Elementary Education Online, 9(1), 170-187, 2010.
İlköğretim Online, 9(1), 170-187, 2010. [Online]: http://ilkogretim-online.org.tr

A Comparison of Differential Scoring Methods For Multiple Choice Tests in Terms of Classical Test and Item Response Theories*

Göksu GÖZEN ÇITAK**

ABSTRACT. The purpose of this research is to determine the effects of binary (1-0) scoring, judgement-based (a priori) option weighting and empirical option weighting on the reliability and validity of a multiple-choice test regarding Classical Test and Item Response theories. Data were collected through the administration of a multiple-choice test of verbal ability to 1593 students attending several departments at Hacettepe and Gazi Universities. Research findings showed that regarding Item Response Theory, "1-0" scoring estimates the parameters within different intervals on the ability scale more precisely than weighted scoring and binary scoring is superior to weighted scoring in terms of validity. In case of Classical Test Theory, results indicated that empirical option weighting estimates the highest reliability and all scoring methods cause an identical effect on test validity.

Key Words: multiple choice tests, partial credit model, scoring methods, option weighting

SUMMARY

Purpose and significance: The purpose of this research is to determine the effects of binary (1-0) scoring, judgement-based option weighting and empirical option weighting on the reliability and validity of a multiple-choice test regarding Classical Test and Item Response theories. It is considered that presenting the effects of differential scoring methods on test statistics contributes not only to the decisions related to the assessments of performance in educational diagnosis, but also to determine more reliable and valid scoring methods.

Methods: Data were collected through the administration of an 18-itemed multiple-choice test of verbal ability to 1593 students attending several departments at Hacettepe and Gazi Universities. In case of data analysis, basic assumptions of Item Response theory were checked first and data collected under assumptions of test theories were analyzed by using statistical techniques including differential reliability and validity estimation methods.

Results: Results of the analysis carried out to check the basic assumptions of Item Response Theory were accepted as an evidence for normality of the distribution, unidimensionality of test items, and local independence of the responses. Findings show that α coefficient (0,69) estimated by empirical option weighting is superior to the reliability coefficients derived by binary scoring (0,64) and judgement-based scoring (0,68) regarding Classical Test Theory and the reliability estimated for binary data by Lord's coefficient is (0,88) superior to the marginal reliability coefficients (0,71) estimated for other methods regarding Item Response theory. It is assessed that in terms of the maximum value of criterion validity coefficient, empirical weighting yields higher validity than the other methods regarding Classical Test Theory and binary scoring yields higher validity than the other methods regarding Item Response Theory.

Discussion and Conclusions: Results showed that under Item Response Theory, "1-0" scoring estimated the parameters within different intervals on the ability scale more precisely than weighted scoring and it is concluded that binary scoring is superior to weighted scoring in terms of validity. In case of Classical Test Theory, empirical option weighting estimated the highest reliability in comparison with the other methods and it is concluded that all of the scoring methods cause an identical effect on test validity. Based on the results; researches discussing alternative scoring methods, based on tests of differential domains and study groups and designed for presenting the effects of scoring methods in responding behaviors are recommended. Also, it is recommended that item writing techniques should be examined to benefit from advantages of weighted scoring methods.

* This study is a summary of researcher's doctoral dissertation prepared in the Institute of Educational Sciences of Ankara University under the advisory of Prof. Dr. Ezel Tavşancıl in 2007.

** Lecturer, PhD., Ankara University, Faculty of Educational Sciences, gozen@education.ankara.edu.tr

Klasik Test ve Madde Tepki Kuramlarına Göre Çoktan Seçmeli Testlerde Farklı Puanlama Yöntemlerinin Karşılaştırılması*

Göksu GÖZEN ÇITAK**

ÖZ. Bu çalışmada, çoktan seçmeli bir testte yer alan madde seçeneklerinin iki kategorili (1-0) puanlama, uzman yargısına dayalı seçenek ağırlıklandırma ve deneysel seçenek ağırlıklandırma yöntemleriyle puanlanması durumunda testin güvenirliğinin ve geçerliğinin Klasik Test Kuramı ve Madde Tepki Kuramı'na göre nasıl etkilendiğinin belirlenmesi amaçlanmıştır. Araştırma verisi çoktan seçmeli bir sözel yetenek testinin, Hacettepe ve Gazi Üniversitesi'nin çeşitli bölümlerinde okuyan toplam 1593 öğrenciye uygulanmasıyla elde edilmiştir. Araştırmanın bulguları, Madde Tepki Kuramı'nda "1-0" puanlamanın kullanıldığı durumda yetenek ölçeği üzerindeki parametrelerin ağırlıklı puanlamaların kullanıldığı duruma göre daha doğru kestirildiğini göstermiş, bu yöntemin test geçerliği açısından da daha etkili olduğu sonucuna ulaşılmıştır. Klasik Test Kuramı'nda ise deneysel ağırlıklandırmanın kullanıldığı durumda güvenirliğin daha yüksek kestirildiği, ancak tüm yöntemlerin test geçerliği açısından benzer etkiyi yaptığı belirlenmiştir.

Anahtar Sözcükler: çoktan seçmeli testler, kısmi puan modeli, puanlama yöntemleri, seçenek ağırlıklandırma

GİRİŞ

Başarı, ilgi, yetenek vb. doğrudan gözlenemeyen özelliklerin çeşitli araçların kullanılmasıyla nicelleştirilmesi ölçme ve değerlendirme çalışmalarının temelini oluşturur. Bu özelliklerin nicel verilere dönüştürülmesine puanlama yöntemleri hizmet eder. Bu yöntemler aynı zamanda kullanılan ölçme araçlarının psikometrik özellikleri üzerinde de çeşitli etkilere sahiptir.

Klasik Test ve Madde Tepki kuramlarıyla ilgili yürütülen ilk çalışmalarda bireyin bir testte yer alan maddelere verdiği cevapların doğru veya yanlış olarak sınıflandırıldığı modeller üzerinde durulmuştur (Lord ve Novick, 1968; Hambleton ve Swaminathan, 1985; Crocker ve Algina, 1986). Geleneksel puanlama, "1-0" puanlama ya da doğru cevap sayısı yöntemi olarak adlandırılan bu modellerde bir maddeyi anahtara göre doğru işaretleyen öğrenci o maddeyi doğru cevaplamış kabul edilir ve bireye o madde için "1" puan verilir; maddeyi boş bırakan veya çeldiricilerden birini veya seçeneklerden birkaçını işaretleyen öğrenci ise o maddeyi yanlış cevaplamış kabul edilir ve bireye o madde için "0" (sıfır) puan verilir. Buradaki temel ilke, sadece tam bilgiye sahip bireylerin doğru cevaplarına ve tahminle verilen doğru cevaplara puan atamaktır. Tamamen yanlış bilgiye sahip olan, kısmi bilgiye sahip olan ve tahminle cevaplama davranışında bulunarak yanlış cevap veren bireylerin cevapları ise yanlış cevap kategorisinde değerlendirilmektedir (Jaradat ve Tollefson, 1988). Bu yöntemde maddelere eşit ağırlıklar verilmekte, tüm maddelerin başarıyı eşit şekilde sınadığı ve tüm çeldiricilerin eşit düzeyde çekiciliğe sahip olduğu varsayılmaktadır (Haladyna, 1990). Bu yöntem yaygın olarak kullanılmakla beraber tüm cevaplayıcı-madde ilişkilerinin iki kategorili modellerle incelenmesinin uygun olmayacağı fikrine dayalı olarak da eleştirilmektedir. Örneğin müzik, dans ve konuşma performansını belirlemek veya başarıyı/yeteneği ölçmede kullanılan kısmi olarak doğru yanıtlanmış farklı türdeki maddeleri puanlamak için ikiden fazla kategori içeren modellere ihtiyaç duyulur.

Başarıyı ve/veya yeteneği ölçmede çoktan seçmeli testler sıkça kullanılan araçlardır ve bunlar için yaygın olarak kullanılan yöntem yine "1-0" yöntemiyle puanlamadır. Coombs, Milholland ve Womer (1956), Frary (1980) ve Ben-Simon, Budescu ve Nevo (1997) çoktan seçmeli bir madde üzerinde cevaplayıcının sahip olabileceği bilgi düzeylerini tartışırken, bireyin doğru seçeneği çeldiricilerle sınıflaması sonucu basit bir mantıkla çeldiricilerden birinin doğru olduğuna inanabileceği örneğini vermektedir. Buna zıt şekilde, doğru seçenek bir veya daha fazla çeldiriciyle de sınıflanabilir ve kalan iki veya daha fazla çeldirici arasından tahminde bulunulabilir. Bu farklı durumlar cevaplayıcının bilgisindeki tam bilgiden bilginin yokluğuna uzanan ve süreklilik gösteren değişkenliği (Hutchinson, 1982), başka bir deyişle kısmi bilginin değişik düzeylerdeki varlığını ortaya koyar. Çoktan seçmeli maddelerin seçenekleri üzerinde yanlış yanıtların dağılımları yetenek düzeyleri arasında farklılaştığından, cevaplayıcının doğru yanıtı bildiğini veya tesadüfi olarak yanlış seçeneklerden birini işaretlediğini varsayan bir modelden öte

* Bu çalışma, 2007 yılında Ankara Üniversitesi Eğitim Bilimleri Enstitüsü'nde Prof. Dr. Ezel Tavşancıl danışmanlığında yürütülmüş olan doktora tezinin özetidir.

** Öğretim Görevlisi Dr., Ankara Üniversitesi, Eğitim Bilimleri Fakültesi, gozen@education.ankara.edu.tr

tüm seçeneklerden bilgi elde edilebilen bir modelin kullanılmasına gereksinim vardır (De Ayala, 1993). Böylelikle, gerçek bilgi durumuna daha yakın sonuçlar elde edebilmek için cevaplayıcıların kısmi bilgilerini yansıtan puanlama yöntemlerinin geliştirilmesi ve kullanılması önemli hale gelmiştir (Ben-Simon, Budescu ve Nevo, 1997).

Klasik Test ve Tepki Kuramı'nda kısmi bilgiyi ortaya koymak için önerilmiş olan tüm puanlama modellerinin temelinde "bireylere ilişkin ölçümlerin bileşik bir ölçüm oluşturmak veya bir ölçütü kestirmek için birleştirilmesiyle, aynı ölçümlerin toplanması veya ortalamasının alınması durumunda elde edilenden daha güvenilir veya geçerli bir ölçüm elde edilebilir" beklentisi yatmaktadır. Bileşen ölçülerinin eşit güvenirliğe ve varyanslara sahip olması, eşit interkorelasyonlar göstermesi ve ölçülen örtük değişken veya kestirilecek dış ölçütle eşit düzeyde korelasyon göstermesi zor olsa da Wang ve Stanley (1970) ve Waters (1976) bileşen özelliklerinden her birinin bileşik ölçüme yansıtılacağını, böylelikle kısmi bilgiyi önemseyen puanlamaların işlevsel olacağını belirtmektedir. Maddelerin eşit düzeyde bilgi içermediğini ve seçeneklerin farklı düzeylerde kullanışlılığa sahip olduğunu varsayan bu puanlama yöntemleri "çok kategorili/ağırlıklı puanlama" olarak adlandırılmaktadır (Haladyna, 1990).

Çok kategorili/ağırlıklı puanlama yöntemleri

Crehan ve Haladyna (1994)'nın tanımlamalarına göre "yanlış cevap seçeneklerinde mevcut olan farklı düzeylerdeki bilginin kullanımını" içeren çok kategorili puanlama yöntemleri için pek çok farklı sınıflama önerilmektedir. Ancak bunların tümü temelde Lord ve Novick (1968) tarafından çoktan seçmeli testler için belirlenen üç cevaplama-puanlama yöntemi altında yer alır; (1) Yeni madde yapıları ve/veya cevaplama yöntemleri, (2) Farklı madde puanlama yöntemleri (seçenek ağırlıklandırma yöntemleri) ve (3) Farklı test puanlama yöntemleri (madde ağırlıklandırma yöntemleri). Bunlar içerisinde yer alan madde puanlama yöntemlerini Frary (1989) iki başlık altında ele almaktadır. Bunlar cevaplayıcı yargısına dayalı yöntemler ile doğrudan cevaplama yöntemleridir.

