895**

TARTIŞMA ve SONUÇ

Klasik Test Kuramı'na göre geliştirilen testin puan ortalaması ile Madde Tepki Kuramı'na göre geliştirilen testin yetenek kestirimleri arasında hesaplanan yüksek korelâsyon katsayıları, puan ve yetenek kestirimi dağılımlarının birbirine benzer olduğunun bir kanıtı olarak düşünülebilir. Bu bulgu, iki ayrı testten iki ayrı kurama göre hesaplanan puanların ve yetenek ölçülerinin bireyleri sözel yeteneklerine göre benzer şekilde sıraya koyduğunu, biri yerine diğeri kullanılması durumunda sonuçlarda çok değişiklik meydana gelmeyeceğini göstermektedir. Bulunan katsayı, yapılan birçok çalışmada (Başarır-Erden, 1997; Baykul, 1979; Courville, 2004; Fan, 1998; Hwang, 2002; Macdonald ve Paunonen, 2002; Nartgün, 2002; Stage, 1997a, 1997b, 1997c, 1997d; Stage 1998a, 1998b; Stage 1999a, 1999b; Stage, 2003; Warrens, Gruijter ve Heiser, 2007) aynı test üzerinden elde edilen test puanı yetenek kestirimi korelâsyonlarından çok farklı değildir.

Klasik Test Kuramı'na göre geliştirilen testin test puanlarında sınıf düzeyi arttıkça meydana gelen artışın, Madde Tepki Kuramı'na göre geliştirilen testten elde edilen yetenek puanlarında da benzer şekilde olduğu görülmektedir. Her iki karşılaştırma için yapılan varyans analizi sonucunda sınıf düzeylerine göre test puanı ortalamaları arasındaki fark istatistiksel olarak manidardır. Test puanı ve yetenek kestirimleri arasındaki farkların sınıf düzeylerine göre değişim göstermemesi sınıf ve test türü ortak etkisinin olmadığını, diğer bir deyişle sınıf düzeylerine göre gözlenen farkların testlere göre değişmediğini göstermektedir. Sınıf düzeyine göre yapılan karşılaştırmalarda her iki testten elde edilen test puanı ve yetenek kestirimi değerlerinin aynı sınıf düzeylerinde farklılaştığı görülmektedir. Özdemir'in (2002) çalışmasında da sınıf düzeyine göre farklılaşma hem test puanları hem de aynı testten elde edilen yetenek kestirimlerinde benzer olduğu bulunmuştur. Başarır-Erden'in (1997) çalışmasında da benzer sonuçlara ulaşılmıştır. Klasik Test Kuramı'na göre geliştirilmiş testten alınan puanların sınıf düzeyine göre değişiminin Madde Tepki Kuramı'na göre geliştirilen testten elde edilen yetenek kestirimleri ile benzer olması iki test puanının geçerliliğinin benzer olduğunu göstermektedir.

Tüm sınıf düzeylerinde ve grubun genelinde bahsedilen notlarla her iki test ve yetenek puanları arasında pozitif yönlü ve istatistiksel olarak manidar korelâsyonlar bulunmuştur. Nijenhuis, Evers ve Mur (2000), lise başarı notlarıyla standart yetenek testleri arasındaki korelâsyonların ortalamasının 0.50 olduğunu ifade etmektedir. Jensen (1998) ise en yüksek geçerlilik katsayılarını ortaokul öğrencilerinde ve 0.60 ve 0.70 arasında bulmuş, lise öğrencilerinde katsayıların 0.50-0.60 aralığına gerilediğini işaret etmiştir. Bu çalışmada da grubun tümünün Klasik Test Kuramı'na göre geliştirilen testten aldıkları puanlar ile Türkçe dersi başarı not ortalaması arasındaki korelâsyon r=0.669; Madde Tepki Kuramı'na göre geliştirilen teste ait yetenek kestirimleriyle Türkçe dersi not ortalamaları arasındaki korelâsyon ise r=0.652 olarak bulunmuştur. İki korelâsyon katsayısı arasındaki farkın manidar olmaması her iki testin Türkçe dersi yılsonu başarı notları ve yılsonu genel başarı not ortalamaları ölçütlerine göre hesaplanan geçerlilikleri bakımından fark bulunmadığı göstermektedir. Benzer durum test puanları ve yetenek kestirimleri ile yılsonu başarı not ortalaması arasındaki ilişkilerde de söz konusudur. Bu bulgular Baykul'un (1979) çalışmasıyla da paraleldir. Diğer sınıf düzeylerinden farklı olarak 7. sınıf düzeyinde Klasik Test Kuramı'na göre geliştirilen testin geçerlilik katsayısının yüksek olması ve genel olarak Klasik Test Kuramı'na göre geliştirilen testin daha yüksek geçerlilik katsayıları vermesinin, ölçüt olarak kullanılan okul notlarının da Klasik Kuram'la belirleniyor olmasından kaynaklandığı savunulabilir.

