Elementary Education Online, 11(1), 251-263, 2012.
İlköğretim Online, 11(1), 251-263, 2012. [Online]: http://ilkogretim-online.org.tr

Comparing the Test Information Obtained through Multiple-Choice, Open-Ended and Mixed Item Tests Based on Item Response Theory

Selda Gültekin* and Nükhet Çıkrıkçı Demirtaşlı**

ABSTRACT: The purpose of this study is to find out whether there is a difference between item-test functions and the level of their relative efficiency which are estimated according to the Item Response Theory among the tests which measure similar cognitive skills related to success in mathematics, and in which multiple-choice and constructed response items are employed together with varying percentages. For that purpose, item and test information functions were estimated and relative efficiency indices were calculated through responses given to four mixed item tests, among the items within TIMSS 2007 Maths test booklet #2, the test length (number of items: 15 and 25) and open-ended item percentages (20% and 40%) of which were different. Parameter estimations were carried out via BILOG-MG software. Research data were obtained from 320 eight-grade students who participated in TIMMS 2007 maths application from Turkey and who were given the booklet #2, via the official webpage of OECD (http://timss.bc.edu/). At the end of the research, it is determined that the test composed of constructed response items yields more information in terms of students' level of competence in mathematics than the one composed of multiple-choice items. Relative efficiency values indicate that the test composed of open-ended items is the most efficient of all tests.
Key Words: Mixed Item Tests, Item Response Theory, Test Information Function, Relative Efficiency

SUMMARY

Purpose and Significance: Problems arising from the popular usage of multiple-choice items in large-scale tests (success by chance, inability to check higher order cognitive skills, etc) have resulted in a tendency to use different item types along with multiple-choice items. Recently, it is observed that open-ended and other item types (true/false and restricted open-ended items and both) are being used along with multiple-choice items in large-scale testing. Despite the variety in the item types that are globally used in large-scale tests, they are not preferred in national exams in Turkey. This research is attempting to answer whether combined usage of multiple-choice and constructed response item types results in differences in the level of information obtained from the test. Findings of this research are expected to shed light on developments related to the application of nation-wide exams. Popularization of mixed item test applications is considered to be of importance in terms of adapting modern education and measurement/evaluation systems.

Methods: Data of the research were obtained from 320 eight-grade students who participated in Trends in International Science and Mathematics Study (TIMSS) implemented in 2007 and who were given booklet #2 of the maths test. In order to be compared within the scope of the research, six tests – four mixed item tests, an open-ended and a multiple-choice item test – were composed of the items from TIMSS 2007 Maths Test booklet #2. In mixed item tests, the test length was limited to 15 and 25 items, and the open-ended item percentage to 20% and 40%. The item and test information function pertaining to the question "at what level are the item and test information function and relative efficiency indices of the research" were obtained through expected a posteriori method via BILOG MG software.

Results: At the end of the research, item information functions which indicated the level of information obtained from test items proved that the amount of information obtained from open-ended test items was the highest. It was found out that the average information obtained from the open-ended items in different tests was more than the amount obtained from multiple-choice items. The relative efficiency values obtained through the ratio of test information functions showed that although test with open-ended items included fewer number of items, it had more efficiency of measurement than longer tests. The short mixed item test with a high percentage of open-ended items was found to yield more information at low order skills whereas the open-ended item test did so at higher order skills levels.

Discussion and Conclusions: Based on research findings, and taking the possible difficulties in scoring due to the composition of the test merely of open-ended items into account for the tests to be used in large-scale test applications, the usage of long mixed item tests with an open-ended item percentage closer to 50% is considered to be beneficial in terms of the amount of information obtained from the test. According to this, the usage of two formats with approximately equal percentages would prevent the inclusion of possible guessing errors due to the implementing of multiple-choice items and would increase the level of measured cognitive skills provided that they are qualified questions. In this case, decisions on selection or competence to be taken based on the results of these tests would prove more valid and reliable for monitoring of learning, selection and placement of students/persons to educational programs/jobs.

* Exp. Selda Gültekin, TED Ankara College Foundation Private Highschool, e-mail: kibarselda@gmail.com
** Assoc. Prof. Dr. Nükhet Çıkrıkçı Demirtaşlı, Ankara University Faculty of Educational Sciences Department of Measurement and Evaluation, e-posta: rnukhet@yahoo.com

Çoktan Seçmeli, Açık Uçlu ve Karma Testlerden Sağlanan Bilginin Madde Tepki Kuramına Dayalı Olarak Karşılaştırılması

Selda Gültekin[**] ve Nükhet Çıkrıkçı Demirtaşlı[***]

ÖZ. Bu çalışmanın amacı, Matematik başarısına ilişkin olarak benzer bilişsel becerileri ölçen çoktan seçmeli ve yanıtı sınırlı açık uçlu maddelerin tek başına ve birlikte kullanıldığı, ve bu iki madde tipinin test içinde yer alma yüzdelerinin farklı olduğu testlerden Madde Tepki Kuramına göre kestirilen madde ve test bilgi fonksiyonlarının (item and test information function), göreli etkililik (relative efficiency) düzeylerinin farklı olup olmadığını araştırmaktır. Bu amaç doğrultusunda, TIMSS 2007 matematik testi iki numaralı kitapçıkta yer alan maddelerden test uzunluğu (madde sayısı: 15 ve 25) ve açık uçlu madde yüzdesi (%20 ve %40) farklı dört karma testte verilen yanıtlardan madde ve test bilgi fonksiyonları kestirilmiş, göreli etkinlik indeksleri hesaplanmıştır. Parametre kestirimleri, BILOG-MG programı kullanılarak yapılmıştır. Araştırma verileri, TIMSS 2007 matematik uygulamasına Türkiye'den katılan ve matematik testinin 2. kitapçığını alan 320 8.sınıf öğrencisine ait olan ve OECD'nin resmi web sayfasından (http://timss.bc.edu/) elde edilmiştir. Araştırma sonucunda, yanıtı sınırlı açık uçlu maddelerden oluşan testin öğrencilerin matematik başarısı hakkında çoktan seçmeli maddelerden oluşan teste göre daha fazla bilgi verdiği sonucuna ulaşılmıştır. Göreli etkililik değerleri, açık uçlu maddelerden oluşan testin diğer testlerden etkili olduğunu ortaya koymuştur. Bu doğrultuda, geniş ölçekli test uygulamalarında, çoktan seçmeli maddelerin yanı sıra, yanıtı sınırlı açık uçlu maddelere de yer verilmesi, hem çoktan seçmeli maddelerden gelen şans başarısı kaynaklı hataların puanlara karışmasını önleyebilecek, nitelikli sorular olmak koşuluyla ölçülen zihinsel becerilerin düzeyi yükselebilecek ve bu testlerin sonuçlarına göre verilen seçme veya yeterlik kararlarının daha geçerli ve güvenilir olması sağlanabilecektir.

Anahtar sözcükler: Karma Testler, Madde Tepki Kuramı, Test Bilgi Fonksiyonu, Göreli Etkililik

GİRİŞ

Öğrenci merkezli öğrenme yaklaşımlarının benimsendiği değişen öğretim programlarında, sınıf içi değerlendirme etkinliklerinin doğası değişmiş; öğretmenin hem öğretim hem de değerlendirme etkinliklerini çeşitlendirmesi ve birleştirmesini gerektirmiştir. Tek bir soru formatı ya da değerlendirme şeklini kullanmak öğrencilerde tek yönlü bir çalışma alışkanlığı geliştirerek; öğrenme çıktıları açısından yine tek yönlü özellikler ağırlık kazanmaktadır (Berberoğlu, 2006).