Doğrudan cevaplama yöntemlerinde cevaplayıcılar en doğru olduğuna inandıkları cevabı işaretlerler. Bu sınıfta; doğruyu bulana dek cevaplama yöntemi, çoklu doğru seçenek yöntemi, seçenek ağırlıklandırma yöntemleri ve Madde Tepki Kuramı yöntemleri yer almaktadır (Frary, 1989). Seçenek ağırlıklandırma yöntemlerinde, kısmi bilginin yanlış yanıtlara farklı ağırlıklar verilmesiyle ölçülebileceği varsayılarak her madde için bireye, işaretlediği seçeneğe bağlı olarak bir puan atanır ve geleneksel puanlamalarda önemsenmeyen toplam test puanları varyansına katkıda bulunulur (Lord ve Novick, 1968; Waters, 1976). Lord ve Novick (1968) ve Frary (1989) seçenek ağırlıklandırma yöntemlerini iki grup altında ele almıştır: (1) uzman yargısına dayalı önsel (a priori) seçenek ağırlıklandırma ve (2) önceki uygulamalara dayalı deneysel ağırlıklandırma.

Uzman yargısına dayalı ağırlıklandırma, seçenek puanlarının belirlenmesinde uzman görüşüne başvurmayı gerektirir. Seçenekler doğru cevaba yakınlık derecelerine göre sıralanabilir ve sıralamalarla orantılı ağırlıklar maddelere puan olarak atanabilir. Ağırlıklandırma, ağırlık ortalamalarını gösteren bir puanlama cetveliyle gerçekleştirilir (Echternacht, 1976). Bu tür ağırlıklarının kullanıldığı çalışmalar incelendiğinde, elde edilen sonuçlar arasında tutarlılık olmadığı gözlenmiştir. Örneğin, Nedelsky (1954), Hambleton, Roberts ve Traub (1970) ve Patnaik ve Traub (1973)'un yürüttüğü çalışmalarda önsel ağırlıklandırma farklı puanlamalarla karşılaştırılmış, sonuçlar önsel ağırlıklandırmanın kullanıldığı testlerin güvenirliğinin diğer yöntemlerin kullanıldığı testlerden manidar biçimde yüksek olduğunu göstermiştir. Buna ters düşecek biçimde, Kansup ve Hakstain (1975) uzman yargısına dayalı ağırlıklandırmanın geleneksel puanlamalara göre avantaj sağlamadığını belirlemiş, benzer sonuçlar Echternacht (1976), Downey (1979) ve Cross, Ross ve Geller (1980) tarafından da elde edilmiştir.

Deneysel ağırlıklandırma yönteminde ise seçenek ağırlıkları verilen cevaplara dayalı olarak belirlenir. Kullanılan deneysel ağırlıklar genelde seçeneklerin çekicilikleri, seçeneği işaretleyen cevaplayıcının ortalama standartlaştırılmış puanları ve işaretlenen seçenek ile diğer maddelerden elde edilen toplam puan arasındaki korelasyonlardır (Ben-Simon, Budescu ve Nevo, 1997). Bu yöntemler, doğrusal ve doğrusal olmayan yöntemler üzerine kuruludur (Haladyna, 1990). Doğrusal yöntemler, seçenek ağırlıklarının ve madde cevaplarının doğrusal bileşimlerini göz önünde bulundurur ve bu kategoride "iki yönlü (reciprocal) ortalamalar", "seçenek-toplam korelasyonu" ve "oran ağırlıkları" yöntemleri yer almaktadır. Doğrusal olmayan yaklaşımlar ise bir dizi Madde Tepki Kuramı modelini içerir ve seçenekler için doğrusal olmayan dönüşümler gerektirir (Thissen, 1976).

Doğrusal yöntemler üzerinde yürütülen çalışmalar (Guilford, 1941; Guttman, 1941; Davis ve Fifer, 1959; Echternacht, 1976; Cross ve Frary, 1978; Downey, 1979; Cross, Ross ve Geller, 1980) deneysel seçenek ağırlıklandırmanın "1-0" puanlama yöntemiyle karşılaştırıldığında iç tutarlılık

anlamındaki güvenirliği anlamlı bir şekilde artırdığını göstermiştir. Diğer yandan, çoğu araştırmacının bu yöntemin geçerliğe etkisini ya değerlendirmemiş olduğu ya da bu yöntemin geçerliği düşürdüğü sonucunu elde ettiği (Sabers ve White, 1969; Echternacht; 1976) gözlenmektedir.

Doğrudan cevaplama yöntemleri içerisinde yer alan tek boyutlu çok kategorili Madde Tepki Kuramı modellerinin tümü yine seçenek ağırlıklandırmayı temel almaktadır ve cevaplayıcının bir maddeyi farklı performans düzeylerinde tamamlayabilmesine olanak veren test durumları için geliştirilmiştir (Glas ve Verhelst, 1989). Bu modeller işlem basamaklarının ardışık bir sıra ile tamamlanmasını gerektiren ve doğruluk derecesine göre sıralanabilen cevaplar için uygundur. Bu modellerden biri, Samejima (1969) tarafından geliştirilen derecelendirilmiş cevap modelleri olarak adlandırılan Madde Tepki Kuramı modelleri arasında yer alan ve Masters (1982) tarafından önerilen kısmi puan modelidir.

Kısmi puan modeli (Partial credit model)

Kısmi puan modeli her bir maddenin kendi oranlı ölçek yapısına sahip olduğu bir modeldir. Wright (1999) bu modelin, cevapların belli oranda bilgiyi içerdiği ve cevaplayıcının cevabın doğruluğu oranında kısmi puan aldığı çoktan seçmeli testler için oldukça kullanışlı olduğunu belirtmektedir. Modelde cevabın kısmi doğruluğu maddeden maddeye farklılaşmakta, cevaplayıcı maddedeki işlemin tamamlandığı performans düzeyine eşit bir puan almaktadır. Masters (1988) modelde tanımlanan madde parametrelerinin ve cevap kategorilerinin sıraları arasında bir ilişki olması zorunluluğunun bulunmadığını, işlem basamaklarının aynı güçlükte olması şartı aranmadığı gibi, basamakların güçlüklerine göre sıralanmasının da gerekmediğini vurgulamaktadır. Rasch ailesinin ölçme modellerinin bir üyesi olarak kısmi puan modeli, Rasch modelinin her madde için bireyden bağımsız parametreler kestiren ve tatmin edici istatistikler elde edilmesine olanak veren özelliklerini içinde barındırmaktadır. Bir cevaplayıcının yeteneğini değerlendirmek için kullanılan puanlama kuralının test etmenin temel amaçlarından biri olan tanı sürecindeki işlevselliğinden (Adams, 1988) dolayı kısmi puan modelinin çok kategorili diğer örtük özellik modellerine göre üstünlüğe sahip olduğu savunulmaktadır (Samejima; 1969; Dodd,1984; Dodd ve Koch; 1987).

Test kuramlarının varsayımlarına dayalı olarak puanlama yöntemlerinin farklı matematiksel modellere sahip olduğu ve testlerin psikometrik niteliklerinde de farklı katkılara neden olduğu gözlenmiştir. "1-0" puanlama yönteminin sınırlılıkları, kısmi bilginin ölçülmesini temel alarak önerilen farklı ağırlıklandırma yöntemlerindeki ve bunların kullanıldığı araştırma sonuçlarındaki farklılıklar göz önünde bulundurularak, çoktan seçmeli bir testte yer alan maddelerin seçeneklerinin ağırlıklı puanlanmasının testin güvenirliğine ve geçerliğine nasıl etki ettiğinin Klasik Test ve Madde Tepki kuramlarına göre belirlenmeye çalışılmasında gereklilik görülmüştür. Yurt dışında bu konuda pek çok araştırma (Corey, 1930; Odell, 1931; Coombs, Milholland ve Womer, 1956; Wang ve Stanley, 1970; Bayuk, 1973; Waters, 1976; Frary, 1980; Jaradat ve Tollefson, 1988; Sympson ve Haladyna, 1988; Haladyna, 1990; Ben-Simon, Budescu ve Nevo, 1997; Backhoff, Tirado ve Larrazolo, 2001) yürütülmüş olmasına rağmen, Türkiye'de bu konu üzerinde yürütülen araştırmaların (Akkuş, 2000, Özdemir, 2002; Saygı, 2004; Gözen, 2006) sayısı oldukça azdır. Bu araştırma ile, kısmi bilginin varlığına, işlevine, önemine ve ölçülmesine yönelik çalışmalara ilgi çekerek Türkiye'de bu konuda yapılmış olan az sayıda çalışmaya bir yenisini eklemek hedeflenmiştir.

Amaç

Bu araştırmanın genel amacı, çoktan seçmeli bir testte, seçeneklerin iki kategorili ("1-0") ve ağırlıklı puanlama (uzman yargısına dayalı seçenek ağırlıklandırma ve deneysel seçenek ağırlıklandırma) yöntemleriyle puanlanması durumunda testin psikometrik özelliklerinin (güvenirlik ve geçerlik) Klasik Test Kuramı ve Madde Tepki Kuramı'na göre nasıl etkilendiğini belirlemektir.

Belirlenen amaç doğrultusunda bu çalışma kapsamında yanıt aranan sorular şunlardır:

1) Çoktan seçmeli bir testte, seçeneklerin Klasik Test Kuramı'na göre "1-0" puanlama, uzman yargısına dayalı seçenek ağırlıklandırma ve deneysel seçenek ağırlıklandırma yöntemleriyle puanlanmasının testin güvenirliğine ve geçerliğine etkisi nedir?
2) Çoktan seçmeli bir testte, seçeneklerin Madde Tepki Kuramı'na göre "1-0" puanlama, uzman yargısına dayalı seçenek ağırlıklandırma ve deneysel seçenek ağırlıklandırma yöntemleriyle puanlanmasının testin güvenirliğine ve geçerliğine etkisi nedir?

YÖNTEM

Çalışma grubu

Araştırmanın verileri, sözel akıl yürütme yeteneğini ölçmeyi amaçlayan, 4 seçenekli 18 maddelik bir çoktan seçmeli testin, 2006-2007 eğitim-öğretim yılı bahar döneminde Hacettepe ve Gazi Üniversitesi'nin Ankara ili merkezinde bulunan çeşitli fakültelerinin farklı bölümlerinde lisans programlarına devam eden toplam 1593 öğrenciye tek oturumda uygulanmasıyla elde edilmiştir. Çalışma grubunun heterojenliğini sağlamak amacıyla bölümlerin çeşitlilik göstermesine dikkat edilmiş, bu bölümlerin ÖSS'de hangi puan türüne (Sayısal, Sözel veya Eşit Ağırlık) göre öğrenci aldığı göz önünde bulundurularak çalışma grubunda her puan türü için yüksek, orta ve düşük puanlarla öğrenci alan bölümlere yer verilmesine özen gösterilmiştir.

Veri toplamada kullanılan test araştırmacı tarafından geliştirilmiştir. Bu nedenle, önce aracın deneme uygulaması yapılmış, deneme formu 2006-2007 eğitim-öğretim yılı güz döneminde Ankara Üniversitesi'nin Ankara ili merkezinde bulunan çeşitli fakültelerinin farklı bölümlerinde okuyan toplam 468 öğrenciye 60 dakikalık tek bir oturumda uygulanmıştır.

Veriler ve toplanması

Kısmi bilgiyi ölçmeyi amaçlayan madde yapıları geleneksel yapıdaki maddelerden farklı olarak, yanlış seçeneklerden bir ya da birkaçını içeren doğru yanıtları barındıran (Waters, 1976) ya da o maddeyle ölçülen davranışa sahip olma açısından sıralanabilir (De Ayala, 1993) düzeylerde bilgi içeren nitelikte olmalıdır. Bu özelliklere sahip bir aracın bulunmaması nedeniyle araştırmacı tarafından seçenekleri belli düzeyde doğruluk içeren maddelerden oluşan ve sözel akıl yürütme yeteneğini ölçmeyi amaçlayan bir çoktan seçmeli test geliştirilmiştir.