Her iki kurama göre geliştirilen testlerin güvenilirlik katsayıları incelendiğinde Klasik Test Kuramı'na göre geliştirilen test için hesaplanan KR-20 katsayısı ile Madde Tepki Kuramı'na göre hesaplanan Lord'un Güvenilirlik katsayısının çok yakın değerler aldığı görülmektedir. Wilson ve diğerleri'nin (2006) elde ettiği aynı test maddeleri için hesaplanan güvenilirlik katsayılarının benzerliği sonucuna, aynı uygulama sonucunda farklı kuramlara göre ayrı ayrı oluşturulmuş iki test için de ulaşılmıştır. Her ne kadar, bu katsayılar iki kurama göre farklı yöntemlerle elde ediliyor ve farklı anlamlara geliyor olsa da sonuçta güvenilirlik kanıtı olarak kullanılmaktadırlar ve güvenilir test geliştirme yönüyle de iki kuramın aynı noktada buluşmakta olduğu görülmektedir.

Klasik Test Kuramı, daha yaygın olarak bilinen ve kullanılan kuramdır. Madde Tepki Kuramı ise son yıllarda kullanımı artan bir kuram olarak karşımıza çıkmaktadır. İki kuramdan da haberdar

olan bir test geliştiricinin aklına gelebilecek olan "testimi hangi kurama göre geliştirmeliyim" sorusuna verilebilecek yanıt, amaca göre değişebilecektir. Eğer test geliştiricinin amacı, ilgilendiği özelliği ölçecek geçerli ve güvenilir bir araç ortaya koymaktan ileri gitmiyor ise seçilecek kuramın hiç fark etmeyeceği, bu çalışmayla ortaya konmuş olan bir sonuçtur. Böyle bir durumda, seçilecek kuramdan çok, seçilen kuramın gereklerini yerine getirmek önem kazanmaktadır.

Eğer test geliştiricinin, Madde Tepki Kuramı'nın Klasik Kuram'dan farklı olarak sağladığı düşünülen, test puanlarını eşitleme, bireye uyarlanmış test geliştirme, madde yanlılıklarını belirleme gibi amaçları söz konusu ise tercih edeceği kuram, her ne kadar bu konulara getirdiği çözümlerin sağlamlığı konusunda çalışmalar devam etmekte ve kesin bir sonuca ulaşılamamış ise de, Madde Tepki Kuramı olmalıdır.

KAYNAKÇA

Başarır-Erden, D.B. (1997). *Örtük Özellikler ve Klasik Test Teorisi Yaklaşımına Dayalı Olarak Geliştirilen Likert Tipi Tutum Ölçeğinin Psikometrik Özelliklerinin Karşılaştırılması.* Yayımlanmamış doktora tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Baykul, Y. (1979). *Örtük Özellikler ve Klasik Test Kuramları Üzerine Bir Araştırma.* Yayımlanmamış doktora tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Baykul, Y. (2000). *Eğitimde ve Psikolojide Ölçme: Klasik Test Teorisi ve Uygulaması.* Ankara: ÖSYM Yayınları.

Bruning, J.L., Kintz, B.L. (1993) *İstatistik*. Çev.: Dönmez, A. Gündoğan Yayınları, Ankara.

Büyüköztürk, Ş. (2007).*Sosyal Bilimler İçin Veri Analizi El Kitabı*. 8. Baskı, Pegem A Yayınları, Ankara.

