Eğitim Bilimleri Araştırmaları Dergisi
ULUSLARARASI E-DERGİ
Cilt:4 Sayı:1 Nisan 2014

EBAD-JESR

Journal of Educational Sciences Research
INTERNATIONAL E-JOURNAL
Vol:4 No:1 April 2014

Bilgisayar Ortamında Bireye Uyarlanmış Test Stratejilerinin Karşılaştırılması[1]

Fatih KEZER[2] & Nizamettin KOÇ[3]

ÖZET

Bu araştırmada, bilgisayar ortamında bireye uyarlanmış test (BOBUT) yöntemi ile geleneksel kâğıt kalem test yönteminin karşılaştırılması ve BOBUT yöntemine ilişkin farklı stratejilerin karşılaştırılması amaçlanmıştır. Temel araştırma modelindeki araştırmanın verileri, Ankara Üniversitesi Yabancı Diller Yüksekokulu bünyesinde, hazırlık sınıfında öğrenim görmekte olan toplam 1166 öğrenciden toplanmıştır. Farklı stratejilerin karşılaştırılması amacıyla R programı ile simülatif veriler de oluşturulmuştur. Araştırmada veri toplama aracı olarak İngilizce Kelime Testi kullanılmıştır. BOBUT uygulamalarının yapılabilmesi için araştırmacı tarafından bir çevrimiçi ortam geliştirilmiştir. Araştırma sonucunda, çevrimiçi ortam kullanılarak yapılan bireye uyarlanmış test uygulamasında, kâğıt kalem testine göre madde sayılarında büyük oranda tasarruf sağlandığı saptanmıştır. Bireye uyarlanmış ve kâğıt kalem test uygulamalarından elde edilen yetenek parametreleri arasında da pozitif yönde yüksek korelâsyon katsayıları bulunmuştur. Farklı stratejiler ve kağıt kalem testinden elde edilen yetenek parametreleri arasında pozitif yönde yüksek korelasyon katsayıları bulunarak, araştırma kapsamında ele alınan 18 farklı strateji ile kağıt kalem testlerinin çok benzer yetenek parametrelerinin kestirildiği ortaya konmuştur. Aynı şekilde farklı stratejiler ile kestirilen yetenek parametrelerinin kendi aralarında pozitif yönde yüksek korelasyon katsayıları elde edilmiştir. Sonlandırma kuralları dikkate alındığında, farklı stratejilerden elde edilen yetenek kestirimlerinin gerek kağıt kalem testinden elde edilen yetenek parametreleri ile arasında gerekse kendi aralarında en düşük korelasyon katsayılarının sonlandırma kuralı olarak standart hatanın 0.50'den küçük olması durumunda elde edildiği saptanmıştır. ML, EAP ve MAP yetenek kestirim yöntemlerinden kaynaklı, kestirilen yetenek parametrelerinde farklılık olmadığı görülmüştür.

Anahtar Sözcükler: Bilgisayar ortamında bireye uyarlanmış test, Test stratejileri, Madde tepki kuramı

DOI Number: http://dx.doi.org/10.12973/jesr.2014.41.8

[1] Bu makale, Fatih Kezer'in, Ankara Üniversitesi Eğitim Bilimleri Enstitüsü'nde, Prof. Dr. Nizamettin KOÇ danışmanlığında yapmış olduğu "Bilgisayar Ortamında Bireye Uyarlanmış Test Stratejilerinin Karşılaştırılması" (2013) adlı doktora tezinden üretilmiştir.

[2] Dr. - Kocaeli Üniversitesi Eğitim Fakültesi - fatihkezer@yahoo.com
[3] Prof. Dr. - Ankara Üniversitesi, Eğitim Bilimleri Fakültesi - nkoc@ankara.edu.tr