*Testin deneme formunun geliştirilmesi ve uygulanması***.** Bu aşamada öncelikle, sözel akıl yürütme yeteneğini ölçen sorularla yoklanan davranış örnekleri ve araştırmada kullanılan puanlama yöntemlerinin temel özellikleri dikkate alınarak testte yer alan maddelerin "en doğru cevabı istenen" türde olmasına ve ölçülen davranışların şu çerçevede olmasına karar verilmiştir:

1. Cümledeki boşluğu anlam açısından uygun biçimde tamamlayan sözcüğü bulma,
2. Metinde belirtilen duruma dayanarak ulaşılabilecek yargıları bulma,
3. Metindeki ana düşünceyi, asıl söylenmek isteneni ya da parçanın temel amacını bulma,
4. Cümlenin taşıdığı duygu, düşünce ve anlam yönünden eş ya da yakın anlamlısını bulma,
5. Parçaya uygun başlığı seçme,
6. Parçada kullanılan dile, anlatıma dayanarak yazarın tutumunu, yaklaşımını belirleme,
7. Verilen sözcük grubuyla eş anlamlı/ yakın anlamlı seçeneği bulma,
8. Parçada değinilen bilgiye/ görüşe anlamca en yakın görüşü bulma,
9. Bir durumdan çıkarılabilecek sonucu veya verilen duruma neden olabilecek durumu bulma.

Ölçülecek davranışlar belirlendikten sonra, Lisansüstü Eğitim Sınavı (LES), Milli Eğitim Bakanlığı Ortaöğretim Kurumları Öğrenci Seçme ve Yerleştirme Sınavı (OKS), Kamu Personeli Seçme Sınavı (KPSS) ve Öğrenci Seçme Sınavı (ÖSS) ile ilgili hazırlık, deneme sınavı ve geçmiş yıllardaki sınav soruları incelenmiştir. Bunlar arasından ilk aşamada 75 madde seçilerek üzerlerinde değiştirme/geliştirme çalışmaları (yeni madde kökü veya seçenek yazılması) yürütülmüştür. Bu aşamada, seçenek puanlarını birbirinden bağımsız şekilde belirleyen uzmanlar arasında tutarlılık sağlamada yaşanan güçlükler ve madde yazarı tarafından seçeneklerin küçük bir uzman grubuyla önceden ağırlıklandırılarak uzman görüşüne sunulması önerileri (Echternacht, 1976) dikkate alınarak, formda yer alan tüm madde seçenekleri bu iki uzman yardımıyla puanlanmıştır.

Her bir madde için seçenek puanlamaları "1-4" puan arasında yapılmıştır ve "4" puan o madde için en doğru cevabı içeren seçeneği, "1" puan ise yine doğru olan ancak doğruluğu açısından en uzak olan cevabı içeren seçeneği göstermektedir. Puanlamalar iki dil uzmanı ve araştırmacıdan oluşan üç kodlayıcı tarafından hem ayrı zamanlarda, hem de aynı kodlayıcılar tarafından iki farklı zamanda yapılmıştır. Kodlayıcılar arasında tutarlılığın gözlenemediği ve dilbilgisi ve yazım kuralları açısından kusurlu olduğu düşünülen 13 madde ile maddeleri soru yazma teknikleri ve psikometrik teknikler açısından inceleyen dört ölçme ve değerlendirme uzmanının görüşü doğrultusunda 6 madde test formundan çıkarılmıştır. Hazırlanan 56 maddelik son test formu, beşi Türkçe öğretmeni ve beşi Türk Dili ve Edebiyatı uzmanı olmak üzere on uzmana verilmiş, uzmanlardan her madde için verilen ağırlıklı puanların uygun olup olmadığını belirlemeleri, uygun olmadığını düşündükleri seçenek puanları için yeni puanlar önermeleri istenmiştir. Uzmanlardan alınan ağırlıklı seçenek puanları arasındaki uyum Kendall'ın W uyum katsayısı ile incelenmiş, güvenirlik analizi sonuçları Tablo 1'de sunulmuştur.

Tablo 1. *Uzman yargısına dayalı seçenek ağırlıklandırma yöntemi için güvenirlik analizi*

Madde No	Kendall W	N	sd	χ^2	p	Madde No	Kendall W	N	sd	χ^2	p
1	0,60	10	3	18,00	0,00	29	0,94	10	3	28,08	0,00
2	0,96	10	3	28,92	0,00	30	0,88	10	3	26,40	0,00
3	0,92	10	3	27,72	0,00	31	0,96	10	3	28,92	0,00
4	0,60	10	3	18,12	0,00	32	0,94	10	3	28,08	0,00
5	0,86	10	3	25,68	0,00	33	0,85	10	3	25,56	0,00
6	0,83	10	3	24,84	0,00	34	1,00	10	3	30,00	0,00
7	0,94	10	3	28,08	0,00	35	0,47	10	3	14,04	0,03
8	0,93	10	3	27,84	0,00	36	0,94	10	3	28,08	0,00
9	0,71	10	3	21,36	0,00	37	0,94	10	3	28,08	0,00
10	0,71	10	3	21,24	0,00	38	0,85	10	3	25,56	0,00
11	0,36	10	3	10,92	0,01	39	0,94	10	3	28,08	0,00
12	0,61	10	3	18,36	0,00	40	0,50	10	3	14,88	0,02
13	0,96	10	3	28,92	0,00	41	0,92	10	3	27,48	0,00
14	0,84	10	3	25,32	0,00	42	0,96	10	3	28,92	0,00
15	1,00	10	3	30,00	0,00	43	0,94	10	3	28,08	0,00
16	0,92	10	3	27,72	0,00	44	0,92	10	3	27,48	0,00
17	0,88	10	3	26,52	0,00	45	0,80	10	3	23,88	0,00
18	0,96	10	3	28,92	0,00	46	0,80	10	3	23,88	0,00
19	0,76	10	3	22,68	0,00	47	0,94	10	3	28,08	0,00
20	0,78	10	3	23,52	0,00	48	1,00	10	3	30,00	0,00
21	0,75	10	3	22,44	0,00	49	1,00	10	3	30,00	0,00
22	0,75	10	3	22,44	0,00	50	1,00	10	3	30,00	0,00
23	0,84	10	3	25,32	0,00	51	0,74	10	3	22,32	0,00
24	0,93	10	3	27,84	0,00	52	0,94	10	3	28,08	0,00
25	0,88	10	3	26,40	0,00	53	0,93	10	3	27,84	0,00
26	0,94	10	3	28,08	0,00	54	1,00	10	3	30,00	0,00
27	0,56	10	3	16,92	0,01	55	0,96	10	3	28,92	0,00
28	1,00	10	3	30,00	0,00	56	0,71	10	3	21,36	0,00

Tablo 1 incelendiğinde, tüm W uyum katsayılarının 0,05 düzeyinde manidar olduğu, bir başka deyişle puanlayıcıların madde seçenekleri için yaptığı ağırlıklandırmaların genel olarak tutarlı olduğu gözlenmektedir. Ancak, bu puanlamada bir seçenek için önerilen puanların ortalaması o seçeneğin ağırlıklı puanı (1,2,3,4) olarak kullanıldığından, iki ya da daha fazla seçeneğinin ağırlıklı puan ortalamaları binişik olan altı madde test formundan çıkarılmıştır. Böylece testin 50 maddelik deneme formu oluşturulmuş, Ankara Üniversitesi'nin farklı bölümlerinde okuyan 468 öğrenciye uygulanmıştır.

Madde seçimi. Nihai test için madde seçiminde veri, hem Klasik Test Kuramı'na hem de Madde Tepki Kuramı'na göre çözümlenmiştir. KR-20 güvenirliği 0,82 olarak kestirilen deneme formunda yer alan maddeler için, Klasik Test Kuramı kapsamında madde güçlük indeksi (p_j) ve çift serili korelasyon katsayısına (r_{bis}) dayalı madde ayırıcılık gücü indeksi (r_{jx}) değerleri elde edilmiş; p_j değeri 0,50 komşuluğunda olan ve r_{jx} değeri 0,30 ve daha büyük olan maddelerin seçimine dikkat edilmiştir.

Bu çalışma kapsamında Madde Tepki Kuramı'nın çok kategorili modellerinden biri olan kısmi puan modeli dikkate alınmaktadır. Dolayısıyla, verinin Madde Tepki Kuramı'na göre çözümlenmesinde öncelikle madde olasılıklarını örtük özelliğin düzeyi (θ) ve b_i madde güçlük parametresi yardımıyla yordayan (Embretson ve Reise, 2000, 60) bir parametreli lojistik model dikkate alınmıştır. Ancak, bu kurama göre madde seçiminde b_i parametresinin yanı sıra a_i ayırıcılık gücü parametresi değeri de önemli bir ölçüttür (Hambleton, Swaminathan ve Rogers, 1991, 15). Nihai testte, hem bir parametreli hem de iki parametreli lojistik model analizine göre madde-model uyumunun ölçüsü olan χ^2 değeri manidar olmayan (α=0,01), bir başka deyişle her iki modele de uyumlu olan maddeler seçilmiştir. Bunlara ek olarak, madde güçlükleri dağılımının normal olmasına (Baykul, 2000, 345) da dikkat edilmiş; nihai test için, 2'si çok zor (dağılımın yaklaşık %1,5'i), 2'si zor (dağılımın yaklaşık %13'ü), 11'i orta güçlükte (dağılımın yaklaşık %68'i), 2'si kolay ve 1'i çok kolay maddelerden oluşan toplam 18 madde seçilmiştir. Madde analizi sonuçlarına Tablo 2'de yer verilmiştir.

Tablo 2. *Deneme uygulamasından elde edilen madde analizi sonuçları*

Madde No	Klasik Test Kuramı		Madde Tepki Kuramı			Seçilen Maddeler	Madde No	Klasik Test Kuramı		Madde Tepki Kuramı			Seçilen Maddeler
			1P*	2P**						1P	2P		
	p_j	r_{jx}	b_i	b_i	a_i			p_j	r_{jx}	b_i	b_i	a_i	
1	0,62	0,51	-0,77	-0,57	0,67		26	0,53	0,33	-0,21	-0,25	0,37	✓
2	0,65	0,43	-0,93	-0,85	0,48	✓	27	0,54	0,39	-0,23	-0,22	0,48	✓
3	0,49	0,30	0,08	0,09	0,24		28	0,91	0,24	-3,58	-4,28	0,35	
4	0,46	0,29	0,23	0,26	0,35		29	0,85	0,17	-2,66	-3,66	0,30	
5	0,68	0,39	-1,16	-1,06	0,48	✓	30	0,50	0,44	0,03	0,00	0,52	✓
6	0,66	0,38	-1,05	-0,96	0,48	✓	31	0,69	0,43	-1,24	-1,05	0,53	
7	0,59	0,30	-0,59	-0,71	0,25		32	0,82	0,29	-2,35	-2,40	0,41	
8	0,74	0,65	-1,58	-0,90	1,02		33	0,70	0,51	-1,29	-0,90	0,71	
9	0,73	0,39	-1,51	-1,37	0,48	✓	34	0,86	0,28	-2,73	-3,15	0,36	
10	0,92	0,22	-3,75	-4,21	0,37		35	0,24	0,20	1,84	3,57	0,20	
11	0,65	0,30	-0,93	-1,09	0,35		36	0,12	0,21	3,07	4,15	0,30	
12	0,93	0,20	-3,89	-4,77	0,33		37	0,40	0,42	0,64	0,54	0,50	✓
13	0,34	0,44	1,01	0,80	0,55		38	0,34	0,39	1,04	0,99	0,43	✓
14	0,71	0,50	-1,40	-1,02	0,66		39	0,94	0,20	-4,04	-5,90	0,33	
15	0,59	0,33	-0,55	-0,60	0,39	✓	40	0,90	0,26	-3,33	-3,48	0,40	
16	0,53	0,31	-0,17	-0,22	0,35	✓	41	0,34	0,35	1,05	1,07	0,41	✓
17	0,36	0,19	0,92	1,94	0,18		42	0,84	0,29	-2,49	-2,80	0,37	
18	0,71	0,47	-1,35	-1,00	0,64	✓	43	0,52	0,35	-0,12	-0,14	0,40	✓
19	0,51	0,30	-0,06	-0,09	0,33	✓	44	0,57	0,37	-0,47	-0,46	0,45	✓
20	0,87	0,45	-2,95	-2,12	0,65		45	0,91	0,16	-3,47	-4,78	0,30	
21	0,24	0,33	1,82	1,86	0,41		46	0,78	0,25	-1,95	-2,43	0,38	
22	0,95	0,30	-4,38	-4,04	0,46		47	0,58	0,34	-1,52	-1,57	0,39	
23	0,81	0,27	-2,18	-2,44	0,37		48	0,90	0,39	-3,26	-2,64	0,55	
24	0,85	0,38	-2,66	-2,32	0,50		49	0,45	0,40	0,30	0,27	0,45	✓
25	0,31	0,11	1,24	3,25	0,15		50	0,59	0,49	-0,59	-0,47	0,61	✓

*Bir Parametreli Lojistik Model **İki Parametreli Lojistik Model

Verilerin çözümlenmesi

Veriler için öncelikle Madde Tepki Kuramı modellerinin ortak varsayımları olan normallik, tek boyutluluk, yerel bağımsızlık ile bir parametreli lojistik model açısından önemli olan madde ayırıcılık gücü parametrelerinin eşitliği ve şans başarısının minimum düzeyde olması test edilmiştir.