Courville, T.G. (2004). An Empirical Comparison of Item Response Theory and Classical Test Theory Item/Person Statistic, Yayımlanmamış doktora tezi, Texas A&M University, Educational and Psychological Measurement. [Online]: Retrieved on 21-November-2007, at URL:Web: http://txspace.tamu.edu/bitstream/handle/1969.1/1064/etd-tamu-2004B-EPSY-Courville-2.pdf;jsessionid= 9921EB67B88437D40E179EC843EAFE4C?sequence=1.

Crocker, L., Algina, J. (1986). *Introduction Classical and Modern Test Theory.* USA: CBS College Publishing Company.

Çıkrıkçı-Demirtaşlı N. (1998). Test Geliştirmede Yeni Yaklaşımlar: Örtük Özellikler Kuramı - Temel Özellikleri, Varsayımları, Model ve Sınırlılıkları. *Ankara Üniversitesi Eğitim Fakültesi Dergisi*, 2 (28), 161-173.

Fan, X. (1998). Item response theory and classical theory: an empirical comparison of their item-person statistic. *Educational and Psychological Measurement,* v:58 n:3, 357-381.

Hambelton, R.K. (1990). Item response theory: introduction and bibliography. *Psicothema*, 2, 1, 97-107.

Hambelton, R.K. (1994). Item Response theory: a broad psychometric framework for measurement advances. *Psicothema*, 6, 3, 535-556.

Hambelton, R.K., Swaminathan H., and Rogers, H.J. (1991). *Fundamentals of Item Response Theory.* California: Sage Publications Inc.

Hambleton, R.K., Swaminathan H. (1985). *Item Response Theory: Principles and Application.* Kluwer, Nijhoff Publishing a Member of the Kluwer Academic Publisher Group.

Harvey, R.J., Hammer A.L. (1999). Item response theory. *Counseling Psychologist,* Vol:27, No: 3, 353-374.

Hwang, D.Y. (2002). Classical Test Theory and tem Response Theory: Analytical and Empirical Comparison. Speeches/meeting paper, presented at the Annual Meeting of the Southwest Educational Research Association (Austin, Feb. 14-16, 2002)

Jensen, A.R. (1998). The g Factor: the Science of Mental Ability. (Westport, Preager). Akt.: Nijenhuis, N.T., Evers, A., and Mur, J.P. (2000).

Kelecioğlu, H. (2001). Örtük Özellikler Teorisindeki b ve a Parametreleri ile Klasik Test Teorisindeki p ve r İstatistikleri Arasındaki İlişki. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi,* 20: 104-110.

Kuşçu, H., Kuşçu, Ü. (1997). *Atasözleri ve Deyimler Sözlüğü* (Genişletilmiş Yeni Baskı). Altın Kitaplar Yayınevi, İstanbul.

Linden, W.J., Hambleton, R.K. (1997) *Handbook of Modern Item Response Theory.* Springer-Verlag New York Inc.

Linn, R.L. (1998). *Educational Measurement*. New York: Macmillan Publishers.

Lord, F.M., Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. Addison Wesley Publishing Company, Educational Testing Service.

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, 921-943. [Online]: Retrieved on 24-November-2007, at URL: http://epm.sagepub.com./cgi/content/ abstract/62/6/921.

Mendeş, M., Karabayır, A. (2003). Bağımlı İki veya Daha Fazla Korelâsyon Katsayısının Karşılaştırılmasında Kullanılan Test Yöntemleri. *Hayvansal Üretim*. 44 (2): 91-98.

Mislevy, R.J., Bock, R.D. (1990). *Item Analysis and Test Scoring with Binary Logistic Models*. 2. Ed., Scientific Software, Inc., Mooresville.

Nartgün, Z. (2002). *Aynı Tutumu Ölçmeye Yönelik Likert Tipi Ölçek ile Metrik Ölçeğin Madde ve Ölçek Özelliklerinin Klasik Test Kuramı ve Örtük Özellikler Kuramına Göre İncelenmesi*. Yayımlanmamış doktora tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Nijenhuis, N.T., Evers, A., and Mur, J.P. (2000). Validity of the differential aptitude test fort the assessment of immigrant children. *Educational Psychology*, 20, 1, 99-115.