GİRİŞ

Bilgisayar Ortamında Bireye Uyarlanmış Test (BOBUT) yöntemi, birçok avantajıyla bugün dünyada pek çok ülkede giderek artan bir kullanım alanı bulmaktadır. BOBUT, Madde Tepki Kuramı'nin (MTK) güçlü algoritması ile geçerli ve güvenilir bir test etme süreci sağlamaktadır. Özellikle MTK'nın getirilerinden biri olan madde karakteristik eğrisi doğrultusunda kullanılan soruların niteliğine ilişkin elde edilen detaylı bilgi ile bireyin yetenek düzeyinin belirlenmesinde maddeler daha verimli kullanılabilmektedir. Mevcut ölçme ve değerlendirme yaklaşımları göz önüne alındığında gruba bağlı madde parametreleri ve maddelere bağlı yetenek parametreleri kestirimi sınırlılığını ortadan kaldıran değişmezlik özelliği ile de bireylerin yeteneği hakkında daha geçerli ve güvenilir bir kestirim imkânı sağlamaktadır (Crocker & Algina, 1986; Hambleton, Swaminathan & Rogers 1991). Farklı testlerden elde edilen yetenek parametrelerinin karşılaştırılabilirliğini sağlayan bu avantaj ile test standardizasyonu da kolaylaşmaktadır.

BOBUT'ta temel işleyiş, bir testte bireyin karşısına en uygun maddeyi çıkarmak ve yetenek düzeyini (θ) bu doğrultuda belirlemektir. Yetenek düzlemi boyunca herhangi bir noktada bireyler hakkında bir maddenin sağladığı bilgi sadece bu maddenin parametrelerine bağlıdır (Bejar 1983; Hambleton, Swaminathan & Rogers 1991; Folk & Smith, 2002). Böylelikle bireyin yetenek düzeyi için madde parametreleri kullanılarak en fazla bilgi veren madde belirlenebilir. Her madde sonrasında, sonlandırma kuralı gerçekleşene kadar, yetenek kestirimi döngüsüyle bireyin karşısına belirlenen yetenek düzeyi için en çok bilgi sağlayan madde çıkarılmaktadır.

BOBUT'ta öncelikle bireyin yetenek düzeyine ilişkin bir tahminde bulunulur. Bu öncül test ya da bireyin daha önceki performanslarına dayalı olabilmektedir (Segall, 2003). Birey bazında daha önceki performanslara göre yetenek tahmini yapılabileceği gibi, grup ortalaması alınarak her birey için aynı başlangıç yetenek tahmini de zaman zaman kullanılmaktadır. Her yetenek düzeyi için nitelikli ölçme yapabilmek adına madde havuzunda her yetenek düzeyine hitap eden ve yüksek ayırtedicilikte çok sayıda madde bulunmalıdır. Bireye uyarlanmış test en etkili olarak çok sayıda yüksek derecede ayırt ediciliği olan ve güçlük-özellik düzeyinde (b-θ) eşit biçimde temsil edilen maddelerden oluşan bir madde havuzuyla çalışmaktadır (Weiss, 1985; Georgiadou, Triantafillou & Economides, 2006; Veldkamp & Linden, 2010; Weiss, 2011). Madde havuzu birçok madde yanıt formatı içerebilir (Embretson & Reise, 2000; Wainer vd., 2000; Sukamolson, 2002).

Yetenek kestiriminde bulunmak için farklı matematiksel yöntemler bulunmasına karşın başlıca iki yaklaşım kullanılmaktadır: "En Çok Olabilirlik (Maximum Likelihood-ML)" yöntemi ile Bayes/Bayesçi (Bayesian) istatistiğe dayanarak geliştirilen "En Yüksek Posteriori/Sonsal Maksimum Kestirim (Maximum A Posteriori-MAP)" ve "Posteriori İçin Beklenen Değer/ Sonsal Beklenti Kestirimi (Expected A Posteriori-EAP)" yöntemleridir (Embretson & Reise, 2000; Baker & Kim, 2004; Linden & Pashley, 2010). ML yöntemi, birey hakkında en fazla bilgi veren maddeyi seçmeye dayanmaktadır. Bu seçimi yaparken de *olabilirlik fonksiyonunu* kullanır. Olabilirlik fonksiyonu, bireyin maddelere verdiği yanıtlar birbirinden bağımsız olmak üzere, yanıt olasılıklarının çağrımı biçiminde tanımlanabilir. Olabilirlik fonksiyonunu en yüksek yapan θ değeri bulunarak en çok olabilirlik kestirimi yapılabilir. ML yöntemi ile ilgili en büyük sıkıntı, bireyin maddelere tümüyle doğru ya da tümüyle yanlış yanıt verdiği durumda yetenek tahmini yapamamasıdır. Tümüyle doğru yanıt örüntüsü pozitif sonsuzlukta monoton artan; tümüyle yanlış örüntüsü de negatif sonsuzlukta monoton azalan bir fonksiyona sahiptir. Böyle bir durumda fonksiyonu en yüksek yapan değeri bulmak mümkün değildir. Bir başka problem ise madde karakteristik