Güvenirlik kestirimlerinde, Klasik Test Kuramı'na göre ağırlıklı puanlamalar için α, "1-0" puanlama için KR-20, Madde Tepki Kuramı'na göre ağırlıklı puanlamalar için marjinal güvenirlik (Thissen, 1991), "1-0" puanlama için Lord'un güvenirlik katsayısından (Lord 1980, 52) yararlanılmıştır.

Farklı puanlama yöntemlerinin test geçerliğine etkisini belirleyebilmek için Klasik Test Kuramı kapsamında cevaplayıcıların araştırmada kullanılan testten aldıkları puanlar ile Türk Dili Sözlü Anlatım dersinden aldıkları notlar arasındaki ilişkiye bakılmış, Madde Tepki Kuramı kapsamında ise cevaplayıcıların θ yetenek düzeyleri ile ölçüt puanlar arasındaki ilişki incelenmiştir. Bu ilişkinin yönü ve düzeyi ile ilgili bilgi Pearson Momentler Çarpımı Korelasyon Katsayısı ile elde edilmiştir .

Çalışmada kullanılan bir diğer geçerlik kanıtı, ÖSS-Sözel ve ÖSS-Sayısal puan türüne göre öğrenci alan bölümlerde okuyan öğrencilerin Klasik Test Kuramı kapsamında elde edilen test puanları ortalamaları arasında manidar bir fark olup olmadığının farklı puanlama yöntemleri açısından incelenmesidir. Aynı karşılaştırma, öğrencilerin Madde Tepki Kuramı kapsamında elde edilen θ yetenekleri için de yapılmıştır. İlişkisiz örneklemler için t testinden yararlanarak yapılan bu karşılaştırmadan elde edilen sonuçların yapı geçerliğinin göstergesi olabileceği düşünülmüştür. Analizlerde 0,01 manidarlık düzeyi benimsenmiş, verinin farklı kuramlara göre çözümlenmesinde EXCEL 7.0, ITEMAN 3.5, SPSS 12.00, STATISTICA 5.00, BILOG 3.00 ve MULTILOG 7,00 programlarından yararlanılmıştır.

BULGULAR

Madde Tepki Kuramı'nın varsayımlarının test edilmesi

Madde Tepki Kuramı modellerinin avantajları, test verisinin bu modellerin uygulamaları altında yatan varsayımlarla tutarlılığı ölçüsünde elde edilebilmektedir (Embretson ve Reise, 2000, 233; Hambleton, Swaminathan ve Rogers, 1991, 9). Bu nedenle, model-veri ve madde-veri uyumunun değerlendirilmesinde öncelikle verinin kuramın varsayımlarını karşılayıp karşılamadığı test edilmiştir.

Dağılımın normalliğinin test edilmesi. Dağılımın normalliğinin test edilmesinde, uygulamalardan elde edilen bazı betimsel istatistikler ile Kolmogorov-Smirnov tek örneklem testinden yararlanılmıştır.

Tablo 3. *Deneme uygulaması ve nihai test verilerine ilişkin betimsel istatistikler*

	Deneme Formu	Nihai Test
Madde Sayısı	50	18
Cevaplayıcı Sayısı	468	1593
Aritmetik Ortalama	31,69	9,49
Standart Sapma	6,65	3,24
Ortanca	32,00	10,00
Mod	31,00	10,00
En Düşük Puan	14,00	1,00
En Yüksek Puan	48,00	18,00
Testin Ortalama Güçlüğü	0,63	0,53
Bağıl Değişim Katsayısı	20,81	34,14
Çarpıklık Katsayısı	0,01	-0,06
Basıklık Katsayısı	-0,34	-0,19

Basıklık ve çarpıklık katsayılarının 0,00'a çok yakın değerler alması ile birlikte Tablo 3'te yer alan diğer betimsel istatistikler göz önüne alındığında, her iki uygulamadan elde edilen verinin dağılımının normale yakın olduğu kabul edilebilir. Bununla birlikte, Kolmogorov-Smirnov testinin sonuçları, $D>D_{tablo}$ durumunda gözlenen dağılımla teorik dağılımın birbiriyle uyumlu olduğunu (Siegel, 1956/1977, 53) söyleyen H_0 hipotezinin reddedilmesi kuralına göre yorumlandığında, hem deneme uygulaması hem de nihai test için elde edilen dağılımın teorik dağılımla uyumlu olduğu ($p<0,01$) gözlenmiştir.

Tek boyutluluk varsayımının test edilmesi. "Test performansını etkileyen baskın bir bileşen veya faktörün varlığı"nı gerektiren (Hambleton ve Swaminathan, 1985, 16) bu varsayımın test edilmesinde maddeler arası tetrakorik korelasyon matrisine dayalı temel bileşeler analizinden yararlanılmıştır. Bu analize göre, deneme uygulaması verilerine ilişkin özdeğerler λ_1:9,13, λ_2: 3,78, λ_3: 3,12, λ_4: 2,39, λ_5: 2,23, nihai test verisinden elde edilen özdeğerler ise sırasıyla λ_1:3,74, λ_2: 1,32 , λ_3: 1,15, λ_4: 1,12 ve λ_5: 1,05'tir. Özdeğer eğrileri, her iki test için de ilk faktörden sonra belirgin bir düşüş olduğunu göstermiş, ilk faktöre ait özdeğerin ikinci özdeğerlerden yaklaşık olarak üç kat büyük olması, ikinci özdeğerle diğer özdeğerler arasında büyük fark bulunmaması ve faktör yüklerinin büyüklüğü için 0,30 değerinin ölçüt olarak kabul edilmesi durumuna göre maddelerin büyük çoğunluğunun her iki test için de birinci boyutta yüksek yüke sahip olması (deneme formuna yer alan 43 madde, nihai testte yer alan 17 madde) (Lord, 1980, 21) göz önünde bulundurularak veri kümesinin tek boyutlu olarak ele alınabileceği kabul edilmiştir.

Yerel bağımsızlık varsayımının test edilmesi. Bu varsayım ile, test performansını etkileyen yetenek sabit tutulduğunda, bir madde çiftine verilen cevapların istatistiksel olarak bağımsız olduğu kabul edilmektedir. İki değişkenin ilişkili olması, aralarında ortak bir özelliğin söz konusu olduğu anlamına gelir ki faktör analizinin temelinde yatan prensip, değişkenlerin ilişkili olmasına neden olan ortak faktörün sabitlenmesiyle değişkenlerin ilişkisiz konuma gelmesidir. Benzer şekilde, Madde Tepki Kuramı'na göre yetenek faktörü sabitlendiğinde, cevaplayıcının maddelere verdiği yanıtların bağımsız olması beklenir. Bu nedenle tek boyutluluk varsayımı karşılandığında yerel bağımsızlık varsayımı da karşılanmış olur (Hambleton, Swaminathan ve Rogers, 1991, 11). Bu bilgiye dayanarak araştırmada tek boyutluluk varsayımının karşılanmasıyla yerel bağımsızlık koşulunun da karşılandığı kabul edilebilir. Bununla birlikte farklı yetenek grupları için maddeler-arası korelasyonlara ilişkin betimsel değerler de karşılaştırılmış (Hambleton, Swaminathan ve Rogers, 1991), test puanı aralıklarının belirlenmesinde puan dağılımının alt ve üst uçlarındaki %20'lik puan dilimleri ölçüt olarak kullanılmıştır.

Tablo 4. *Farklı yetenek gruplarından elde edilen maddeler arası korelasyonlara ilişkin karşılaştırmalar*

	Yetenek Grupları	Puan Aralıkları	N	$\overline{X}$	Minimum	Maksimum	Ranj	S_x^2
Tüm	Deneme Formu	14-48	468	0,07	-0,13	0,29	0,42	0,01
	Nihai Test	1-18	1593	0,09	-0,02	0,20	0,22	0,00
Üst	Deneme Formu	41-48	45	-0,02	-0,43	1,00	1,43	0,02
	Nihai Test	15-18	52	-0,04	-0,18	0,61	0,80	0,01
Alt	Deneme Formu	14-21	26	-0,01	-0,70	0,75	1,45	0,04
	Nihai Test	1-4	63	-0,04	-0,04	-0,31	0,26	0,57

Tablo 4'te yer alan değerler incelendiğinde, üst ve alt gruptan elde edilen korelasyonların ortalamasının tüm gruba ilişkin ortalamadan düşük ve 0,00'a oldukça yakın olduğu gözlenmektedir. Bu bulgular, yerel bağımsızlık koşulunun karşılandığı yönünde destekleyici bir bilgi olarak kabul edilmiştir.

Madde ayırıcılık gücü (a_i) parametrelerinin eşitliğinin test edilmesi. Hambleton, Swaminathan ve Rogers (1991, 56) ranj veya standart sapma değerlerinin incelenerek ayırıcılık gücü indekslerine ilişkin dağılımın homojen olduğuna ilişkin elde edilecek bulguyla, a_i parametrelerinin eşitliğini savunan bir model kullanmanın model-veri uyumu açısından uygun olabileceğini belirtmektedir. Bu çalışmada a_i parametrelerinin eşitliği nihai testteki 18 madde için test edilmiş, öncelikle ayırıcılık gücü indekslerinin ranjları incelenmiştir. Bu değerler incelendiğinde, r_{pbis} yardımıyla elde edilen en düşük madde ayırıcılık gücü indeksi değerinin 0,30, en yüksek değerin ise 0,48 olduğu, r_{bis} yardımıyla elde edilen en düşük madde ayırıcılık gücü indeksi değerinin 0,38, en yüksek değerin ise 0,62 olduğu gözlenmiştir.

Dağılımla ilgili daha fazla bilgi elde edebilmek için madde ayırıcılık gücü indekslerinin ortalamaları ve standart sapmaları hesaplanmıştır. Nokta çift serili korelasyon katsayısının kullanıldığı durumda elde edilen ortalamanın 0,38, standart sapmanın 0,05 olduğu, çift serili korelasyon katsayısının kullanıldığı durumda elde edilen ortalamanın 0,48, standart sapmanın ise 0,06 olduğu gözlenmiştir. Her iki korelasyon katsayısı yardımıyla hesaplanan madde ayırıcılık gücü indekslerinin ranjlarının ve standart sapmalarının çok küçük olması, bu değerlere ilişkin dağılımın oldukça homojen olduğu şeklinde değerlendirilmiş; bu bulgu, a_i parametrelerinin eşitliğinin kabul edilebileceği yönünde yorumlanmıştır.

Şans başarısının minimum düzeyde olmasının test edilmesi. Testte yer alan maddelere verilen cevapların, tahminle cevap verme davranışının bir sonucu olup olmadığının belirlenmesinde düşük test puanı alan öğrencilerin zor test maddeleri üzerindeki performansı dikkate alınmıştır. Nihai test maddeleri üzerindeki şans başarısının minimum düzeyde olduğunu kanıtlayabilmek için; test puanı 4'ün altında olan 63 kişilik alt grubun ve test puanı 15'in üstünde olan 52 kişilik üst grubun testte yer alan en zor beş maddeyi doğru cevaplandırma yüzdeleri Tablo 5'te sunulmuştur.