Eğitimde öğrenmeleri izleme, teşhis, seçme, yerleştirme gibi çeşitli amaçlarla kullanılan ölçme araçlarında yer alan farklı madde tiplerinin güçlü ve zayıf yönleri bulunmaktadır. Çoktan seçmeli maddelerden oluşan testler, daha kısa sürede daha fazla sayıda beceri ölçülebilmesi, puanlanma kolaylığı ve objektifliği gibi üstünlüklere sahiptir (Haladyna, 1997). Çoktan seçmeli madde türüyle, bilişsel alanın bazı düzeyindeki davranışları (bilgi, kavrama, uygulama, analiz) ölçmek mümkündür (Tekin, 1991). Bununla birlikte, son yıllarda eğitim-öğretimde, okuduğunu anlama, kritik etme, yorumlama, bilgiyi toplayıp analiz edebilme, bir sonuca ulaşma, grafik ya da tablo halinde verilen bilgiden sonuç çıkarma, uzaysal muhakeme, gözlem yapma, gözlemlerden sonuca ulaşma, günlük hayatta sıkça karşılaşılan problemleri çözebilme, araştırma yapma gibi bireyleri sosyal yaşama daha çok hazırlayan becerilerin ağırlık kazandığı görülmektedir (Berberoğlu, 2006). Bu türden üst düzey düşünme becerilerinin sadece çoktan seçmeli maddelerden oluşan testlerle yoklanması zordur.

Çoktan seçmeli maddelerin önemli diğer bir sınırlılığı da bireylerin tahminle puan kazanma olanağının bulunmasıdır. Çoktan seçmeli madde formatının yapısı gereği seçenekler içermesi, o maddeyle ölçülen özelliğe sahip olmayan veya kısmen sahip olan yanıtlayıcıyı, şansını kullanarak doğru yanıtı bulmaya yöneltebilir. Şans başarısı olarak tanımlanan bu durum testin geçerlik ve güvenirliğini olumsuz etkilemektedir (Tekin, 1991).

[**] Ölçme Değerlendirme Uzmanı, TED Ankara Koleji Vakfı Özel Lisesi, e-posta: kibarselda@gmail.com
[***] Doç. Dr., Ankara Üniversitesi Eğitim Bilimleri Fakültesi, Ölçme ve Değerlendirme Anabilim Dalı, e-posta: rnukhet@yahoo.com

Çoktan seçmeli maddelerin geniş ölçekli testlerde yaygın olarak kullanılması sonucunda ortaya çıkan bu tartışmalar çoktan seçmeli maddelerin yanı sıra farklı madde türlerini de kullanma eğilimini ortaya çıkarmıştır. Farklı madde tiplerinin birbirlerine olan üstünlüğünü avantaja dönüştürmek için ölçme araçlarında farklı madde tiplerinin bir arada kullanılması yaygınlık kazanmaktadır. Dünyada "National Assessment of Educational Progress (NAEP), Massachusetts Comprehensive Assessment System (MCAS), Test of English as a Foreign Language (TOEFL), Programme for International Student Assessment (PISA), Trends In International Mathematics and Science Study (TIMSS), Advanced Placement Test (AP), CITO Türkiye Öğrenci izleme Sistemi-ÖİS" gibi farklı madde türlerinin bir arada kullanıldığı, akademik başarı ve yeterlik belirlemeye yönelik ölçme uygulamaları bulunmaktadır. Uygulandıkları ülkelerde yüzbinlerce bireyin katılımıyla gerçekleşen bu tür uygulamalar, karma test maddelerinin puanlanmasına yönelik geliştirilmiş özel yazılımlarla (IntelliMetric, E-rater, Intelligent Essay Assessor, Project Essay Grade vb.) puanlanmaktadır. (Wainer&Thissen, 1993; Bastari, 2000; Shermis&Burstein, 2003; MEB, 2007; Demirtaşlı, 2010).

Dünyada geniş ölçekli testlerde kullanılan madde türlerindeki çeşitlilik ulusal sınavlarda tercih edilmemektedir. Bu araştırmayla çoktan seçmeli ve yanıtı sınırlı açık uçlu madde tiplerinin birlikte kullanılmasının testten elde edilen bilgi düzeyinde farklılıklara yol açıp açmadığı ortaya konmaya çalışılmıştır. Araştırmanın bulgularının ulusal düzeydeki merkezi sistem sınavlarının uygulanmasıyla ilgili gelişmelere ışık tutacağı düşünülmektedir. Nitekim, ölçme araçlarında sadece çoktan seçmeli soru formatının kullanılmasının eğitim sistemi üzerinde yarattığı olumsuz etkiler sadece eğitimciler arasında değil karar vericiler düzeyinde de tartışılmaya başlanmıştır. Berberoğlu (2009), Öğrenci İzleme Sistemi (ÖİS) kapsamında öğrencilerin akademik başarısını etkileyebilecek diğer duyuşsal ve öğretimsel özelliklerin yoklandığı Öğrenci Sosyal Gelişim Programı çerçevesinde uygulanan anketlerle, öğrencilerin sınıf düzeyi arttıkça daha çok ezberleme stratejilerini kullandıklarını ortaya konmuştur. Bu durum, öğretmenlerin sınıfta "test çözme"yi bir öğretim etkinliği gibi kullanma ve bu türden testlerin ev ödevi gibi verilmesi davranışının bir sonucudur. Öğrenciler, konuları kavramak yerine, soruların çözüm algoritmalarını ezberlemektedirler. Soru üzerinden eğitimin yanlış bir süreç olduğu tartışmaları son dönemde oldukça artmıştır. Karma test uygulamalarının yaygınlaşmasının, modern çağa uygun eğitim-öğretim ve ölçme-değerlendirme sistemlerine geçiş açısından önemli olacağı düşünülmektedir.

Bu uygulamalarda kullanılmaya başlanan açık uçlu maddeler, kendi içerisinde yanıtı sınırlandırılmış ve yanıtı serbest bırakılmış sorular olmak üzere iki grupta ele alınmaktadır. Yanıtı sınırlandırılmış açık uçlu maddede öğrenciden yanıtın niteliğine, uzunluğuna ya da organizasyonuna yönelik sınırlamalar yapması istenir. Yanıtı sınırlandırılmış açık uçlu maddeler, hem daha çok soru sormaya olanak sağlaması hem de puanlamanın kolay olması nedeniyle daha çok tercih edilmesi gereken bir formattır (Kubizh ve Borich, 2003). Matematikte problem çözme becerisi yoklanırken kısa ama çok sayıda problem sorulması şans faktörünü büyük ölçüde azaltarak güvenirliğin artmasını sağlar. "açık uçlu" olarak tanımlanan bu tip sorularda yazma ve ifade becerileri puanlamada etkili olmayacak ve yine çoktan seçmeli maddeler gibi 0 ve 1 şeklinde puanlanabilecektir (Umay, 1997).

Yukarıda sayılan bu ölçme uygulamalarından biri de Uluslararası Eğitim Başarısı Değerlendirme Kuruluşu (International Association for the Evaluation of Educational Achievement – IEA) tarafından, katılımcı OECD ülkelerinin dört yılda bir 4. ve 8. sınıf düzeylerinde uygulanan, öğrencilerin matematik ve fen başarılarını ölçmeyi amaçlayan TIMSS (Trends in International Mathematics and Science Study)'tir. TIMSS uygulamalarının dördüncü ve sonuncusu olan TIMSS 2007, katılımcı ülkelerin öğrencilerin matematik ve fen bilimleri alanındaki başarı durumları değerlendirilip; öğretim programları, öğretmen ve okulların özellikleri, öğrenci özellikleri ve eğitim sistemi hakkında da bilgi sahibi olunmaktadır. 49 ülkenin katıldığı uygulamaya Türkiye sadece 8. sınıf düzeyinde 146 okulda toplam 4498 öğrenciyle katılmıştır. 4. sınıf düzeyinde katılım olmamıştır (IEA, 2008-1; MEB, 2011).