eğrilerinin sıfır değerini de içeriyor olmasıdır. Bu problemi aşmak için çoğunlukla log-olabilirlik (loglikelihood-logL) tercih edilmektedir. Log-L, madde karakteristik eğrisinin doğal logaritmasının alınmasıyla elde edilmektedir. Log-L fonksiyonu da madde karakteristik eğrilerinin toplanmasıyla oluşmaktadır. Yetenek düzeyi kestirimi yapılırken Log-L fonksiyonunu en büyük yapan değer alınmaktadır.

Bayesçi yöntemler ise bireyin sonraki yetenek kestirim aralığını en aza indirecek soruyu seçmektedir (Lord & Stocking, 1988; Hambleton & Swaminathan, 1989; Rudner, 1998; Baker & Kim, 2004). ML yönteminin tümü doğru ya da tümü yanlış yanıt örüntülerinde kararlı sonuçlar vermemesi, büyük örneklemlerde sonuç vermesi gibi sıkıntılar bu yöntemin her durumda koşulsuz şartsız kullanılmasının önüne geçmiştir. ML yönteminin sınırlılıkları öncül bir dağılım ile giderilmeye çalışılmıştır. MAP yöntemi, öncül dağılım kullanılarak kestirim yapan bir yöntemdir. Burada kullanılan öncül dağılımlar, dağılımın birey yetenek düzeyi kestirimini öncülün ortalamasına doğru çekmesi ve kestirimin standart hatasını düşürmesi bakımından kısa testlerde etken rol oynamaktadır. Bu durum bir taraftan avantaj iken diğer taraftan bu yöntemin zayıf yönü olarak dikkat çekmektedir. Madde sayısı 20'den az olduğunda nispeten daha yanlı sonuçlar verdiği iddia edilmektedir (Embretson & Reise, 2000). Yanlış öncül kullanılması durumunda da bireye ilişkin yetenek kestirimi gerçekten sapmakta ve yanıltıcı olmaktadır.

EAP yöntemi ise ML ve MAP yöntemleri gibi iteratif bir yöntem değildir. Tüm tepki örüntüleri için sonlu bir yetenek düzeyi kestirimi yapmaktadır. İteratif olmaması, kolay ve hızlı hesaplanabilir olması güçlü yönü iken, madde sayısı sonlu olduğundan yanlı olması bu yöntemin zayıf tarafıdır.

BOBUT algoritmasında kullanılan farklı sonlandırma kuralları bulunmaktadır. Yetenek kestirimine ait standart hatanın yeterince küçük olması sık kullanılan sonlandırma kurallarından biridir. Her madde sonunda kestirilen yetenek parametresine ilişkin elde edilen standart hata belli bir değerin altına düşmüş ise yeterli keskinlikte ve kararlılıkta yetenek kestirildiği için test sonlandırılmaktadır. Standart hatanın 0.20'nin altında olması yaygın olarak kullanılan eşik değerdir. Yine standart hata için 0.25, 0.30 ve 0.50 gibi sınır değerler de kullanılmaktadır. Yetenek kestirimine ilişkin bir başka kararlılık göstergesi de standart hatalar arasındaki farkın giderek küçülmesidir. Son iki yetenek kestirimine ilişkin standart hata değerleri arasındaki fark 0.01'den küçük olduğunda da yine test sonlandırılabilmektedir. Farklı durumlar için sabit uzunluk tercih ediliyorsa testin belirlenmiş uzunluğa erişmesi de sonlandırma kuralı olarak kullanılmaktadır. Bunun dışında küçük madde havuzu ile çalışıldığında madde havuzunda madde kalmaması, adayın test dışı davranışlar sergilemesi de testi sonlandıracak kurallar arasında değerlendirilmektedir (Hambleton, Swaminathan & Rogers, 1991; Segall, 2003; Way, 2006; Tian, Miao, Zhu & Gong, 2007; Choi, Grady & Dodd, 2011; Weiss, 2011).