Tablo 5. *Alt ve üst yetenek grubunun nihai testte yer alan zor maddeler üzerindeki performansı*

Madde No	Madde Güçlük Düzeyi	Maddenin Doğru Cevaplandırılma Yüzdesi	
		Alt Grup (n<4)	Üst Grup (n>15)
6	0,35	0,06	0,83
9	0,42	0,09	0,79
13	0,35	0,06	0,87
14	0,32	0,11	0,96
17	0,43	0,09	0,92

Alt grubun testte yer alan en zor beş madde üzerindeki performansının oldukça düşük olduğu gözlenmektedir. Zor maddeler için bu gruptaki 0,00'a yakın cevaplandırılma yüzdeleri ve üst grubun aynı maddeler üzerindeki 1,00'e yakın doğru cevap yüzdeleri göz önünde bulundurulduğunda şans başarısının minimum düzeydeki varlığı kabul edilebilir.

Model-veri uyumunun değerlendirilmesi. Madde Tepki Kuramı'nda model varsayımlarının karşılanıp karşılanmadığını belirlemek için yapılan testler uyum iyiliği testleri olmadığı halde, uyum iyiliği testleriyle elde edilen bulguların yorumlanmasında kullanımı ve model seçimi konusundaki temel rolü nedeniyle model-veri uyumu araştırmalarında öncelikli olarak ele alınmaktadır. Bu durum göz önünde bulundurularak; verinin Madde Tepki Kuramı'nın varsayımlarıyla tutarlılığına ilişkin buraya

kadar yapılan çalışmalarla ortaya konan tüm olumlu sonuçların model-veri uyumunun genel bir göstergesi olarak kabul edilebileceği düşünülmektedir. Bununla birlikte, "madde-model" uyumunun incelenmesi (Embretson ve Reise, 2000) çalışması da yürütülmüş, deneme formundan elde edilen verilerin bir parametreli lojistik modele göre analiz edilmesi sonucunda, 50 maddeden 40'ının uyum ölçüsü olan χ^2 değerinin 0,01 düzeyinde manidar olmadığı, bir başka deyişle bu maddelerin bir parametreli modele uyumlu olduğu belirlenmiştir.

Madde-veri uyumunun yanı sıra model-veri uyumunun belirlenmesi söz konusu olduğunda önerilen bir yol, "*(S-1) – 2n(r-1)"* serbestlik derecesinde χ^2 dağılımı gösteren –2 log olabilirlik (–2 log λ) istatistiğinden yararlanmaktır (Bock, 1997). *–2 log λ / (S-1) – 2n(r-1) ≤ 3,00* koşulunu sağlayan uyum değerlerinin, model için tatmin edici bir uyumun göstergesi olarak kabul edilebileceği belirtilmektedir (Drasgow, Levine, Tsien, Williams ve Mead, 1995; Bock, 1997). Uzman yargısına dayalı ve deneysel ağırlıkların kullanıldığı durumda verinin kısmi puan modeline göre analiz edilmesi sonucu elde edilen *-2log λ* istatistiği değeri ise 34746,4'tür. Bu değerler verilen bilgiler ışığında, nihai uygulamayla elde verinin "1-0" puanlama yöntemi için bir parametreli lojistik modelle, ağırlıklı puanlama yöntemleri için ise kısmi puan modeli ile uyumunun bir göstergesi olarak değerlendirilebilir.

Betimsel istatistikler

Sözel yetenek testine ilişkin elde edilen bazı betimsel istatistikler Tablo 6'da sunulmuştur. Uzman yargısına dayalı ve deneysel yöntemler için elde edilen betimsel istatistiklerin tümü, testte yer alan 18 maddenin seçenek ağırlıklandırma yöntemlerine göre puanlanmasıyla elde edilmiştir.

Tablo 6. *Çalışma grubuna ilişkin farklı puanlama yöntemleriyle elde edilen betimsel istatistikler*

Puanlama Yöntemleri	N	En Düşük Puan	En Yüksek Puan	$\overline{X}$	S_x
'1-0' Puanlama	1593	1,00	18,00	9,49	3,24
Uzman Yargısına Dayalı Puanlama		33,00	72,00	58,66	6,36
Deneysel Puanlama		33,00	72,00	59,03	6,29

Araştırmada kullanılan ağırlıklı puanlama koşulları altında cevaplayıcı grubundan elde edilebilecek puan aralığı 18≤X≤72, "1-0" puanlama yöntemine göre elde edilebilecek puan aralığı ise 1≤X≤18'dir. Kullanılan ağırlıklı puanlama yöntemlerinde test puanları ranjı için söz konusu olan bu genişleme, Haladyna (1990) tarafından da vurgulandığı gibi, test maddelerine verilen cevaplardan elde edilen bilginin düzeyinin ağırlıklı puanlama yöntemi lehindeki artışının bir göstergesidir.

Klasik Test Kuramı'na göre puanlama yöntemlerinin güvenirlik ve geçerliğe etkisi

Güvenirliğe ilişkin bulgular

Tablo 7'de, uygulamada kullanılan çoktan seçmeli test için Klasik Test Kuramı'na göre farklı puanlama yöntemlerinden elde edilen güvenirlik katsayıları verilmiştir.

Tablo 7. *Klasik Test Kuramı'na göre farklı puanlama yöntemlerinden elde edilen güvenirlik katsayıları*

Puanlama Yöntemleri	N	Güvenirlik Kestirme Yöntemleri	Güvenirlik Katsayıları
"1-0" Puanlama	1593	KR-20	0,64
Uzman Yargısına Dayalı Puanlama		α	0,68
Deneysel Puanlama		α	0,69

Tablo 7 incelendiğinde, Klasik Test Kuramı kapsamında her üç puanlama yöntemiyle elde edilen güvenirlik katsayısının da fazla yüksek olmadığı, bununla beraber deneysel seçenek ağırlıklandırma yöntemiyle kestirilen α katsayısının en yüksek güvenirliği (0,69) verdiği gözlenmektedir.

Güvenirlik, bir testin gerçek puanları kestirmedeki doğruluk derecesidir, böylelikle bir testin içerdiği alt test sayısı arttıkça her bir cevaplayıcının gerçek puanına yaklaşma olasılığı da artacak, testin içerdiği alt test sayısı (madde sayısı) azaldıkça güvenirlik de azalacaktır (Magnusson, 1966, 68). Yukarıda verilen bilgiler ışığında, çalışmada kullanılan testin güvenirliğinin yüksek olmamasının madde sayısına (18) bağlanabileceği düşünülmektedir. Bununla birlikte hata varyansının testin gerçek puanları doğrudan ortaya koyamamasından kaynaklandığı ve güvenirlik katsayısının değerinin gerçek puanlar örnekleminin heterojenliğine dayandığı düşünülürse, teste ilişkin elde edilen varyansın çok büyük olmamasının, bir

başka deyişle grubun fazla heterojen olmamasının Klasik Test Kuramı varsayımları altında testin güvenirliğinin düşük olmasının diğer bir nedeni olabileceği düşünülmektedir.

Klasik Test Kuramı varsayımları altında farklı puanlama yöntemlerine göre kestirilen her üç güvenirlik katsayısının değeri de birbirine oldukça yakındır (0,64, 0,68 ve 0,69) ve gerçekte elde edilen güvenirlik katsayıları arasında belirgin bir niceliksel fark gözlenememektedir. Ancak, α katsayısının test güvenirliği için bir alt sınır değeri verdiği (Lord ve Novick, 1968) göz önünde bulundurularak, madde puanlarını iki değerli bir değişken yapısından kurtararak bu çalışma kapsamında dört değerli bir değişken yapısına dönüştüren, madde varyanslarının büyümesine yol açan ve daha hassas ölçmeler yapılmasına izin veren ağırlıklı puanlama yöntemlerinin güvenirlik katsayılarında gerçekleştirdiği küçük farkın çalışma açısından önemli olduğu düşünülebilir.

Ağırlıklı puanlama yöntemlerinin test güvenirliği üzerindeki etkisinin çözümlenmesiyle ilgili elde edilen bu bulgu, Kansup ve Hakstain'ın (1975) uzman yargısına dayalı seçenek ağırlıklandırmanın geleneksel puanlama yöntemine göre avantaj sağlamadığı şeklindeki, Sympson ve Haladyna (1988) ve Frary (1989)'nin deneysel seçenek ağırlıklandırma yöntemlerinin kestirilen güvenirlik katsayılarında artışla sonuçlandığı ancak bu artışın anlamlı olmadığı şeklindeki ve Ben-Simon, Budescu ve Nevo (1997) tarafından yürütülen çalışmanın ağırlıklı puanlamaların iç tutarlılık güvenirliğini artırmada tek başına üstünlük sağlamadığı şeklindeki bulgularıyla tutarlılık göstermektedir.

Geçerliğe ilişkin bulgular

Bir testin kapsam yönünden ölçme konusu davranışları yeterli derecede temsil etmesi önemlidir. Bu anlamda öncelikle testin kapsam geçerliğinin belirlenmesi ile ilgili kanıtlar değerlendirilmiştir. Kapsam geçerliğinin belirlenmesinde kullanılan yollardan biri uzman yargısına başvurarak her test maddesinin ilgili davranışı yoklayıp yoklamadığı konusundaki görüşleri dikkate almaktır. Bu çalışmada kullanılan uzman yargısına dayalı seçenek ağırlıklandırma yöntemi doğrultusunda, testin deneme formunun hazırlanması aşamasında seçeneklere puan vermek üzere birden fazla uzman grubunun görüşüne başvurulmuş, uzmanlar sözel akıl yürütme yeteneğine ilişkin maddeleri hem kapsam hem de ağırlıklandırmaya uygunlukları açısından değerlendirmişlerdir. Uzmanların görüşleri doğrultusunda test formundan çıkarılan maddelerle ve verilen ağırlıklı puanlarla teste son biçimi verilmiştir. Böylelikle, geliştirilen sözel yetenek testinin kapsam geçerliğine sahip olduğu söylenebilir.

Bir testin geçerliğinin belirlenmesinde diğer bir yol, bir ölçme aracıyla ölçülmek istenen yapının o araçla ortaya konma derecesi olarak tanımlanan "yapı geçerliği" ile ilgili çalışmalar yürütmektir (Magnusson, 1966). Yapı geçerliğinin belirlenmesinde çoğunlukla testle ölçülen değişkenlerin boyutluluğunu ortaya koymaya yönelik faktör analizi çalışmalarından yararlanılır. Bu araştırmanın, Madde Tepki Kuramı'nda tek boyutluluk varsayımının karşılanıp karşılanmadığını belirlemede başvurulan temel bileşenler analizi sonuçlarına dayalı olarak varsayımın karşılandığının kabul edilebileceği yönündeki bulgusu dikkate alınarak, kullanılan testin yapı geçerliğine sahip olduğu ileri sürülebilir.

Testle ölçülmek istenen özelliğe üst düzeyde ve alt düzeyde sahip olması beklenen gruplardan elde edilen test puanı ortalamalarının karşılaştırılması yöntemi de testin yapı geçerliği ile ilgili kanıtlar elde etmede kullanılabilmektedir. Bu doğrultuda, özelliği bilinen iki grubun testle ölçülen özellik açısından karşılaştırılması yoluna başvurulmuş, ÖSS-Sözel puan türüne göre bölümlere yerleşen öğrencilerin ÖSS-Sayısal puan türüne göre bölümlere yerleşen öğrencilere göre sözel yetenek testiyle ölçülmek istenen özelliğe daha üst düzeyde sahip olabileceği düşünülerek, bu iki gruptan elde edilen test puanı ortalamaları Tablo 8'de karşılaştırılmıştır.