Test geliştirme ve diğer ölçekleme uygulamaları belli kuramsal temellere dayalı olarak yürütülür. Psikometride ölçme araçlarının geliştirilmesi, puanlanması ve psikometrik nitelikleriyle ilgili sorunların ele alınmasında yaygın olarak iki test kuramından yararlanılır. Biri, klasik test kuramı (KTK) diğeri de kullanımı giderek yaygınlaşan Madde Tepki Kuramı (MTK)'dır. MTK, olasılıklı ölçme modelleri ile ölçme uygulamalarına bazı avantajlar sağlamıştır. Bunlar arasında; farklı amaçlar için test geliştirme, puanların eşitlenmesi, madde yanlılığının belirlenmesi ve bilgisayarlı bireye

uyarlanmış testlerde yetenek puanlarının kestirimi gibi çeşitli ölçme problemlerinin çözümü için kullanışlı bir çerçeve sağlar (Hambleton, Swaminathan ve Rogers, 1991, Baker&Kim, 2004, Zhao, 2008). Madde Tepki Kuramı, kişinin ölçülen özellikteki (yeterlik) düzeyi ile verdiği yanıtlar arasında bir ilişki olduğunu kabul eder, bu ilişkiyi matematiksel bir fonksiyon ile açıklayan olasılıklı bir model önerir. (Embretson ve Reise, 2000). Madde Tepki Kuramı iki temel kabule dayanır: (a) doğrudan gözlenemeyen örtük özellik ya da yetenek/yeterlik olarak adlandırılan psikolojik yapı, bireylerin test maddelerindeki gözlenen performansından kestirilebilir, (b) bireylerin maddelerdeki performansı ile madde performansından sorumlu olan özellik arasındaki ilişki, madde karakteristik fornksiyonu/eğrisi olarak adlandırılan doğrusal olmayan bir fonksiyonla açıklanabilir (Hambleton, Swaminathan ve Rogers, 1991). Kuram, belli varsayımlar (tekboyutluluk, yerel bağımsızlık, model-veri uyumu) altında maddelerin özelliklerinden bağımsız yetenek parametreleri ve yanıtlayıcı örnekleminden bağımsız madde parametreleri kestirebileceğini iddia eder. MTK bu özelliği farklı amaçlar test geliştirmeyi, paralel ve bireye uyarlamalı test geliştirmeyi, test eşitleme çalışmalarını daha mümkün kılmaktadır.

Bireylerin geleceğini etkileyen SBS, ÖSS gibi geniş ölçekli testler sadece çoktan seçmeli maddelerden oluşmaktadır. Bu testler de yukarıda özetlendiği üzere çoktan seçmeli testlerin sınırlılıklarını taşımaktadır. Bu noktadan hareketle benzer yapıları ölçen çoktan seçmeli ve yanıtı sınırlı açık uçlu madde formatlarının bir arada kullanıldığı durumlarda testlerin psikometrik niteliklerinin nasıl farklılaştığının ortaya konması önem kazanmıştır. Bu kapsamda ölçme uygulamalarında KTK'nın sınırlılıklarını gideren MTK'ye dayalı ölçme model ve yöntemlerinin kullanımı yaygınlaşmıştır

Madde Tepki Kuramında, test maddelerini seçmede ve testleri karşılaştırmada kullanılan en önemli parametre, "madde ve test bilgisi"dir. Madde bilgisi, maddenin ölçtüğü özellik hakkında ne denli güvenilir bilgi verdiği gösterir. Temelde madde bilgisi parametresi, maddenin güçlük ve ayırıcılık parametreleri ile ilişkilidir. Eşitlik 1 ve 2 madde bilgisini veren fonksiyonu ve diğer madde parametreleri ile ilişkisini vermektedir.

$$I_i(\theta) = \frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)\, Q_i(\theta)} \qquad (1)$$

Eşitlik (1)'de; $I_i(\theta)$: i. maddenin madde bilgi fonksiyonunu, $P_i(\theta)$: θ'ya bağlı madde tepki fonksiyonunu, $Q_i(\theta)$: $1–P_i(\theta)$'yı ve $P_i'(\theta)$: $P_i(\theta)$'nın birinci türevini göstermektedir. İki kategorili puanlanan maddelerde üç parametreli lojistik model için $P_i(\theta)$'nın açılımı bu eşitlikte yerine konduğunda aşağıdaki eşitlik elde edilmektedir:

$$I(\theta) = \frac{2{,}89\, a_i^2 (1 - c_i)}{\left[c_i + \exp(1{,}7\, a_i\, (\theta - b_i)\right]\left[1 + \exp(-1{,}7\, a_i\, (\theta - b_i)\right]^2} \qquad (2)$$

Bu eşitlikte, madde bilgi fonksiyonunun madde parametreleriyle ilişkisi açıkça görülmektedir. Bir madde, madde güçlük parametresi (b) hitap ettiği yetenek düzeyi θ'ya yaklaştıkça, madde ayırıcılığı (a) arttıkça ve şans parametresi (c) sıfıra yaklaştıkça daha fazla bilgi vermektedir.

θ yetenek düzeyinde bir testin sağladığı bilgi, o yetenek düzeyinde madde bilgi fonksiyonlarının toplamından elde edilir.

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) \qquad (3)$$

Bu durum, her bir test maddesinin katkısının, testteki diğer maddeler bilinmediğinde de tanımlanabildiğini ortaya koymaktadır. Test bilgi fonksiyonu, Klasik Test Kuramındaki güvenirlik katsayısından farklı olarak, testin uygulandığı örneklemden tamamen bağımsız olduğunu iddia etmektedir.

Bir test bilgi fonksiyonu, yetenek düzeyi ölçeğinde bazı noktalarda en yüksek değerini alırken; tüm yetenek düzeylerinde eşit olmayan değerler alabilir. Böyle bir testin, yetenek düzeyleri test bilgi fonksiyonunun en yüksek değere ulaştığı noktanın yakınlarındaki bireylerin yetenek kestiriminde iyi olduğu söylenmektedir. Bazı testlerde ise test bilgi fonksiyonu yetenek düzeyinin bazı noktalarında daha basık bir dağılım gösterebilir. Test bilgi fonksiyonu bu testlerin, o yetenek aralığındaki bireylere hitap ettiğini gösterir. Test bilgi fonksiyonu yorumlanırken; yetenek düzeyi ve test bilgi fonksiyonu arasındaki ilişki göz önünde tutulmalıdır. (Hambleton, Swaminathan ve Rogers, 1991, Embretson ve Reise, 2000, Baker, 2001).

Test bilgi fonksiyonu, test geliştirme ve madde seçme açısından önem taşımaktadır. Ancak; test ve madde bilgi fonksiyonu, kesin olarak yorumlanamadığında, testlerin göreli olarak karşılaştırılması mümkündür. Genel olarak, $I_A(\theta)$ ve $I_B(\theta)$ iki teste ilişkin aynı θ düzeyinde test bilgi fonksiyonlarını gösterdiğinde; göreli etkililik (relative efficiency):

$$RE(\theta) = I_A(\theta) / I_B(\theta)$$

formülüyle hesaplanmaktadır. Örneğin; A ve B gibi aynı özelliği ölçmek üzere geliştirilmiş iki testin test bilgisi sırasıyla, $I_A(\theta)=25$ ve $I_B(\theta)=20$ olduğunda, $RE(\theta)=1.25$ olarak bulunmaktadır. Bu durumda, "Belirli θ düzeyinde, Test A, Test B'den %25 daha uzunmuş gibi bilgi sağlamaktadır." yorumu yapılabilmektedir. Bu kavram, test modeli ve puanlama formülünün seçiminde önemli rol oynamaktadır. (Lord ve Novick, 1968, Hambleton & Swaminathan, 1985, Hambleton, Swaminathan ve Rogers, 1991).

Farklı madde tiplerinin bir arada kullanıldığı ve madde tepki kuramı kapsamında ele alan çeşitli araştırmalar vardır. Bu çalışmalarda, kullanılan farklı madde formatlarına ilişkin puanlama süreçleri ve bunların ölçme sonuçlarının geçerlik ve güvenirliğini artırmaya etkisi olduğu üzerinde durulmuştur (Wainer&Thissen, 1993; Bastari, 2000; Schaeffer ve diğerleri, 2002; Baker&Kim, 2004; Shin, 2007). Madde Tepki Kuramına dayalı olarak yapılan bu çalışmalarda, açık uçlu maddelerden sağlanan bilginin çoktan seçmeli maddelere göre daha yüksek olduğu; açık uçlu maddeler şans başarısı içermediğinden düşük yetenek düzeyindeki bireylerde de ölçülen özelliğe ilişkin yüksek bilgi sağlanabildiği belirlenmiştir. Madde sayısı ve açık uçlu madde oranının artmasının testten sağlanan bilgi, güvenirlik kestirimi ve ölçme duyarlığını artırdığı bulgusuna ulaşılmıştır. (Lukhele ve diğerleri, 1994; Ercikan ve diğerleri, 1998; Kinsey 2003; Uyeno, 2004)

Yukarıda değinilen ilgili alan yazında farklı madde formatlarının birbirine üstünlüğünün araştırıldığı ve son yıllarda farklı madde formatlarının bir arada kullanıldığı karma test uygulamalarında testlerin psikometrik niteliklerinin farklılaşıp farklılaşmadığı araştırmalara konu olmaktadır. Farklı madde formatlarının birbirlerine olan üstünlüğüne dair pek çok araştırma olmasına karşın; özellikle son yıllarda dünyada sıklıkla geniş ölçekli test uygulamalarında yer bulmaya başlayan karma testlerin sadece tek madde formatından oluşan testlere karşı üstün olup olmadığına ilişkin yeterli araştırma olmaması bu araştırmada karma testlerin madde ve test bilgi fonksiyonlarının, çoktan seçmeli veya açık uçlu maddelerden oluşan testlere göre farklılık gösterip göstermediğinin araştırılmasına ihtiyaç duyulmuştur.