BOBUT, temel aldığı MTK doğrultusunda güçlü varsayımlara dayanması, donanım ve yazılım gereksinimi, geniş madde havuzu ihtiyacı gibi sınırlılıklardan dolayı bugüne kadar Türkiye'de yaygın olarak kullanım imkânı bulamamıştır. Merkezi sınavlar, hali hazırda Klasik Test Kuramı'na dayalı geleneksel kâğıt kalem testi olarak uygulanmaktadır. Söz konusu avantajları göz önüne alındığında özellikle geniş ölçekli testler için BOBUT uygulamalarının yaygınlaştırılması gerekmektedir. Bu anlamda BOBUT'un yapısının en iyi şekilde anlaşılması ve bu yönde çalışmaların yapılması gerekmektedir.

Bu araştırmada, geleneksel kâğıt kalem testlerinin aksine bireylerin yetenek düzeyleri farklılığını esas alarak bireylere özgü test imkânı sunan ve MTK'nın bir uygulaması olan bilgisayar ortamında bireye uyarlanmış test yöntemi ile kâğıt kalem test yönteminin

karşılaştırılması ve bilgisayar ortamında bireye uyarlanmış test yöntemine ilişkin farklı stratejilerin bir İngilizce Kelime Testi çerçevesinde karşılaştırılması amaçlanmıştır. Bu doğrultuda BOBUT uygulamasında madde sayısının dağılım, BOBUT ve kâğıt kalem testi uygulamalarında kestirilen yetenek parametreleri arasında manidar bir ilişki olup olmadığı, benzer şekilde simülatif veriler doğrultusunda farklı başlatma ve sonlandırma kuralları ile yetenek kestirim yöntemlerinin farklılığına göre yetenek parametreleri arasında manidar bir ilişki olup olmadığı incelenmiştir.

YÖNTEM

Araştırma Modeli

Bu araştırmada MTK'nın bir uygulaması olan bilgisayar ortamında bireye uyarlanmış test stratejileri karşılaştırılmıştır. Araştırma, BOBUT uygulamalarında kullanılan başlatma ve sonlandırma kuralları ile yetenek kestirim yöntemlerine ilişkin karşılaştırmaları içermesi bakımından temel araştırma niteliğindedir.

Çalışma Grubu

Araştırmanın çalışma grubunu, Ankara Üniversitesi Yabancı Diller Yüksekokulu bünyesinde, 2012-2013 eğitim öğretim yılında hazırlık sınıfında öğrenim görmekte olan öğrenciler oluşturmaktadır. Veriler, Yabancı Diller Yüksekokulu bünyesinde her öğretim yılı başlangıcında üniversiteyi kazanan öğrencilere yapılan "Seviye Tespit Sınavı" sonucu A1, A2 ve B1 olarak adlandırılan üç farklı düzeyden, oluşturulan testin düzeyine uygun olacak şekilde A1 ve A2 düzeylerinde okuyan öğrencilerden toplanmıştır.

Araştırma verileri, üç farklı aşama ile elde edilmiştir. Her aşamada faklı öğrencilerden veri toplanmıştır. İlk aşamada, kullanılacak olan "İngilizce Kelime Testi" nin iki farklı ön deneme uygulaması yapılmıştır. Testte kullanılan maddelerin niteliğini belirlemek, maddelerin anlaşılabilirliğini kontrol etmek, test süresini tespit etmek gibi amaçlarla toplam 105 öğrenci ile ön deneme uygulamaları yapılmıştır. Öncelikle A1 ve A2 düzeylerinde bulunan 29'u erkek ve 34'ü kız toplam 63 öğrenciden birinci ön deneme uygulaması ile veri toplanmıştır. İkinci ön deneme uygulaması için yine A1 ve A2 düzeylerindeki sınıflardan 19'u erkek, 23'ü kız olmak üzere toplam 42 öğrenciyle uygulama yapılmıştır.