Tablo 8. *Klasik Test Kuramı'na göre farklı puanlama yöntemleri için elde edilen test puanı ortalamalarının karşılaştırılması*

Puanlama Yöntemleri	Gruplar	N	$\overline{X}$	S_x	sd	t	p
"1-0" Puanlama	Sözel	288	9,85	2,99	1015	3,69*	0,000
	Sayısal	729	9,04	3,24			
Uzman Yargısına Dayalı Puanlama	Sözel	288	59,29	5,78		3,34*	0,001
	Sayısal	729	57,82	6,53			
Deneysel Puanlama	Sözel	288	59,64	5,71		3,33*	0,001
	Sayısal	729	58,19	6,44			

*p<0,01

Tablo 8'den görülebileceği gibi, farklı iki gruba uygulanan sözel yetenek testinden alınan test puanı ortalamaları arasında, tüm puanlama yöntemleri için ÖSS-Sözel puan türüyle bölümlere yerleşen öğrencilerin puanlarının ortalaması lehinde 0,01 düzeyinde manidar bir fark elde edilmiştir. Bu bulgu, hem "1-0" hem de ağırlıklı puanlama yöntemlerinin sözel ve sayısal ağırlıklı bölümlerde okuyan öğrencilerin sözel yetenek düzeyleri arasında ayrım yapabildiği, böylece her üç puanlama yönteminin de geçerli olduğu şeklinde yorumlanabilir.

Çalışma kapsamında kullanılan üç farklı puanlama yönteminin test geçerliğine etkisini araştırmak amacıyla başvurulan diğer bir geçerlik kanıtlama yöntemi ölçüte dayalı geçerliktir. Çalışmada ölçüte dayalı geçerlik kanıtı olarak uygunluk geçerliğinden yararlanılmış, ölçüt puan olarak uygulamaya katılan öğrencilerin, lisans programlarının ilk yılında zorunlu olarak tüm bölümlere okutulan Türk Dili Sözlü Anlatım dersinden aldıkları öğretmen notlarının kullanılabileceği düşünülmüştür. Uygulamaya katılan öğrenciler arasından 2. ve daha üst sınıflarda okuyan toplam 168 öğrencinin Türk Dili Sözlü Anlatım dersi notları ile sözel yetenek testinden farklı puanlama yöntemlerine göre aldıkları puanlar arasındaki Pearson Momentler Çarpımı korelasyon katsayısı değerleri Tablo 9'da sunulmuştur.

Tablo 9. *Klasik Test Kuramı'na göre farklı puanlama yöntemleri için elde edilen geçerlik katsayıları*

Puanlama Yöntemleri	N	K	Geçerlik Katsayısı (r_{jX})
"1-0" Puanlama	168	18	0,55
Uzman Yargısına Dayalı Puanlama			0,52
Deneysel Puanlama			0,52

Tablo 9 incelendiğinde, "1-0" puanlama yöntemin kullanıldığı durumda elde edilen geçerlik katsayısının (0,55) ağırlıklı puanlama yöntemlerinin kullanıldığı durumda elde edilen geçerlik katsayılarından (0,52) yüksek olduğu gözlenmektedir. "1-0" puanlama yönteminin kullanıldığı durumda elde edilen geçerlik katsayısının diğer puanlamalara göre yüksek elde edilmesi, ölçüt olarak kullanılan ders notlarının da büyük olasılıkla geleneksel puanlamaları kullanan değerlendirme sistemleriyle verilmiş olması ile açıklanabilir. Buna benzer bir bulgu, Waters'ın (1976) çalışmasıyla elde edilmiş, çalışmanın sonucuna göre ağırlıklı puanlama yöntemlerinin kullanıldığı tüm durumlarda bir ölçüte dayalı geçerlik katsayıları ile ilgili bir düşüşün olduğu rapor edilmiştir. Ancak görülmektedir ki, ağırlıklı puanlama yöntemlerinin kullanıldığı durumda geçerlik katsayısında gözlenen düşüş niceliksel bir öneme sahip değildir. Puanlama yöntemleri açısından büyük bir fark göstermeyen geçerlik katsayıları, kullanılan tüm yöntemlerin test geçerliği üstünde aynı etkiyi yaptığı şeklinde yorumlanabilir.

Genel olarak, çoktan seçmeli sözel yetenek testine ilişkin geçerlik katsayısının, üç puanlama yönteminin kullanıldığı durumda da yüksek olmadığı görülmektedir. Geçerlik katsayılarının düşük elde edilmesinin nedenleri; ölçüt olarak kullanılan Türk Dili Sözlü Anlatım dersi notlarının elde edilmesinde kullanılan değerlendirmelerin içeriğinin çalışmada kullanılan testin içeriği ile bütünüyle örtüşmemesi, araştırmada kullanılan testin ölçtüğü sözel yetenek alanlarının sınırlılığı, öğretmen yapımı testlere dayalı olarak elde edilen başarı ölçülerine öznel bir takım değerlendirmelerin de karışabileceği dikkate alındığında, ölçüt puanların güvenirliğinin bu durumdan olumsuz etkilenebilmesi ve/veya çalışmada kullanılan sözel yetenek testine ilişkin elde edilen güvenirlik katsayısının istenilen düzeyde yüksek olmaması (0,64) olabilir. Bunlara ek olarak, farklı yöntemlere göre puanlanan testlere ilişkin geçerlik katsayılarının elde edilmesinde kullanılan 168 öğrenciden oluşan grubun homojenliğinin de bu sonucu doğurması mümkündür.

Yordamanın gücü ile yordayıcının güvenirliği arasındaki ilişki dikkate alındığında, bir testin bir ölçüte göre geçerliğinin, bu testin güvenirliğinin kare kökünden büyük olamayacağı bilinmektedir. Testin ölçüte dayalı geçerliği ile güvenirliği arasındaki ilişki $\rho(X,Y) \leq \sqrt{\rho(X,X')}$ ile gösterilmekte ve bu ilişki bir testin ölçüt geçerliğinin alabileceği değerin üst sınırını vermektedir (Gulliksen, 1950, 23). Bu ilişkiden yararlanılarak elde edilen geçerlik katsayılarının alabileceği en yüksek değerler Tablo 10'da verilmiştir.

Tablo 10. *Klasik Test Kuramı'na göre farklı puanlama yöntemleri için geçerliğin üst sınırı*

Puanlama Yöntemleri	N	Geçerliğin Üst Sınırı
"1-0" Puanlama	1593	0,80
Uzman Yargısına Dayalı Puanlama		0,82
Deneysel Puanlama		0,83

Tablo 10 göstermektedir ki, uygun bir ölçütün kullanılması durumunda, Klasik Test Kuramı kapsamında deneysel seçenek ağırlıklandırma yöntemine göre puanlanan test diğer puanlama yöntemlerine göre daha geçerli (0,83) sonuç verebilmektedir. Ancak, geçerliğin alabileceği en yüksek değerler arasındaki bu niceliksel fark kullanılan yöntemlerden birinin diğerine göre daha üstün olduğu şeklinde değerlendirme yapmak için yeterli değildir. Bu açıdan, Klasik Test Kuramı kapsamında farklı puanlamaların test geçerliği üstünde aynı etkiyi yaptığı şeklindeki yorum burada da yinelenebilir.

Farklı puanlama yöntemlerinin test geçerliği açısından tartışıldığı bu alt amacın çözümlenmesiyle elde edilen bulgular; Sabers ve White (1969)'ın "1-0" ve deneysel seçenek ağırlıklandırma yöntemleriyle elde edilen geçerlik kestirimlerinin neredeyse aynı olduğu ve Cross, Ross ve Geller'in (1980) deneysel seçenek ağırlıklandırma yöntemiyle elde edilen geçerliğin "1-0" puanlamayla elde edilene ancak eşit olabildiği yönündeki bulgularıyla örtüşmektedir. Benzer bulgular, uzman yargısına dayalı puanlamaların söz konusu olduğu durumlar için Hambleton, Rogers ve Traub'un (1970) çalışmalarında da elde edilmiştir.

Madde Tepki Kuramı'na göre puanlama yöntemlerinin güvenirlik ve geçerliğe etkisi

Güvenirliğe ilişkin bulgular

Tablo 11'de, uygulamada kullanılan çoktan seçmeli test için Madde Tepki Kuramı'na göre farklı puanlama yöntemleriyle elde edilen güvenirlik katsayıları verilmiştir.

Tablo 11. *Madde Tepki Kuramı'na göre farklı puanlama yöntemlerinden elde edilen güvenirlik katsayıları*

Puanlama Yöntemleri	N	Güvenirlik Kestirme Yöntemleri	Güvenirlik Katsayıları
"1-0" Puanlama	1593	Lord'un Güvenirlik Katsayısı	0,88
Uzman Yargısına Dayalı Puanlama		Marjinal Güvenirlik	0,71
Deneysel Puanlama		Marjinal Güvenirlik	0,71

Madde Tepki Kuramı'nda farklı puanlama yöntemlerine göre elde edilen güvenirlik katsayıları incelendiğinde, en yüksek güvenirliğin "1-0" puanlama yöntemi için Lord'un güvenirlik katsayısıyla (0,88) kestirildiği gözlenmektedir. Uzman yargısına dayalı ve deneysel puanlama yöntemleri için teste ilişkin kestirilen marjinal güvenirlik katsayıları (0,71) ise eşittir.

Üzerinde çalışılan gruba bağımlı olarak kestirilen ve bu nedenle genellenebilme özelliğinden yoksun olan klasik güvenirlik tanımlamaları, Madde Tepki Kuramı'nda yerini madde ve test bilgi fonksiyonlarıyla açıklanan ölçmenin doğruluğu kavramına bırakmıştır. θ yetenek düzeyinde bir testle elde edilen bilgi, bu yetenek düzeyindeki madde bilgi fonksiyonlarının toplamıyla tanımlandığından ve kestirilen yeteneğin standart hatasıyla ters yönlü olarak ilişkilendirildiğinden Tablo 11'de verilen marjinal güvenirlik değerleri (0,71) bir ortalama değer olarak ele alınmaktadır. Marjinal güvenirlikleri kestirilen seçenekleri ağırlıklandırılmış teste ilişkin θ yetenek aralıkları ve test bilgi fonksiyonları incelendiğinde, ölçeğin -3.00 ve -1,40 yetenek aralığındaki bireylere ilişkin parametreleri daha yüksek düzeyde bilgi ile kestirdiği, bir başka deyişle daha az hata ile ölçtüğü gözlenmektedir.

"1-0" yöntemine dayanan puanlamalarda kullanılmak üzere Lord (1980, 52) tarafından önerilen güvenirlik katsayısı ise marjinal güvenirlikten farklı olarak güvenirliği yetenek düzeyleri üzerindeki ortalama standart hataya dayalı olarak açıklayan Klasik Test Kuramı'yla ilişkilendirmektedir. Buna göre, "1-0" puanlama yönteminin kullanıldığı durumda, bir bireyin doğru cevap sayısına dayanan gerçek puanı bu bireyin gözlenen puanının beklenen değeridir ve bu değer belli bir θ yetenek düzeyindeki tüm bireyler için aynıdır. θ yetenek düzeyindeki bir cevaplayıcının i maddesini doğru cevaplandırma olasılığı $P_i(\theta)$ ile gösterildiğinde, $P_i(\theta)$ θ'nın artan bir fonksiyonu ise bir bireyin doğru cevap sayısına dayalı gerçek

puanı da yeteneğin artan bir fonksiyonudur. Başka bir deyişle, gerçek puan ve θ yeteneği farklı ölçek düzeylerinde aynı şeyi ifade ederler (Lord, 1980, 46). Aralarındaki tek ve en önemli fark, gerçek puan ölçeğinin testteki maddelere dayanırken θ yetenek ölçeğinin testteki maddelerden bağımsız olmasıdır.

Tablo 11'de verilen güvenirlik katsayıları ölçmenin doğruluğu anlamında değerlendirildiğinde, "1-0" puanlamanın kullanıldığı durumda farklı θ yetenek düzeylerindeki parametrelerin uzman yargısına dayalı ve deneysel seçenek ağırlıklandırmanın kullanıldığı duruma göre daha doğru kestirildiği belirtilebilir. Güvenirlik katsayıları arasındaki bu niceliksel farkın marjinal güvenirlik katsayısı ile klasik güvenirlik tanımıyla ilişkili Lord'un güvenirlik katsayısının elde edilmesinde kullanılan yaklaşımlar arasındaki yukarıda değinilen farklılıklardan kaynaklandığı düşünülmektedir.