Bu çalışmada, Matematik başarısına ilişkin olarak benzer yapıları ölçen iki madde tipinin (çoktan seçmeli ve yanıtı sınırlı açık uçlu) ayrı (tek başına) ve birlikte kullanıldıkları koşullar ile iki madde tipinin test içinde yer alma yüzdelerinin test formlarında farklı olduğu durumlarda MTK'ye göre kestirilen madde ve test bilgi fonksiyonlarının (test information function) ve göreli etkinlik (relative efficiency) düzeylerinin farklı olup olmadığını araştırmak amaçlanmıştır. Bu amaç çerçevesinde; "Madde sayısının 25 ve 15; açık uçlu madde yüzdesinin %40 ve %20 olduğu dört karma testten (KT1:25,%40; KT2:25,%20; KT3:15,%40; KT4:15,%20), açık uçlu (AUT) ve çoktan seçmeli (ÇST) testlerden kestirilen madde ve test bilgi fonksiyonları ile göreli etkinlik (relative efficiency) indeksleri hangi düzeydedir?" sorusuna yanıt aranmıştır.

YÖNTEM

Araştırma Modeli

Bu araştırmada, çoktan seçmeli ve açık uçlu madde formatlarının Madde Tepki Kuramına dayalı olarak birlikte ölçeklenmesinin madde ve test bilgi fonksiyonlarında fark yaratıp yaratmadığı araştırılmıştır. Bu yönüyle çalışma, kuramsal ve temel araştırma niteliğindedir.

Araştırma Grubu

Araştırmanın verileri, 2007 yılında uygulanan Uluslararası Fen ve Matematik Eğilimleri Araştırması (Trends in International Mathematics and Science Study – TIMSS 2007) Türkiye'den katılan ve matematik testinin 2. kitapçığını alan 320, 8.sınıf öğrencisinden elde edilmiştir.

TIMSS 2007 uygulamasında yer alan ülkelerdeki katılımcılar, tek biçimli (uniform) örnekleme yaklaşımıyla örnekleme minimum sapmayla dahil olmuştur. Bu yöntem bir kalite standardı sağlayarak, araştırma sonuçlarındaki ülkeler arası farklılıkların örnekleme yönteminden kaynaklanması ihtimalini ortadan kaldırmaktadır. Katılımcı her ülkenin TIMSS 2007 Ulusal Araştırma Koordinatörü (National Research Coordinator), TIMSS & PIRLS Uluslararası Çalışma Merkezi (International Study Center) tarafından onaylanan örnekleme prosedürünün her adımını uygulama ve rapor etme sorumluluğunu üstlenmiştir. Ulusal Araştırma Koordinatörleri *Windows Within-school Sampling Software* yazılımını kullanarak örnekleme dahil olacak öğrencileri belirlemektedir.

TIMSS 2007 uygulaması, dünya çapında 49 ülke ve 7 kıyaslama (benchmarking) katılımcısı ile gerçekleşmiştir. Coğrafi bölge ve okul türü örnekleme tabakası olarak alınmıştır. Türkiye uygulamasında örneklem yedi coğrafi bölgede ve iki okul türüne (devlet ve özel) göre tabakalı olarak her bölgeden oranı ölçüsünde okul; her okuldan da seçkisiz (random) bir sınıf seçilecek şekilde belirlenmiştir. Bu doğrultuda Türkiye'de uygulamaya 146 okuldan bu yolla seçilmiş 4498, 8. sınıf öğrencisi katılmıştır (IEA 2008-1).

Araştırma kapsamında yayımlanan maddelerin bulunduğu 2. kitapçık ele alındığından dolayı, araştırmanın amacına yönelik yapılan analizlerde 2. kitapçığı yanıtlayan 320 öğrencinin yanıtlarından elde edilen veriler kullanılmıştır.

Veriler ve ölçme aracı

Araştırma kapsamında ele alınan veriler, TIMSS & PIRLS Uluslararası Çalışma Merkezine ait internet sitesindeki uluslararası veri tabanından SPSS dosyaları şeklinde alınmıştır (http://timss.bc.edu/TIMSS2007/idb_ug.html). 8. sınıf verilerinin bulunduğu dosyadan Türkiye katılımcılarının matematik testine verdikleri yanıtlar alınarak araştırmada kullanılan veriler elde edilmiştir.

TIMSS matematik değerlendirmeleri iki boyutta tasarlanmıştır: (i) öğrencilerin öğrenmesi beklenen konu ya da kapsam ve (ii) öğrencilerin göstermesi beklenen bilişsel beceriler. Bu doğrultuda matematik testi, 8. sınıf düzeyinde Sayılar, Cebir, Geometri, Veri ve Olasılık konuları kapsamında (content domain); *bilgi*, *uygulama ve akıl yürütme* bilişsel alanlarını (cognitive domain) içermektedir. *Bilgi*, öğrencilerin matematik olguları, kavramları, araçları ve yöntemlerine dayalı bilgisini tanımlamaktadır. *Uygulama*, öğrencinin problem durumundaki kavramsal algılama ve bilgiyi uygulamadaki yeteneğine odaklanır. *Akıl yürütme* ise, benzer olmayan durumları içeren ve çok aşamalı problemler gibi rutin problem çözmenin ötesine geçmek olarak tanımlanmaktadır. (IEA, 2005 ve 2008-2, Gonzales, 2008).

MTK'ye dayalı olarak geliştirilen ve puanlanan TIMSS 2007 8. sınıflar matematik testinde toplamda 214 maddeden oluşan bir soru havuzu bulunmakla birlikte, uygulamada her öğrenci her birinin güçlük düzeyi denk 28-33 sayıda madde bulunan 14 farklı kitapçıktan birini yanıtlamıştır. Her iki kitapçıkta sayısı 12-19 arasında değişen ortak madde (anchor item) bulunmaktadır. Ortak maddeler 14 farklı formun Madde Tepki Kuramına dayalı olarak diğer kitapçıkları almamış öğrencileri o maddeleri de yanıtlamaları durumunda yeterlik kestirimi yapmak amacıyla kullanılmıştır. Madde Tepki Kuramına dayalı olarak eşitleme (equating) yoluyla test kitapçıklarının güçlük açısından denkliği sağlanmış ve bu yolla öğrencinin almadığı testteki yeterlik düzeyinin kestirilebilmesi mümkün olmuştur. Veri dosyasındaki (bsaturm4.sav) tanımlamalardan yararlanılarak TIMSS 2007 8.

sınıf matematik uygulamasındaki 2. kitapçıkta yer alan maddelerin madde tipi ve bilişsel alanlara göre dağılımları belirlenerek Tablo 1'de verilmiştir.