Reise, S.P., Ainsworth, A.T., and Haviland, M.G. (2005). Item Response theory. Fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science*, 14, 2, 95-101.

Stage, C. (1997a). The Applicability of Item Response Models to the SweSAT. A Study of the DTM Subtest. Em No: 21, Report, Sweden Umea University, Department of Educational Measurement. [Online]: Retrieved on 04-December-2007, at URL: http://www.umu.se/ edmeas/publikationer/pdf/EM%21no%2197.pdf.

Stage, C. (1997b). The Applicability of Item Response Models to the SwesSAT. A Study of the ERC Subtest. Em No: 24. Report, Sweden Umea University, Department of Educational Measurement. [Online]: Retrieved on 04-December-2007, at URL: http://www.umu.se/ edmeas/publikationer/pdf/EM%24no%2497.pdf.

Stage, C. (1997c). The Applicability of Item Response Models to the SweSAT a Study of the READ Subtest. Em No: 25. Report, Sweden Umea University, Department of Educational Measurement. [Online]: Retrieved on 04-December-2007, at URL: http://www.umu.se/ edmeas/publikationer/pdf/EM%25no%2597.pdf.

Stage, C. (1997d). The Applicability of Item Response Models to the SweSAT. A Study of the WORD Subtest. Em No: 26. Report, Sweden Umea University, Department of Educational Measurement. [Online]: Retrieved on 04-December-2007, at URL: http://www.umu.se/edmeas/publikationer/pdf/EM%26no%2697.pdf.

Stage, C. (1998a). A Comparison Between Item Analysis Based on Item Response Theory and Classical Test Theory. A Study of the SweSAT Subtest WORD. Report, Sweden Umea University, Department of Educational Measurement. [Online]: Retrieved on 04-December-2007, at URL: http://www.umu.se/edmeas/publikationer/pdf/enr2998sec.pdf.

Stage, C. (1998b). A Comparison Between Item Analysis Based on Item Response Theory and Classical Test Theory. A Study of the SweSAT Subtest ERC. Report, Sweden Umea University, Department of Educational Measurement. [Online]: Retrieved on 04-December-2007, at URL: http://www.umu.se/edmeas/publikationer/pdf/enr3098sec.pdf.

Stage, C. (1999a). A Comparison Between Item Analysis Based on Item Response Theory and Classical Test Theory. A Study of the SweSAT Subtest READ. Report, Sweden Umea University, Department of Educational Measurement. [Online]: Retrieved on 04-December-2007, at URL:http://www.umu.se/edmeas/publikationer/pdf/enr3399sec.pdf.

Stage, C. (1999b). Predicting Gender Differences in WORD Items. A Comparison of Item Response Theory and Classical Test Theory. Report, Sweden Umea University, Department of Educational Measurement. [Online]: Retrieved on 04-December-2007, at URL: http://www.umu.se/edmeas/publikationer/pdf/enr3499sec.pdf.

Stage, C. (2003). Classical Test Theory or Item Response Theory: The Swedish Experience. Em No: 42, Report, Sweden Umea University, Department of Educational Measurement. [Online]: Retrieved on 04-December-2007, at URL: http://www.umu.se/edmeas/publikationer/ pdf/em%20no%2042.pdf.

Türk Dil Kurumu (2005a). *Türkçe Sözlük*. Türk Dil Kurumu, Ankara.

Türk Dil Kurumu (2005b). *İlköğretim Okulları İçin Türkçe Sözlük*. Türk Dil Kurumu, Ankara.

Türk Dil Kurumu (2005c). *İlköğretim Okulları İçin Yazım Kılavuzu*. Türk Dil Kurumu, Ankara.

Warrens, M.J., Gruijter, D.N.M., Heiser, W.J. (2007). A Systematic Comparison Between Classical Optimal Scaling and the Two-Parameter IRT Model. *Applied Psychological Measurement*, 31, 2, 106-120. [Online]: Retrieved on 21-November-2007, at URL: http://apm.sagepub.com./cgi/content/abstract/ 31/2/106.

Wilson, M., Allen, D.D., Li, J.C. (2006). Improving measurement in health education and health behavior research using item response modeling: comparisons with the classical test theory approach. *Health Education Research Theory and Practice. 21,* 19-32.