Bu alt amacın çözümlenmesiyle elde edilen bulgu, Özdemir (2002) tarafından yürütülen çalışmanın, karşılaştırılan puanlama yöntemleri için elde edilen Lord'un güvenirlik katsayısının marjinal güvenirlik katsayısına göre yüksek elde edildiği şeklindeki bulgusuyla tutarlılık göstermektedir.

Geçerliğe ilişkin bulgular

Bu alt amaç için çalışma kapsamında kullanılan üç farklı puanlama yönteminin test geçerliğine etkisini araştırmak amacıyla öncelikle yapı geçerliğine dayalı kanıtlar ortaya konmaya çalışılmış, ÖSS-Sözel puan türüne göre bölümlere yerleşen öğrencilerin θ yetenek ortalamalarıyla ÖSS-Sayısal puan türüne göre bölümlere yerleşen öğrencilerin θ yetenek ortalamaları Tablo 12'de karşılaştırılmıştır.

Tablo12. *Madde Tepki Kuramı'na göre farklı puanlama yöntemleri için elde edilen θ yetenek düzeylerinin karşılaştırılması*

Puanlama Yöntemleri	Gruplar	N	$\overline{\theta}$	S_x	sd	t	p
"1-0" Puanlama	Sözel	288	0,11	0,93		3,67*	0,000
	Sayısal	729	-0,14	1,00			
Uzman Yargısına Dayalı Puanlama	Sözel	288	0,97	0,49	1015	3,39*	0,001
	Sayısal	729	0,85	0,53			
Deneysel Puanlama	Sözel	288	0,98	0,48		3,31*	0,001
	Sayısal	729	0,87	0,51			

*p<0,01

Tablo 12'ye göre, Madde Tepki Kuramı kapsamında kullanılan üç puanlama yöntemi açısından da sözel puan türüyle öğrenci alan bölümlerdeki cevaplayıcıların θ yetenek ortalamalarının, sayısal puan türüne göre öğrenci alan bölümlerdeki cevaplayıcıların θ yetenek ortalamalarından manidar bir şekilde yüksek olduğu gözlenmiştir. Bu bulgu, çalışmada kullanılan puanlama yöntemlerinin tümünün sözel ve sayısal ağırlıklı bölümlerde okuyan öğrencilerin sözel yetenek düzeyleri arasında ayrım yapabildiği ve kullanılan tüm puanlama yöntemlerinin geçerli olduğu şeklinde yorumlanabilir.

Bir ölçüte dayalı geçerlik için kanıt olarak kullanılmak üzere uygunluk geçerliği konusunda da çalışmalar yürütülmüş, bu amaçla 168 öğrencinin Türk Dili Sözlü Anlatım dersi notları ile sözel yetenek testinde yer alan maddelere verdikleri yanıtlardan kestirilen θ yetenek düzeyleri arasındaki ilişkinin yönü ve derecesi incelenmiştir. Bu ilişkiyi gösteren Pearson Momentler Çarpımı korelasyon katsayısı değerleri farklı puanlama yöntemlerine göre Tablo 13'te sunulmuştur.

Tablo 13. *Madde Tepki Kuramı'na göre farklı puanlama yöntemleri için elde edilen geçerlik katsayıları*

Puanlama Yöntemleri	N	K	Geçerlik Katsayısı (r_{jx})
"1-0" Puanlama	168	18	0,54
Uzman Yargısına Dayalı Puanlama			0,54
Deneysel Puanlama			0,53

Tablo 13 incelendiğinde, "1-0" ve uzman yargısına dayalı puanlama yöntemiyle elde edilen geçerlik katsayılarının eşit olduğu (0,54), deneysel puanlamayla elde edilen θ yetenek kestirimlerinin ise Türk Dili Sözlü Anlatım dersi notlarıyla 0,53 düzeyinde korelasyon verdiği gözlenmiştir. Bu alt amaçla ilgili bulgular gözden geçirildiğinde, bunların Klasik Test Kuramı kapsamında aynı etkinin araştırıldığı ilk alt amaçtan elde edilen bulgularla büyük ölçüde benzer olduğu gözlenmektedir. Her iki kurama göre elde edilen uygunluk geçerliğinin göstergesi kabul edilen korelasyon katsayısı değerleri neredeyse aynıdır. Bu benzerlik, Klasik Test Kuramı kapsamında elde edilen gözlenen test puanlarının, Madde Tepki Kuramı kapsamında elde edilen θ yetenek kestirimleriyle ilişkisine dayanmaktadır. Nitekim, cevaplayıcıların test puanları ile θ yetenek düzeyleri arasında tam (1,00) korelasyon bulunmaktadır ve bu durum Lord (1980, 40) tarafından belirtildiği gibi "gerçek puan" ve "θ yetenek" kavramlarının farklı ölçek düzeylerinde aynı şeyi ifade ettiklerinin bir göstergesidir.

Madde Tepki Kuramı kapsamında gözlenen geçerlik katsayısı değerleri, kullanılan ölçüte dayalı olarak "1-0" ve iki ağırlıklı puanlama yönteminin eş düzeyde uygunluk geçerliğine sahip olduğu şeklinde yorumlanabilir. Ancak bu yorumun bu çalışmada kullanılan ölçütle sınırlılığı önemle göz önünde bulundurulmalıdır. Bir ölçüte dayalı geçerlik katsayısı ile ilgili daha etkili bir bilgi elde edebilmek amacıyla testin geçerliği ve güvenirliği, bir başka deyişle yordamanın gücü ile yordayıcının güvenirliği arasındaki ilişkiden yararlanılmıştır. Bu ilişkinin dikkate alınmasıyla elde edilen geçerlik katsayılarının alabileceği en yüksek değerler Tablo 14'te verilmiştir.

Tablo 14. *Madde Tepki Kuramı'na göre farklı puanlama yöntemleri için geçerliğin üst sınırı*

Puanlama Yöntemleri	N	Geçerliğin Üst Sınırı
"1-0" Puanlama	1593	0,94
Uzman Yargısına Dayalı Puanlama		0,84
Deneysel Puanlama		0,84

Tablo 14 incelendiğinde, uygun bir ölçütün kullanılması durumunda "1-0" puanlama yöntemiyle elde edilebilecek geçerlik katsayısı değerinin 0,94'e kadar yükselebileceği gözlenmektedir. Bu katsayının uzman yargısına dayalı ve deneysel seçenek ağırlıklandırma yöntemlerinin kullanıldığı durumda alabileceği maksimum değer 0,84'tür. Bu bulgu, Madde Tepki Kuramı kapsamında ağırlıklı puanlamalar yerine geleneksel puanlamalar kullanılarak daha geçerli kestirimler yapılabileceği şeklinde yorumlanabilir.

Madde Tepki Kuramı'na göre farklı puanlama yöntemlerinin geçerliğe etkisinin incelendiği bu alt amaca ilişkin bulgular, Backhoff, Tirado ve Larrazolo'nun (2001), içerisinde Madde Tepki Kuramı modellerinin de yer aldığı farklı puanlamaları karşılaştırdıkları çalışmalarının, test geçerliğinin ağırlıklı puanlamalara dayalı bir avantaj sağlamadığı şeklindeki bulgusuyla örtüşmektedir. Buna ek olarak, Özdemir (2002) de Madde Tepki Kuramı kapsamında "1-0" puanlama yönteminin kullanıldığı durumda test geçerliği için gözlenen üst sınır değerinin ağırlıklı puanlamalara göre yüksek olduğunu gösterecek benzer bir bulgu elde etmiştir.

TARTIŞMA, SONUÇ ve ÖNERİLER

De Ayala (1993)'nın "bilginin fonksiyonu" ile ilgili yaptığı tanımlamalarda, bir maddeden bir cevaplayıcının yeteneği hakkında belli bir miktarda bilgi elde edilebildiği ve bunun gerçekte "cevaplayıcının yeteneği hakkındaki belirsizliğin maddenin büyük miktarda bilgi ile işleme alınması yoluyla giderilebileceği" anlamına geldiği vurgulanmaktadır. Bir maddeden elde edilebilecek bilgi miktarının artırılması isteği, çok kategorili maddelerin kullanımına yönelimi de arttırmaktadır. Çok kategorili maddeleri içeren testlerin iki kategorili maddeler içeren testlere tercih edilmesinin nedenini Ark (2001), daha az sayıda çok kategorili maddeyle aynı düzeyde test güvenirliğinin elde edilebiliyor olması ve bazı psikolojik özelliklere ilişkin maddelere verilen tepkilerin oranlı ölçeklerle ölçülebiliyor olmasına fırsat vermesi olarak tanımlamaktadır. Bu doğrultuda, bu çalışmada çoktan seçmeli bir test için farklı test kuramlarına göre iki kategorili ve çok kategorili puanlama yöntemleri, testlerin güvenirliği ve geçerliği açısından incelenmiştir.

Klasik Test Kuramı'nın temel varsayımlarından biri gerçek puanlarla ($T_{.j}$), hata puanları ($E_{.j}$) arasındaki ilişkiyle ilgilidir; bir M evrenindeki j ölçmeleri için $\rho(E_{.j}, T_{.j}) = 0$ olduğu, bir başka deyişle, gerçek puanlarla hata puanları arasındaki korelasyonun sıfıra eşit olduğu varsayılır. Bunu kanıtlamada başvurulan yöntem, j ölçmeleri için hata puanlarının τ_{ij} gerçek puanlar üzerine doğrusal regresyonundan yararlanmak suretiyle τ_{ij} 'ye ait hata puanlarının beklenen değerini incelemektir (Magnusson, 1966, 36). Klasik Test Kuramı'nda yer alan tüm ölçme modellerinde hatanın kestirimi bu tanımdan yola çıkılarak yapılır. Öte yandan, Madde Tepki Kuramı'nın da bir test verisine uygulanmasında en önemli aşama, kullanılan Madde Tepki Kuramı'nı karakterize eden parametrelerin kestirilmesidir. Madde Tepki Kuramı modellerinde bir maddenin doğru cevaplandırılma olasılığı cevaplayıcının θ yetenek düzeyine ve madde parametrelerine bağlıdır, ancak yetenek ve madde parametrelerinin her ikisi de bilinmemektedir. Kestirme problemi burada ortaya çıkar.

Klasik Test Kuramı kapsamında kullanılan regresyon analizi de benzer bir mantık içerir; bir değişkenle ilgili gözlenen değerlerle regresyon modeline uygun parametrelerin, yani regresyon katsayılarının kestirilmesi gerekir. Ancak bu noktada, regresyon modelleriyle Madde Tepki Kuramı modelleri arasındaki iki temel farka değinmek gerekir. İlki, regresyon modellerinin genellikle doğrusal olması, Madde Tepki modellerinin ise doğrusal olmamasıdır. İkincisi ve en önemlisi, yordayıcı (bağımsız) değişkenin (test ve madde puanlarının) regresyon analizinde gözlenebilir olmasıdır. Madde Tepki Kuramı'ndaki yordayıcı değişken θ ise gözlenebilir bir değişken değildir. Klasik Test Kuramı kapsamında hata kestirimlerinin dayandığı doğrusal regresyonda, modelle veri arasındaki en iyi uyum "en küçük kareler" (least square) ölçütüne dayalı olarak tanımlanmaktadır. Madde Tepki Kuramı modellerinde ise bu ölçüt kullanılmaz çünkü en küçük kareler kestiriminin özelliklerinin doğrusal olmayan modeller için tanımlanması zordur. Bunun yerine Madde Tepki Kuramı'nda parametreler "en çok olabilirlik" (maximum likelihood) ölçütüne dayalı olarak kestirilir (Hambleton, Swaminathan ve Rogers, 1991, 32-33).