Tablo 1. *TIMSS 2007 8. Sınıflar Matematik Testi 2. Kitapçıkta Yer Alan Maddelerin Madde Tipi ve Bilişsel Alanlara Göre Dağılımları*

Madde Tipi / Bilişsel Alan	Açık Uçlu		Çoktan Seçmeli		Toplam	
	Madde Sayısı	Yüzde	Madde Sayısı	Yüzde	Madde Sayısı	Yüzde
Bilgi	3	27,27	9	45,00	12	38,71
Uygulama	5	45,45	10	50,00	15	48,39
Akıl yürütme	3	27,27	1	5,00	4	12,90
Toplam	11	100,00	20	100,00	31	100,00

TIMSS 2007 8. Sınıflar Matematik Testi 2. kitapçığında yer alan çoktan seçmeli maddeler iki kategorili (0-1) olarak puanlanmıştır. Öte yandan, açık uçlu maddelerden sadece biri (M042220) kısmi puanlamış (0-1-2), diğer 10 madde iki kategorili (0-1) olarak puanlanmıştır. İki uçlu puanlanan maddeleri araştırma kapsamında ele alma sınırlılığından dolayı kısmi puanlanan madde (M042220) iki kategorili puanlanarak teste dahil edilmiştir.

Araştırma kapsamında, TIMSS 2007 Matematik Testi 2. kitapçığındaki maddelerden oluşturulan karma testlerde test uzunluğu 15 ve 25 madde, açık uçlu madde yüzdesi %20 ve %40 ile sınırlı tutulmuştur. İlgili alanyazında, çok farklı sayıda test uzunluğu (10-20-30, 15-25-50, vb.) ve açık uçlu madde yüzdesi (%15-%25, %25-%50, vb.) ile araştırmalar yapılmıştır. Bu araştırmada karma testleri oluşturmada yararlanılan test uzunlukları ve açık uçlu madde yüzdeleri TIMSS 2007 Matematik Testi 2. kitapçığından, bilişsel düzeylere göre dağılımı denk olacak şekilde dört karma test elde etmeye uygun olacak şekilde belirlenmiştir. Bu kitapçıkta yer alan 31 madde arasından, farklı madde sayısı ve açık uçlu madde yüzdelerinde oluşturulan dört karma teste, açık uçlu ve çoktan seçmeli testlere ilişkin bilgiler Tablo 2'de verilmiştir.

Tablo 2. *Karma Testlerdeki Madde Sayılarının Bilişsel Alan ve Madde Tipine Göre Dağılımları*

Karma Test 1 (KT1) (k=25 ; AU Yüzdesi: %40)				
Bilişsel Alan	Açık Uçlu	Çoktan Seçmeli	Toplam	Bilişsel Alandaki Yüzdesi
Bilgi	3	8	11	**44**
Uygulama	5	6	11	**44**
Akıl yürütme	2	1	3	**12**
Toplam madde sayısı	10	15	25	**100**
Karma Test 2 (KT2) (k=25 ; AU Yüzdesi: %20)				
Bilişsel Alan	Açık Uçlu	Çoktan Seçmeli	Toplam	Bilişsel Alandaki Yüzdesi
Bilgi	2	9	11	**44**
Uygulama	1	10	11	**44**
Akıl yürütme	2	1	3	**12**
Toplam madde sayısı	5	20	25	**100**
Karma Test 3 (KT3) (k=15 ; AU Yüzdesi: %40)				
Bilişsel Alan	Açık Uçlu	Çoktan Seçmeli	Toplam	Bilişsel Alandaki Yüzdesi
Bilgi	2	4	6	**40**
Uygulama	3	4	7	**47**
Akıl yürütme	1	1	2	**13**
Toplam madde sayısı	6	9	15	**100**
Karma Test 4 (KT4) (k=15 ; AU Yüzdesi: %20)				
Bilişsel Alan	Açık Uçlu	Çoktan Seçmeli	Toplam	Bilişsel Alandaki Yüzdesi
Bilgi	1	5	6	**40**
Uygulama	1	6	7	**47**
Akıl yürütme	1	1	2	**13**
Toplam madde sayısı	3	12	15	**100**

k: madde sayısı , AU: açık uçlu madde

Verilerin Analizi

Bu bölümde, MTK varsayımlarına ilişkin olarak yapılan analizler ve araştırma sorularına ilişkin analizler olmak üzere iki aşamada açıklanmıştır. Öncelikle MTK varsayımlarından tek boyutluluk ve veri-model uyumu test edilmiştir. Tek boyutluluğu sınamada faktör analizi kullanılmıştır. Yapılan analizle araştırma kapsamında oluşturulan altı testin de tek boyutlu olduğu ortaya konmuştur. Araştırmada tek boyutluluğun sağlanması ilgili kanıt, alanyazında sıkça rastlandığı gibi yerel bağımsızlığın da bir kanıtı olarak değerlendirilmiştir

MTK varsayımlarından model-veri uyumunun üzerinde çalışılacak olan altı test için sağlanıp sağlanmadığı -2loglikelihood istatistiği ölçüt alınarak sınanmıştır. AUT ve KT3'ün 2 PL; ÇST, KT1, KT2 ve KT4'ün 3 PL modele daha iyi uyum sağladığı belirlenmiştir. Model veri uyumu testinin sonuçlarına bağlı olarak KT1, KT2, KT4 ve ÇST için kestirimler 3PL modele göre; KT3 ve AUT için kestirimler 2PL modele göre yapılmıştır.

Araştırmanın "madde ve test bilgi fonksiyonları ile göreli etkinlik indeksleri hangi düzeydedir" sorusuna ilişkin madde ve test bilgi fonksiyonu BILOG MG programında beklenen a posteriori (expected a posteriori) yöntemiyle elde edilmiştir. Beklenen a posteriori (expected a posteriori) yöntemi, diğer kestirim yöntemlerinden (ençok olabilirlik, en yüksek posteriori) ayrı olarak tüm tepki örüntüleri için (tümü doğru - tümü yanlış) sonlu bir yetenek kestirimi ($-3 < \theta < +3.0$) sağlamaktadır. Ayrıca iteratif bir prosedür değildir (Embretson ve Reise, 2000.)

BULGULAR VE YORUMLAR

Test ve madde bilgi fonksiyonları ile test bilgi fonksiyonlarının oranlanmasıyla elde edilen göreli etkinlik değerleri, test geliştiricilere madde ve test seçme açısından kaynaklık etmektedir. Bu doğrultuda, dört karma test, çoktan seçmeli test ve açık uçlu testlerden hangilerinin bireylerin yeterliklerini kestirimede daha fazla bilgi sağladığını ve daha etkin olarak kullanılabileceğini ortaya koymak amacıyla kestirilen madde ve test bilgi fonksiyonları ile göreli etkinlik indekslerine ilişkin bulgular bu bölümde tartışılmıştır. Dört karma test, çoktan seçmeli ve açık uçlu testlerden kestirilen ortalama madde bilgi fonksiyonları Tablo 3'te gösterilmiştir.

Tablo 3'te testlerin ortalama bilgi düzeyleri incelendiğinde, genel olarak birbirlerine yakın oldukları görülmektedir. Testlere ilişkin ortalama bilgi düzeyleri genel olarak yakın olmakla beraber; en fazla bilginin AUT'tan sağlandığı görülmektedir. Sağladığı bilgi bakımından AUT'u KT1 ve KT2 izlemiştir.

Bu noktada, maddelerden sağlanan bilginin madde tiplerine göre farklılaşıp farklılaşmadığını ortaya koyma ihtiyacı doğmuştur. Bu nedenle, ortalama bilgi değerleri madde tipine göre ayrı ayrı değerlendirilmiş ve Tablo 4'te verilmiştir.