Araştırmanın ikinci aşamasında bilgisayar ortamında bireye uyarlanmış test uygulaması için havuzu oluşturacak maddelerin belirlenmesi ve kâğıt-kalem uygulamasına ait yetenek puanlarının kestirilmesi amacıyla A1 ve A2 düzeylerine ait 52 sınıftan 470'i erkek, 524'ü kız olmak üzere toplam 1078 öğrenciden veri toplanmıştır. Optik okuyucu hatası, cevap kâğıdını boş bırakma, soruların çoğunluğuna yanıt vermeme, düzeyin uygun olmaması vb. nedenlerden dolayı toplamda 94 öğrenci veri setinden ayıklanmıştır. Maddelerin psikometrik özelliklerinin belirlenmesi ve BOBUT uygulaması için maddelerin MTK varsayımlarını karşılaması ve kalibrasyonu amacıyla kalan 994 öğrencinin verisi kullanılmıştır.

Araştırma kapsamında belirlenen başlangıç kuralları, yetenek kestirim yöntemleri ve sonlandırma kurallarının karşılaştırılmasındaki simülasyon çalışması için yine aynı 994 öğrencinin verisi kullanılmıştır.

Araştırmanın üçüncü aşamasında simülasyon çalışmaları haricinde gerçek uygulama ile de kağıt kalem testi ve bilgisayar ortamında bireye uyarlanmış test sonuçları karşılaştırılmıştır. Bunun için A1 ve A2 düzeyinden 5 farklı sınıftan 32'si erkek, 45'i kız olmak üzere toplam 77 öğrenciden yararlanılmıştır. Öğrenciler veri toplama aracı olan

'İngilizce Kelime Testi'ni hem kâğıt-kalem formunda hem de bilgisayar ortamında yanıtlamışlardır. Kâğıt-kalem testinin % 70'inden daha azını yanıtlayan (boş bırakan) öğrenciler çalışma dışında bırakılmıştır. Böylelikle toplam 72 sorudan 49 ve altında soru yanıtlamış olan 10 öğrenci çalışma kapsamı dışında bırakılmış ve analizler, kalan 67 öğrencinin verileri üzerinden gerçekleştirilmiştir.

Veri Toplama Araçları

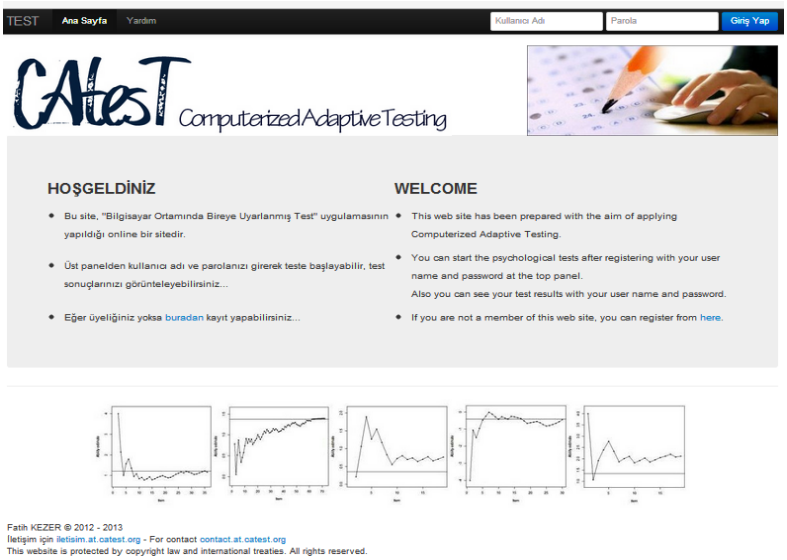
İngilizce kelime testi

Çalışmada kullanılan "İngilizce Kelime Testi" Ankara Üniversitesi Yabancı Diller Yüksekokulu *"Koordinatörlük"* ve *"Sınav Hazırlama ve Değerlendirme"* birimlerinden üç İngilizce uzmanı ile birlikte hazırlanmıştır. İngilizcede en çok kullanılan ilk 3000 kelime arasından öncelikli olarak seçenekler için 500 kelime seçilmiştir. Oluşturulan 100 soruluk testin ilk 35 sorusu '*General Service List*'teki (İngilizcede kullanılan kelimelerin sıklığına göre oluşturulmuş uluslararası bir liste) ilk 1000 kelimeden, ikinci 35 sorusu da yine aynı listedeki ikinci 1000 kelimeden, son 30 sorusu ise "*Corpus of Contemporary American English"*e göre en sık kullanılan üçüncü 1000 kelimeden oluşmaktadır.