Bu çalışmada farklı kuramlara dayalı olarak kestirilen güvenirlik ve geçerlik katsayıları arasında farkların olduğu gözlenmiştir, örneğin "1-0" ile puanlanmış verinin Klasik Test Kuramı'na göre kestirilen güvenirliği 0,64 iken -dolayısıyla geçerliğinin üst sınırı 0,80 iken- aynı veriden Madde Tepki Kuramı kapsamında elde edilen güvenirlik kestirimi 0,88'dir -dolayısıyla geçerliğinin üst sınırı 0,94'tür. Benzer farklar ağırlıklı puanlama yöntemlerinden elde edilen güvenirlik ve geçerlik kanıtları için de söz konusudur. Bu bilgiler ışığında, Klasik Test Kuramı ve Madde Tepki Kuramı kapsamında elde edilen geçerlik ve güvenirlik kestirimleri arasındaki farkın bu iki kuramın parametre kestiriminde kullandığı farklı yaklaşımlardan kaynaklandığı savunulabilir. Bununla birlikte, Klasik Test Kuramı'na dayalı uygulamalar deneysel puanlamanın iç tutarlılık anlamındaki test güvenirliğini artırmaya eğilimli olduğunu, Madde Tepki Kuramı kapsamında ise güvenirliğin kestirilmesinde "1-0" puanlamanın diğer yöntemlere göre daha etkili olduğunu göstermiştir.

Çalışmada kullanılan puanlama yöntemleri açısından çoktan seçmeli teste ilişkin uygunluk geçerliği kanıtları göz önünde bulundurulduğunda, puanlama yöntemlerinin testin uygunluk geçerliğini artırmadığı ya da çalışmada kullanılan hiçbir puanlama yönteminin uygunluk geçerliği anlamında bir diğerinden etkili olmadığı sonucuna ulaşılmıştır. Klasik Test Kuramı'na göre güvenirlik katsayılarına bağlı olarak belirlenen geçerliğin üst sınırı bağlamında deneysel ağırlıklandırmanın "1-0" ve uzman yargısına dayalı ağırlıklı puanlamaya göre daha geçerli sonuçlar verme eğiliminde olduğu gözlenmiştir. Ancak bu eğilim deneysel ağırlıklandırmanın geçerlik açısından üstünlüğünü kanıtlayacak düzeyde değildir. Madde Tepki Kuramı kapsamında ise, testin ölçüte dayalı geçerliği anlamında Klasik Test Kuramı kapsamında puanlama yöntemlerinin birbirine göre üstünlük göstermediği, Madde Tepki Kuramı kapsamında ise "1-0" puanlama yönteminin daha etkili olduğu sonucuna varılmıştır.

Çalışmada elde edilen sonuçlar doğrultusunda farklı Madde Tepki Kuramı modellerinin dikkate alınmasıyla yeni çalışmalar yürütülmesi; ağırlıklı puanlama yöntemlerinden etkili bir şekilde yararlanabilmek için bu yöntemlerle kullanılabilecek madde yazımında etkili teknikler geliştirilme çalışmaları yürütülmesi; ağırlıklı puanlama yöntemlerinin çalışılabilirliğinin cevaplayıcıların bu yöntemlerden yararlanabilme yeteneklerine bağlı olduğu göz önünde bulundurularak bu yöntemler açısından bireylerin ağırlıklandırılmış maddeleri cevaplama davranışları üzerinde puanlama yönergelerinin nasıl bir etkisinin olduğunun araştırılması ve farklı konu alanlarına yönelik başarı testleri, farklı çalışma grupları ve farklı geçerlik ölçütleri kullanılarak benzer çalışmaların yürütülmesi önerilmektedir.

KAYNAKÇA

Adams, R.J. (1988). Applying the partial credit model to educational diagnosis. *Applied Measurement in Education*, 1(4), 347-361.

Akkuş, O. (2000). *Çoktan seçmeli test maddelerini puanlamada, seçenekleri farklı biçimlerde ağırlıklandırmanın madde ve test istatistiklerine olan etkisinin incelenmesi*. Yayımlanmamış yüksek lisans tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Ark, L.A. (2001). Relationships and properties of polytomous Item Response theory models. *Applied Psychological Measurement*, 25(3), 273-282.

Backhoff, E.E., Tirado, F.S., & Larrazolo, N.R. (2001). Differential weighting of items to improve university admission test validity. *Electronic Journal of Educational Research,* 3(1), 21-31.

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik Test teorisi ve uygulaması*. Ankara: ÖSYM Yayınları.

Bayuk, R.J. (1973). The effects of choice weights and item weights on the reliability and predictive validity of aptitude-type tests [Abstract]. *ERIC Digest,* Washington DC: ERIC Clearinghouse on Assessment and Evaluation. (ERIC Document Reproduction Service No: ED078061).

Ben-Simon, A., Budescu, D.V. & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement*, 21(1), 65-88.

Bock, R.D. (1997). The nominal categories model (In W. J. van der Linden & R. K. Hambleton, Eds.), *Handbook of Modern Item Response Theory* (p.33-49), New York Inc.: Springer-Verlag.

Coombs, C.H., Milholland, J.E. & Womer, F.B. (1956). The assessment of partial knowledge. *Educational and Psychological Measurement*, 16, 13-37.

Corey, S.M. (1930). The effect of weighting exercises in a new-type examination. *Journal of Educational Psychology*, 21, 383-385.

Crehan K.D. & Haladyna T.M. (1994). A comparison of three linear polytomous scoring methods. *ERIC Digest,* Washington DC: ERIC Clearinghouse on Assessment and Evaluation. (ERIC Document Reproduction Service No: ED377246).

Crocker, L. & Algina J. (1986). *Introduction to classical and modern test theory*. Orlando: Harcourt Brace Jovanovich Inc.

Cross, L.H. & Frary, R.B. (1978). Empirical choice weighting under "guess" and "do not guess" directions. *Educational and Psychological Measurement*, 38, 613-620.

Cross, L.H., Ross, F.K. & Geller, E.S. (1980). Using choice-weighted scoring of multiple-choice tests for determination of grades in college courses. *Journal of Experimental Education*, 48, 296-301.

Davis, F.B. & Fifer, G. (1959). The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement*, 19, 159-170.

De Ayala, R.J. (1993). An introduction to polytomous Item Response theory models. *Measurement and Evaluation in Counseling and Development***,** 3, 172-189.

De Ayala, R.J., Dodd, B.G. & Koch, W.R. (1992). A comparison of the partial credit and graded response models in computerized adaptive testing. *Applied Measurement in Education*, 5(1), 17-34.

Dodd, B.G. (1984). Attitude scaling: A comparison of the graded response and partial credit latent trait models (Doctoral Dissertation, University of Texas at Austin, 1984). *Dissertation Abstracts International*, 45, 2074A.

Dodd, B.G. & Koch, W.R. (1987). Effects of variations in item stop values on item and test information in the partial credit model. *Applied Psychological Measurement*, 11, 371-384.

Downey, R.G. (1979). Item-option weighting of achievement tests: Comparative study of methods. *Applied Psychological Measurement*, 3, 453-461.

Drasgow, F., Levine, M.V., Tsien, S., Williams, B. & Mead, A.D. (1995). Fitting polytomous Item Response theory models to multiple-choice tests. *Applied Psychological Measurement*, 19(2), 143-165.

Embretson, S.E. & Reise, S.P. (2000). *Item Response theory for psychologists*. New Jersey: Lawrence Erlbaum Associates.

Echternacht, G. (1976). Reliability and validity of item option weighting schemes. *Educational and Psychological Measurement***,** 36, 301-309.

Frary, R. (1980). The effect of misinformation, partial information, and guessing on expected multiple-choice test item scores. *Applied Psychological Measurement*, 4(1), 79-90.

Frary, R. (1989). Partial credit scoring methods for multiple choice tests. *Applied Measurement in Education*, 2(1), 79-96.

Glas, C.A.W. & Verhelst, N.D. (1989). Extensions of the partial credit model. *Psychometrik*a, 54(4), 635-659.

Gözen, G. (2006). Kısa cevaplı ve çoktan seçmeli maddelerin "1-0" ve ağırlıklı puanlama yöntemleri ile puanlanmasının testin psikometrik özellikleri açısından incelenmesi. *Eğitim Bilimleri ve Uygulama,* 5(9), 35–52.

Guilford, J.P. (1941). A simple scoring weight for test items and its reliability. *Psychometrika*, 6(6), 367-374.

Gulliksen, H. (1967). *Theory of mental tests*. New York: John-Wiley & Sons Inc.

Guttman, L. (1941). An outline of the statistical theory of prediction (In P. Horst, Ed.). Prediction of personal adjustment. *Social Science Research Bulletin*, 48, 253-364.

Haladyna, T.M. (1990). Effects of empirical option weighting on estimating domain scores and making pass/ fail decisions. *Applied Measurement in Education*, 3(3), 231-244.
Hambleton, R.K., Roberts, D.M. & Traub, R.E. (1970). A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple-choice test. *Journal of Educational Measurement*, 7, 75-82.
Hambleton, R.K. & Swaminathan, H. (1985). *Item Response theory: Principles and application.* Boston: Kluwer Academic Publishers Group.
Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of Item Response theory*. California: Sage Publications Inc.
Hutchinson, T.P. (1982). Some theories of performance in multiple-choice tests, and their implications for variants of the task. *British Journal of Mathematical and Statistical Psychology*, 35, 71-89.
Jaradat, D. & Tollefson, N. (1988). The impact of alternative scoring procedures for multiple-choice items on test reliability, validity and grading. *Educational and Psychological Measurement*, 48, 627-635.
Kansup, W. & Hakstain, A.R. (1975). A comparison of several methods of assessing partial knowledge in multiple-choice tests: Scoring procedures. *Journal of Educational Measurement*, 12, 219-230.
Lord, F. (1980). *Applications of Item Response theory to practical testing problems.* New Jersey: Lawrence Erlbaum Associates Publishers.
Lord, F. & Novick R.M. (1968). *Statistical theories of mental test scores.* New York: Addison Wesley Publishing Company.
Magnusson, D. (1966). *Test theory.* Stockholm: Addison-Wesley Publishing Company.
Masters, G.N. (1982). A rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-173.
Masters, G.N. (1988). The analysis of partial credit scoring. *Applied Measurement in Education,* 1(4), 279-297.
Nedelsky, L. (1954). Ability to avoid gross error as a measure of achievement. *Educational and Psychological Measurement*, 14, 459-472.
Odell, C.V. (1931). Further data concerning the effect of weighting exercises in new-type examinations. *Journal of Educational Psychology*, 22, 700-704.
Özdemir, D. (2002). *Çoktan seçmeli testlerin Klasik Test teorisi ve örtük özellikler teorisine göre hesaplanan psikometrik özelliklerinin iki kategorili ve ağırlıklandırılmış puanlanması yönünden karşılaştırılması*. Yayımlanmamış doktora tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.
Patnaik, D. & Traub, R.E. (1973). Differential weighting by judged degree of correctness. *Journal of Educational Measurement*, 10, 281-286.
Sabers, D.L. & White, G.W. (1969). The effect of differential weighting of individual Item Responses on the predictive validity and reliability of an aptitude test. *Journal of Educational Measurement*, 6, 93-96.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.
Saygı, B. (2004). "*1-0" ve ağırlıklı puanlama yöntemleri ile puanlanan çoktan seçmeli testlerin madde ve test özelliklerinin karşılaştırılması*. Yayımlanmamış yüksek lisans tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.
Siegel, S. (1977). *Nonparametric statistics* (Çev. Y. Topsever). A.Ü. Dil ve Tarih-Coğrafya Fakültesi Yayınları No: 274. (Eserin orijinali 1956'da yayımlandı).
Sympson, J.B. & Haladyna, T.M. (1988). An evaluation of polyweighting in domain-referenced testing. *ERIC Digest,* Washington DC: ERIC Clearinghouse on Assessment and Evaluation. (ERIC Document Reproduction Service No: ED294 911).
Thissen, D.M. (1976). Information in wrong responses to the raven progressive matrices. *Journal of Educational Measu*rement, 14, 201-214.
Thissen, D.M. (1991). *Multilog user's guide- multiple, categorical item analysis and test scoring using Item Response theory*. Chicago: Scientific Software, Inc.
Wang, M.D. & Stanley, J.C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, 40 (5), 663-705.
Waters, B.K. (1976). The measurement of partial knowledge: A comparison between two empirical option-weighting methods and rights-only scoring. *The Journal of Educational Research*, 69(7), 256-260.
Wright, B.D. (1999). Model selection: Rating scale or partial credit?. *Rasch Measurement Transactions*, 12(3), 641-642.