Tablo 3. *Karma Testler, Çoktan Seçmeli Test ve Açık Uçlu Teste İlişkin Madde Bilgi Fonksiyonları*

Test Adı			KT1		KT2		KT3*		KT4		ÇST		AUT*	
Madde Sayısı			*25*		*25*		*15*		*15*		*20*		*10*	
AU Madde Oranı			*%40*		*%20*		*%40*		*%20*					
Madde Kodu	Format	Düz.	Madde No	Ort. Bilgi	Madde No	Ort. Bilgi	Madde No	Ort. Bilgi	Madde No	Ort. Bilgi	Madde No	Ort. Bilgi	Madde No	Ort. Bilgi
M042003	ÇS	B	1	0.24	1	0.23	1	0.23	1	0.25	1	0.23	--	--
M042079	ÇS	B	2	0.24	2	0.27	--	--	--	--	2	0.26	--	--
M042018	AU	U	3	0.57	--	--	--	--	--	--	--	--	1	0.49
M042055	ÇS	U	4	0.23	3	0.35	2	0.09	2	0.22	3	0.21	--	--
M042039	ÇS	U	--	--	4	0.26	--	--	--	--	4	0.25	--	--
M042199	ÇS	B	5	0.50	5	0.47	3	0.38	3	0.35	5	0.46	--	--
M042301A	AU	B	6	0.25	6	0.25	4	0.31	--	--	--	--	2	0.29
M042301B	AU	AY	7	0.48	7	0.42	5	0.47	4	0.39	--	--	3	0.62
M042301C	AU	AY	8	0.80	8	0.64	--	--	--	--	--	--	4	1.08
M042265	ÇS	AY	9	0.13	9	0.13	6	0.15	5	0.12	6	0.13	--	--
M042137	ÇS	U	10	0.27	10	0.27	7	0.22	6	0.28	7	0.24	--	--
M042148	ÇS	B	11	0.20	11	0.22	8	0.24	7	0.26	8	0.23	--	--
M042254	ÇS	U	12	0.25	12	0.23	9	0.31	8	0.21	9	0.22	--	--
M042250	AU	B	13	0.51	13	0.53	10	0.52	--	--	--	--	5	0.55
M042220	AU	U	14	0.46	--	--	11	0.42	--	--	--	--	6	0.44
M022097	ÇS	B	15	0.28	14	0.28	--	--	--	--	10	0.27	--	--
M022101	ÇS	B	16	0.17	15	0.19	12	0.19	9	0.19	11	0.20	--	--
M022104	ÇS	B	17	0.33	16	0.32	--	--	--	--	12	0.33	--	--
M022105	ÇS	B	--	--	17	0.03	--	--	--	--	13	0.03	--	--
M022106	AU	U	18	0.27	--	--	13	0.28	--	--	--	--	7	0.25
M022108	ÇS	U	--	--	18	0.64	--	--	--	--	14	0.68	--	--
M022110	AU	B	19	0.22	--	--	--	--	10	0.16	--	--	8	0.20
M022181	ÇS	U	20	0.38	19	0.38	--	--	11	0.41	15	0.41	--	--
M032307	AU	U	21	0.42	--	--	--	--	--	--	--	--	9	0.45
M032523	ÇS	U	--	--	20	0.27	--	--	--	--	16	0.26	--	--
M032701	ÇS	U	22	0.31	21	0.33	14	0.26	12	0.39	17	0.37	--	--
M032704	ÇS	U	--	--	22	0.57	--	--	--	--	18	0.67	--	--
M032525	ÇS	B	23	0.49	23	0.52	--	--	13	0.34	19	0.54	--	--
M032579	ÇS	U	24	0.59	24	0.63	--	--	14	0.56	20	0.58	--	--
M032691	AU	U	25	0.40	25	0.40	15	0.46	15	0.53	--	--	10	0.37
Ortalama				0.36		0.35		0.30		0.31		0.33		0.47
Std. Sapma				0.16		0.16		0.12		0.13		0.18		0.25
En Düşük				0.13		0.03		0.09		0.12		0.03		0.20
En Yüksek				0.80		0.64		0.52		0.56		0.68		1.08
Genişlik				0.67		0.62		0.43		0.43		0.65		0.89

* KT3 ve AUT için parametreler 2 Parametreli Liojistik modele göre kestirilmiştir.
ÇS: Çoktan Seçmeli, **AU:** Açık Uçlu, **Düz.:** Maddenin ölçtüğü bilişsel düzey, **B:** Bilgi, **U:** Uygulama, **AY:** Akıl Yürütme.

Tablo 4. *Dört Karma Test ile ÇST ve AUT'ten oluşan testlerden elde edilen Test Bilgisi Ortalamaları*

		Madde Tipi	
		Açık Uçlu	**Çoktan Seçmeli**
Ortalama Bilgi	**KT1**	0.44	0.31
	KT2	0.45	0.33
	KT3	0.41	0.23
	KT4	0.36	0.30
	ÇST	--	0.33
	AUT	0.47	--

Tablo 4 incelendiğinde, farklı testlerdeki açık uçlu maddelerden sağlanan ortalama bilginin çoktan seçmeli maddelerden sağlanan bilgiden daha fazla olduğu görülmektedir.

Farklı testlerin birbirine tercih edilmesi söz konusu olduğunda kıyaslanması amacıyla kullanılan göreli etkinlik (relative efficiency) değerleri, testlere ait ortalama test bilgi fonksiyonlarının birbirlerine oranlanmasıyla hesaplanmış ve Tablo 5'te verilmiştir.

Tablo 5. *Testlere İlişkin Göreli Etkinlik (Relative Efficiency) İndeksi Değerleri*

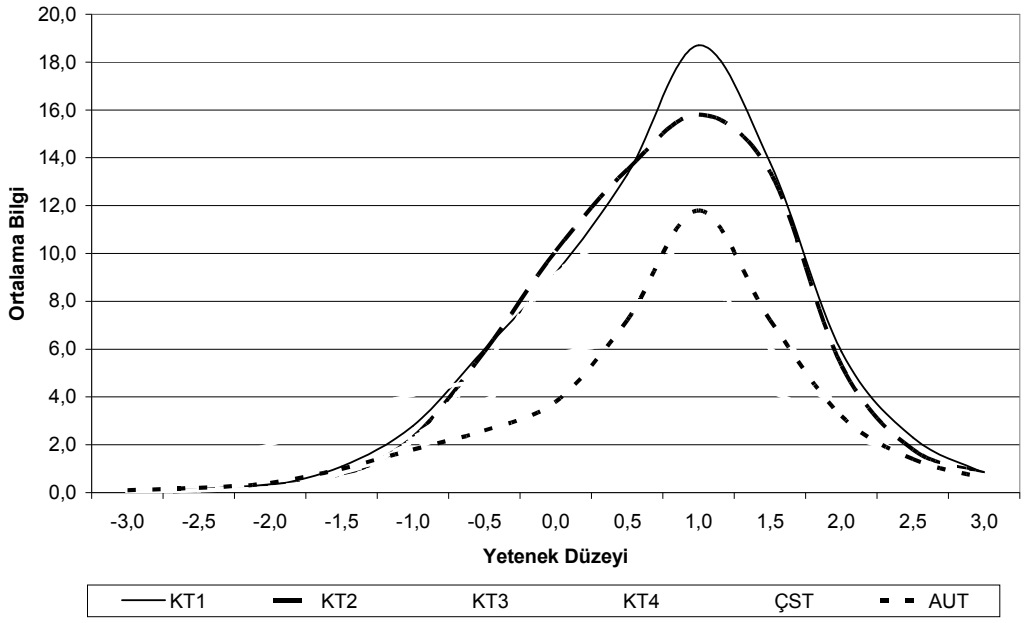
Test Adı	**KT1**	**KT2**	**KT3**	**KT4**	**ÇST**	**AUT**
KT1	1	1.02	1.19	1.16	1.09	0.76
KT2	0.98	1	1.17	1.13	1.07	0.74
KT3	0.84	0.85	1	0.97	0.92	0.63
KT4	0.86	0.88	1.03	1	0.94	0.65
ÇST	0.91	0.93	1.09	1.06	1	0.69
AUT	1.32	1.34	1.57	1.53	1.44	1

Göreli etkinlik indeksi değerlerine göre, KT1; KT2'den %2, KT3'ten %19, KT4'ten %16, ÇST'den %9 daha uzun bir test gibi ölçülen özelliğe ilişkin bireylerarası farklılıkları daha fazla ortaya koyduğu görülmüştür. Öte yandan, AUT; KT1'den %32, KT2'den %34, KT3'ten %57, KT4'ten %53, ÇST'den %44 daha uzun bir test gibi etkili olmuştur. Bu durum, AUT'un madde sayısı az olmasına rağmen, daha uzun testlere göre testle ölçülen özellik bakımından bireylerarası farkları göstermede daha etkili olduğunu göstermiştir. Ayrıca, AU maddelerin karma testler arasında madde sayısı 25 ve açık uçlu madde yüzdesi %40 olan KT1'in, AUT haricindeki diğer testlerden daha etkili olduğu söylenebilir.