Test, oluşturulduktan sonra Ankara Üniversitesi Yabancı Diller Yüksekokulu koordinatörlerinden iki İngilizce uzmanına soruların niteliği, doğruluğu, hedef grubun seviyesine uygunluğu açısından incelettirilmiş ve gerekli düzenlemeler yapılmıştır. Kâğıt kalem uygulaması ve BOBUT uygulamasında kullanılan testteki maddelerin işlerliğini belirlemek, uygulama yapılacak hedef grupta işler olmayan maddelerin düzeltilmesi ya da atılması için iki farklı deneme uygulaması yapılmıştır. İlk deneme uygulaması sonrası madde güçlük indeksleri ve madde ayırtedicilik indeksleri hesaplanmıştır. 100 maddeye ilişkin madde güçlük indeksleri 0.00 ile 0.92 arasında bulunmuştur. Madde ayırt edicilik indeksleri -0.61 ile 0.87 arasında saptanmıştır. Madde ayırtedicilik indeksleri 0.20'nin altında olan 8 maddeyle birlikte -denemek amaçlı-madde ayırt edicilik indeksi 0.20 ile 0.27 arasında olan 6 madde daha değiştirilmiştir. Böylelikle 14 maddenin yerine yeni maddeler yazılmıştır. Yazılan yeni maddeler ile 100 maddelik test doğrultusunda ikinci deneme uygulaması yapılmıştır. Uygulama sonrası madde güçlük indeksleri 0.02 ile 0.95; madde ayırtedicilik indeksleri -0,43 ile 1.00 arasında bulunmuştur. İlk denemede işler olmayan (r_{jx}<0.20) maddelerin yerine yazılan maddeler ve ikinci denemede madde ayırtediciliği 0.20'nin altında olan maddeler incelenerek iki deneme sonrasında ilk denemede 0.20'nin üzerinde madde ayırtediciliğine sahip maddelerde düzenleme yapılarak 5 madde tekrar kullanılmıştır. İki deneme sonrasında madde ayırtedicilikleri 0.20 ile 1.00 arasında değişen maddeler kâğıt kalem ve BOBUT uygulamaları için 100 maddelik teste alınmıştır. Testin geçerlik ve güvenirliğine ilişkin kanıtlar ile testin MTK'ya uygunluğuna ilişkin analizler verilerin analizi bölümünde detaylı olarak anlatılmıştır.

Çevrimiçi ortam

Bilgisayar ortamında bireye uyarlanmış testin uygulanabilmesi için araştırmacı tarafından bir çevrimiçi ortam oluşturulmuştur. Uygulamaların ek donanıma ihtiyaç duymaması, farklı platformlar için kurulum gerektirmemesi, cep telefonu ve tablet-pc gibi mobil cihazlarda kullanılabilmesi, erişiminin kolay olması gibi nedenlerle çevrimiçi ortam tercih edilmiştir. Çevrimiçi ortam, araştırmacı tarafından www.catest.orgadresi altında PHP (Personal Home Page - Kişisel Ana Sayfa) programlama dili ile yazılmıştır. Veri tabanı için MySQL veri tabanı yönetim sistemi kullanılmıştır. Anasayfa görünümü Şekil 1'de verilmiştir.



Şekil 1. *Çevrimiçi ortamı ana sayfası*

Testi alabilmek için öncelikli olarak kullanıcıların sisteme üye olması gerekmektedir. Testlerin güvenliği açısından otomatik olarak pasif modda kayıt olan kullanıcılar, yöneticinin aktif etmesi ile ortamda testlere ulaşabilmektedirler.