Test bilgi fonksiyonu yorumlanırken; yetenek düzeyi ve test bilgi fonksiyonu arasındaki ilişki göz önünde tutulmalıdır (Baker, 2001). Dolayısıyla testten elde edilen bilgi, hangi yetenek düzeylerinde en yüksek değerlerini alıyorsa, o test o yetenek düzeyindeki bireylere hitap ediyor demektir. Bu çıkarımdan hareketle, dört farklı karma testin, çoktan seçmeli ve açık uçlu testten elde edilen bilgi miktarının yetenek düzeyi aralıklarına göre dağılımı Tablo 6 ve Şekil 1'de verilmiştir.

Tablo 6. *Yetenek Düzeylerine Göre Test Bilgi Fonksiyonları Dağılımı*

Testler	**Yetenek Düzeyi**												
	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0
KT1	0.0	0.1	0.3	**1.1**	**2.8**	**6.0**	**9.2**	**13.3**	**18.7**	**13.8**	**5.9**	**2.3**	**0.8**
KT2	0.0	0.1	0.3	**0.7**	**2.3**	**5.9**	**10.1**	**13.5**	**15.8**	**13.4**	**5.3**	**1.8**	**0.8**
KT3	**0.7**	**1.2**	**2.0**	**3.1**	**4.1**	**4.7**	**4.9**	**5.3**	**5.5**	**4.3**	**2.6**	**1.4**	**0.7**
KT4	0.0	0.1	0.2	0.6	**1.7**	**3.8**	**5.7**	**6.4**	**7.8**	**7.0**	**3.4**	**1.2**	0.5
ÇST	0.0	0.1	0.2	**0.8**	**2.2**	**5.4**	**9.1**	**10.6**	**9.2**	**6.1**	**3.3**	**1.4**	**0.7**
AUT	0.1	0.2	0.4	**1.0**	**1.8**	**2.6**	**3.8**	**7.2**	**11.8**	**7.2**	**3.2**	**1.4**	0.6



Şekil 1. Yetenek Düzeylerine Göre Test Bilgi Fonksiyonları Dağılımı

Genel olarak, araştırma kapsamında düzenlenen testlerin tümünün farklı yetenek düzeylerinde bilgi verdiği görülmektedir. Bununla birlikte, Tablo 6 ve Şekil 1 incelendiğinde, KT3'ün (k=15, AU yüzdesi: %40), hemen her yetenek düzeyinde bilgi verdiği, özellikle alt yetenek düzeyinde (-3,-1) diğer testlere göre daha fazla bilgi verdiği görülmektedir. KT3'ün, b değerleri bakımından ortalama olarak en düşük olan test, diğer bir deyişle diğerlerine göre kolay test olduğu önceki kısımlarda ortaya konmuştu. Test bilgi fonksiyonu, madde güçlüğü ve ayırıcılıkla ilişkili olduğundan, alt yetenek düzeylerine hitap eden KT3, alt yetenek düzeylerinde daha fazla bilgi vermiştir.

Diğer testlerin benzer yetenek aralığında bilgi verdiği, sadece KT4'ün (k=15, AU yüzdesi: %20) diğerlerine göre biraz daha dar bir yetenek aralığında bilgi verdiği belirlenmiştir.

Test bilgi fonksiyonu, madde bilgi fonksiyonlarının toplanmasıyla elde edilmektedir (Embretson ve Reise, 2000). Bu durumda madde sayısı fazla olan testten daha fazla bilgi elde edilmesi beklenir. Buna rağmen; en az madde sayısına sahip olan AUT, θ=1 düzeyinde madde sayısı daha fazla olan ÇST, KT3 ve KT4'ten daha fazla bilgi sağlamıştır. Öte yandan, AUT en yüksek bilgiyi (0.5 – 1.5) yetenek aralığında vermiştir. AUT, en zor test olduğundan orta ve üst yetenek düzeylerinde daha fazla bilgi vermiştir.

Test bilgi fonksiyonu, madde ayırıcılığı ve şans parametresiyle de ilişkilidir. Ancak, araştırma kapsamında ele alınan testlerin ayırıcılık düzeyleri ve şans parametreleri (3PL için) birbirine yakın bulunduğundan, bu noktada test bilgi fonksiyonları üzerindeki farklı bir etkisi görülememiştir.

TARTIŞMA, SONUÇ VE ÖNERİLER

Araştırma, benzer yapıları ölçen iki madde tipinin (çoktan seçmeli ve yanıtı sınırlı açık uçlu) ayrı ayrı (tek başına) ve birlikte kullanıldıkları koşullarda oluşturulan testlerde, madde ve yetenek parametreleri; madde ve test bilgi fonksiyonları ile göreli etkinlik indeksleri bakımından farklılıklar olup olmadığı ortaya konmaya çalışılmıştır. Bu amaç doğrultusunda, TIMSS 2007 matematik testi ikinci kitapçığındaki maddelerden test uzunluğu (k= 15 ve 25) ve açık uçlu madde yüzdesi farklı (%20 ve %40) olan testlere ilişkin, madde ve yetenek parametreleri, madde ve test bilgi fonksiyonları, göreli etkinlik indeksleri kestirilmiştir.

Araştırma sonunda test maddelerinden elde edilen bilgi düzeyini gösteren madde bilgi fonksiyonları, en fazla bilginin açık uçlu test maddelerinden sağlandığını ortaya koymuştur. Bu bulgu, Lukhele ve diğerleri (1994) ve Sykes ve diğerleri (2001) yanıtı sınırlı açık uçlu maddelerin çoktan seçmeli maddelerden daha fazla bilgi verdiği bulgusuyla örtüşmektedir. Farklı testlerdeki açık uçlu

maddelerden sağlanan ortalama bilginin, çoktan seçmeli maddelerden sağlanan bilgiden daha fazla olduğu sonucuna ulaşılmıştır.

Test bilgi fonksiyonlarının birbirlerine oranlanmasıyla elde edilen göreli etkililik (relative efficiency) değerleri, AUT'un madde sayısı az olmasına rağmen, daha uzun testlere göre daha etkili olduğunu göstermiştir. Ayrıca, açık uçlu madde yüzdesinin testin %40'ını oluşturduğu ve uzun karma testin açık uçlu test haricindeki diğer testlerden ölçülen özellik bakımından bireylerarası farklılıkları daha etkili ortaya koyduğu sonucuna ulaşılmıştır.

Genel olarak araştırma kapsamında düzenlenen tüm testlerin yetenek düzeyi ölçeğinin (-3, +3) çoğunu kapsayacak şekilde bilgi verdiği belirlenmiştir. Öte yandan, açık uçlu madde yüzdesi yüksek kısa karma testin, alt yetenek düzeylerinde daha yüksek bilgi verdiği; açık uçlu testin ise üst yetenek düzeylerinde daha fazla bilgi verdiği sonucuna ulaşılmıştır.

Araştırmanın bulgularından hareketle, geniş ölçekli test uygulamalarında kullanılacak testlerde, testin tamamının açık uçlu maddelerden oluşmasının puanlamada yaratacağı güçlükler de göz önünde bulundurularak açık uçlu madde yüzdesinin yarıya yakın olduğu ve ölçülen kapsamı (konu ve ölçülen bilişsel beceriler) yeterince iyi temsil eden uzunluktaki karma testlerin kullanılması, testten elde edilen bilgiyi artırma (Lukhele ve diğerleri, 1994, Ercikan ve diğerleri, 1998) bakımından yarar sağlayacağı düşünülmektedir. Bu araştırmada kullanılan testlerde, yanıtı sınırlı açık uçlu madde formatı kullanılmıştır. Bu maddelerin yanıtları yer aldığı TIMSS 2007 Matematik testinde de iki kategorili (0-1) olarak puanlanmıştır. puanlamada kullanılan anahtarlar denemelerden geçerek geliştirilmiş, puanlayıcı güvenirliği sağlanmış anahtarlardır (IEA, 2008-1, s.32). Tekniğine uygun hazırlanmış, üzerinde deneme çalışmalarının yapıldığı dereceli puanlama anahtarları kullanmanın yanıtı sınırlı açık uçlu maddelerin puanlanmasındaki nesnellik sorununu önemli ölçüde azalttığı, bu şekilde güvenilir puanlama sonuçları elde edildiğini gösteren çalışmalar bulunmaktadır (Bennet, 1991; Johnson ve diğerleri, 2000) Ölçme alanyazının da özellikle seçme, yerleştirme, öğrenmeleri izleme amacıyla gerçekleştirilen geniş ölçekli test uygulamalarında tek madde formatı kullanmaktan uzaklaşma farklı madde formatlarını birarada kullanmak daha çok tercih edilmektedir. bu durum her madde formatının tek başına tüm avantajları sağlayamadığı bir formatın dezavantajının diğer formatın avantajı ile giderilebileceği gerçeğinin bir sonucudur. YGS, LYS veya yine çoktan seçmeli maddelerden oluşan ÜDS, ALES gibi testlerde iki formatın birlikte kullanma yönünde deneme çalşmalarının yapılableceği yönünde ipuçları vermiştir. Değinilen tüm noktalar birlikte ele alındığında; bu tür geniş ölçekli testlerde iki madde formatının birbirlerine yakın yüzdelerde kullanılması, çoktan seçmeli maddelerden gelen şans başarısı hatalarının puanlara karışmasını önleyebilecek, nitelikli sorular olmak koşuluyla ölçülen zihinsel becerilerin düzeyi de yükselebilecektir. Bu durumda verilen seçme veya yeterlik kararları daha geçerli ve güvenilir öçme sonuçlarına dayanması sağlanabilecektir.

KAYNAKÇA

Baker, F. B. (2001). *The Basis of Item Response Theory.* USA: ERIC Clearinghouse on Assessment and Evaluation.

Baker, F. B., Kim, S., (2004), *Item Response Theory Parameter Estimation Techniques.* New York:Marcel Dekker, Inc.

Bastari, B., (2000), *Linking Multiple-Choice and Constructed-Response Items to a Common Proficiency Scale.* Doctoral Dissertation. University of Massachusetts Amherst.

Bennett, R.E., and others, (1991), *The Convergent Validity of Expert System Scores for Complex Constructed-Response Quantitative Items.* GRE Research. GRE Board Professional Report No. 88-07bP.

Berberoğlu, G., (2006), *Sınıf İçi Ölçme ve Değerlendirme Teknikleri.* İstanbul: Morpa Kültür Yayınları.

Berberoğlu, G., (2009), CİTO Türkiye Öğrenci İzleme Sistemi (ÖİS) Öğrenci Sosyal Gelişim Programı'na (ÖSGP) İlişkin Ön Bulgular, *CITO Eğitim: Kuram ve Uygulama Dergisi*, Kasım-Aralık Sayısı, 32-42.

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

Demirtaşlı, N. (2010). Açık uçlu soru formatı ve öğrenci izleme sistemi (ÖİS) akademik gelişimi izleme ve değerlendirme (AGİD) modülündeki kullanımı. *Cito Eğitim: Kuram ve Uygulama.* Nisan-Haziran, 21-30.

Embretson, S. E., Reise, S. P., (2000), *Item Response Theory For Psychologists.* New Jersey: Lawrence Erlbaum Associates, Publishers.

Ercikan, K., Schwarz, R.D., Julian, M.W., Burket, G.R., Weber, M.M., Link, V., (1998), Calibration and Scoring of Tests With Multiple-Choice and Constructed-Response Item Types. *Journal of Educational Measurement,* Vol. 35, No. 2, 137-154.

Gonzales, P. (2008), *Highlights From TIMSS 2007, Mathematics and Science Achievement of U.S. Fourth and Eighth Grade Students in an International Context*, National Center for Education Statistics, USA. (http://nces.ed.gov/pubs2009/2009001.pdf adresinden 20.05.2009 tarihinde alınmıştır.)

Haladyna, T. M. (1997). *Writing Test Item to Evaluate Higher Order Thinking*. USA: Allyn & Bacon.

Hambleton, R. K., Swaminathan, H., Rogers, H. (1991), *Fundamentals of Item Response Theory.* Newbury Park CA: Sage Publications.

Hambleton, R. K., Swaminathan, H. (1985), *Item Response Theory. Principles and Applications.* Boston: Kluwer Academic Publishers.

IEA, (2005), TIMSS 2007 Assessment Frameworks, International Study Center, Lynch School of Education, Boston College: USA. (http://timss.bc.edu/ adresinden 03.04.2009 tarihinde alınmıştır.)

IEA, (2008-1), TIMSS 2007 Technical Report, International Study Center, Lynch School of Education, Boston College: USA. (http://timss.bc.edu/ adresinden 03.04.2009 tarihinde alınmıştır.)

IEA, (2008-2), TIMSS 2007 International Mathematics Report, International Study Center, Lynch School of Education, Boston College: USA. (http://timss.bc.edu/ adresinden 03.04.2009 tarihinde alınmıştır.)

Johnson, R.L., Penny, J., Gordon, B. (2000), The Relation Between Score Resolution Methods and Interrater Reliability: An Empirical Study of an Analytic Scoring Rubric. *Applied Measurement in Education*, Vol. 13, Issue 2

Kinsey, T. L. (2003), *A Comparison of IRT and Rasch Procedures in a Mixed-Item Format Test.* University of North Texas. Doctoral Dissertation.

Lord, F. M., Novick, M. R. (1968), *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Lukhele, R., Thissen, D., Wainer, H. (1994), On The Relative Value Of Multiple Choice, Contructed Response, And Examinee Selected İtems On Two Achievement Tests. *Journal Of Educational Measurement*, 31, 231-250.

Milli Eğitim Bakanlığı Talim Terbiye Kurulu Başkanlığı, (2005), *İlköğretim Matematik Dersi Öğretim Programı ve Kılavuzu 1-5. Sınıflar.* Ankara: Devlet Kitapları Müdürlüğü.

Milli Eğitim Bakanlığı Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı, (2007), *PISA 2006 Uluslararası Öğrenci Başarılarını Değerlendirme Programı Ulusal Ön Rapor.* (28.01.2009 tarihinde http://earged.meb.gov.tr/pisa/dokuman/2006/rapor/Pisa_2006_ Ulusal_On_Rapor.pdf adresinden alınmıştır.)

Schaeffer, G. A., Montero, D. H., Julian, M., Bené, N. H., (2002), A Comparison of Three Scoring Methods for Tests With Selected-Response and Constructed-Response Items. *Educational Assessment*, 8(4), 317–340

Shin, D., (2007), *A Comparison of Method of Estimating Subscale Scores for Mixed-Format Tests*. Pearson Educational Measurement Research Reports. (24.12.2008 tarihinde www.pearsonedmeasurement.com/ research/research.htm adresinden alınmıştır.)

Shermis, M. D., Burstein, J. C., (2003), *Automated Essay Scoring: A Cross-Disciplinary Perspective.* Mahwah, NJ: Lawrence Erlbaum Associates.

Sykes, R. C., Truskosky, D., White, H. (11-12 April 2001), *Determining The Representation of Constructed Response Items in Mixed-Item-Format Exams.* Paper presented at Annual Meeting of the National Council on Measurement in Education, Seattle

Tekin, H., (1991), *Eğitimde Ölçme ve Değerlendirme*. Ankara: Yargı Yayınları.

Umay, A. (1997), Yanıtlayıcı Davranışların Analizi Yolu İle Matematikte Problem Çözümleri İçin Bir Güvenirlik ve Geçerlik Araştırması, *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*. 13, 47-56.

Uyeno, R. K. (2004), Assessing The Content Standarts of a Large-Scale, Standards-Based Test: A Psychometric Validty Study of The 2002 Hawai'i State Assessment Grade 8 and Grade 10 Reading Tests. University of Hawai'i. Doctoral Thesis.

Wainer, H., Thissen, D. (1993), Combining Multiple-Choice and Constructed-Response Test Scores: Toward a Marxist Theory of Test Construction. *Applied Measurement in Education*, 6(2), 103-118.

Zhao, Y. (2008). Approaches For Addressing The Fit Of Item Response Theory Models To Educational Test Data. University of Massachusetts: Doctoral Thesis.