Test bittikten sonra yine kullanıcılar otomatik olarak pasif moda geçmekte ve sisteme daha sonraki girişlerinde sadece test sonuçlarına erişebilmektedirler. Çevrimiçi ortam, kullanıcıların birden çok testi aynı anda alabilecekleri şekilde tasarlanmıştır. Araştırma kapsamında sadece "İngilizce Kelime Testi" kullanıldığı için tek test üzerinden işlemler yapılmıştır. Testin her aşamasında havuzdaki maddelerin kullanım sıklığı kontrol edilmekte ve en az kullanılan madde diğer şartlar dâhilinde havuzdan seçilmektedir. Testin başlangıcında madde güçlük parametresi (b) -0.50 ile +0.50 arasında olan bir madde rastgele olarak havuzdan seçilerek kullanıcının karşısına çıkarılmaktadır. İlk yedi maddede kullanıcıya ait yetenek kestirimi yapılmadan sadece kullanıcı yanıtına göre madde güçlüğü doğrultusunda sorular belirlenmektedir.

Kullanıcı mevcut maddeye doğru yanıt vermiş ise daha zor bir madde, yanlış yanıt vermiş ise de daha kolay bir madde havuzdan seçilerek kullanıcının karşısına çıkartılmaktadır. Yedinci maddeden sonra hem yetenek kestirimi yapılmakta hem de yetenek kestirimi ile birlikte madde bilgi fonksiyonları hesaplanmaktadır. Yedinci maddeden sonra kullanıcının yetenek düzeyine göre havuzdaki kalan maddelere ilişkin madde bilgi fonksiyonu hesaplanmakta ve o yetenek düzeyi için en çok bilgiyi veren madde bir sonraki madde olarak seçilmektedir. Testin başlangıcında kullanıcılara ait olan –daha önceden kestirilmiş- yetenek puanları kullanılabileceği gibi tüm kullanıcılar için ortak bir yetenek puanı da kullanılabilmektedir. Araştırmada çalışma grubuna dâhil olan bireyler için başlangıç yetenek puanı sıfır (0) olarak girilmiştir. Çevrimiçi ortam için test algoritması oluşturulurken yetenek kestiriminde En Çok Olabilirlik (Maximum Likelihood-ML) yöntemi kullanılmıştır. Hesaplamalar esnasında log-L fonksiyonunun birinci ve ikinci türevinin hesaplandığı Newton-Raphson metodundan faydalanılmıştır. En çok olabilirlik yönteminin bir sınırlılığı sonucu olarak kullanıcıların maddelere tümüyle doğru yanıt vermesi ya da tümüyle yanlış yanıt vermesi durumunda yetenek puanının -∞/+∞ doğru hızlı bir ivmeyle azalması/artması durumundan dolayı kararlı bir kestirim yapılabilmesi için kullanıcı yanıtlarında en az bir doğru, en az bir yanlış şartı aranmıştır. Böylelikle test algoritmasındaki

birçok kontrol koşulundan biri de yanıtlarda en az bir doğru ve en az bir yanlış cevap aranması olmuştur.

Test araştırmacı tarafından belirlenen sonlandırma kuralları gerçekleşene kadar, donanım ve internetle ilgili bir sorun olmadığı takdirde devam etmektedir. Her madde sonunda kullanıcılara ait P(θ), θ, SE(θ) ve maddelere ait I(θ)'lar hesaplanmaktadır. Test bitmeden sistemde bir sorun oluşması, mevcut donanımla ilgili istenmeyen ve hemen müdahale edilemeyen bir durum oluşması halinde kullanıcıya ait test bitirilmemiş olarak sonlandırılmaktadır ve kullanıcı otomatik olarak pasif duruma geçmektedir. Algoritmaya ilişkin bu müdahaleler kontrolün tamamen sistem yöneticisinde (test yapan kişi) olması için düzenlenmiştir. Çevrimiçi ortamda iki farklı sonlandırma kuralı kullanılmıştır. Her madde sonrasında kestirilen yeteneğe ilişkin elde edilen standart hatanın belli bir kararlılığa ulaşmasının göstergesi olarak standart hatalar arasındaki farkın giderek küçülmesi dikkate alınarak son iki yetenek kestirimine ilişkin standart hata değerleri arasındaki fark 0.01'den küçük olduğunda test sonlandırılmaktadır. Diğer bir sonlandırma kuralı olarak, yeteneğe ilişkin elde edilen standart hata değerinin 0.50'nin altında olması kullanılmıştır. İki sonlandırma kuralından biri sağlandığında test sonlandırılmaktadır. Bununla birlikte havuzdaki maddelerin tamamının bitmesi durumunda da testin sonlandırılması algoritmanın stabilliği açısından testi sonlandıracak başka bir faktör olarak göz önüne alınmıştır.

Test bitiminde kullanıcıya test sonuçları; kaç madde yanıtladığı, yanıtladığı maddelerin kaç tanesine doğru, kaç tanesine yanlış yanıt verdiği ve kestirilen yetenek düzeyi şeklinde sonuç ekranında gösterilmektedir. Teste ilişkin detaylı bilgiler kullanıcı ile paylaşılmayıp yönetici panelinde yer almaktadır. Yönetici her kullanıcıya ilişkin test detaylarına ve yetenek kestirimine ilişkin grafiğe rapor ekranında ulaşabilmektedir. Bununla birlikte yönetici, tüm kullanıcıların kaç madde yanıtladıkları, yetenek kestirimleri ve standart hatalarına ilişkin bilgi veren toplu bir rapor da alabilmektedir.

Simülatif verilerin üretilmesi

Araştırmanın amacı doğrultusunda BOBUT'a ilişkin test stratejileri karşılaştırılırken simülatif veriden faydalanılmıştır. MTK'nın varsayımları ve BOBUT uygulamasının esasları dikkate alındığında; geniş madde havuzu oluşturmanın zorluğu, karşılaştırmaları yaparken uygulamalar için çok fazla sayıda kişiye ihtiyaç duyulması sebebiyle araştırmada simülatif veri de kullanılmıştır. Alanyazında da bu tür çalışmalarda sıklıkla simülatif veriye başvurulduğu görülmektedir (McDonald, 2002; Scullard, 2007; Barrada, Olea, Ponsoda & Abad, 2010; Evans, 2010; Kalender, 2011; Smits, Cuijper & Straten, 2011; Bulut & Kan, 2012; Wang, Kuo, Tsai & Liao, 2012; Zitny, Halama, Jelinek & Kveton, 2012; Patton, Cheng, Yuan & Diao, 2013).

BOBUT uygulamalarında kullanılan farklı başlangıç kuralları, yetenek kestirim yöntemleri ve sonlandırma kuralları çerçevesinde araştırma için 18 farklı durum oluşturulmuştur. Testin başlangıcında bireylerin başlangıç yetenek puanları, bir durum için 0 (sıfır), diğer bir durum için ise daha önceden kestirilmiş yetenekler olarak alınmıştır. Yetenek kestirimi için MTK'da mevcut olan üç kestirim yöntemi de ML, MAP ve EAP üç farklı durum olarak alınmıştır. BOBUT uygulamalarında sıklıkla kullanılan sonlandırma kurallarından standart hatanın 0.50'nin altında olması durumu, standart hatanın 0.30'un

altında olması durumu ve sabit uzunluk durumu üç sonlandırma kuralı olarak araştırmada kullanılmıştır. Böylelikle 2x3x3 şeklinde toplam 18 farklı durum oluşmuştur (Tablo 1).

Tablo 1. *BOBUT stratejileri*

Başlangıç Kuralı	*Yetenek Kestirim Yöntemi*	*Sonlandırma Kuralı*
θ; sıfır (0)	ML	Sabit uzunluk
		SE<0.50
		SE<0.30
	EAP	Sabit uzunluk
		SE<0.50
		SE<0.30
	MAP	Sabit uzunluk
		SE<0.50
		SE<0.30
θ; önceden kestirilmiş	ML	Sabit uzunluk
		SE<0.50
		SE<0.30
	EAP	Sabit uzunluk
		SE<0.50
		SE<0.30
	MAP	Sabit uzunluk
		SE<0.50
		SE<0.30

Simülatif verilerin üretilmesi için R açık kaynaklı istatistik programının "catR" kütüphanesinden faydalanılmıştır. Kütüphane, BOBUT çerçevesinde farklı başlatma kuralları, madde seçim prosedürleri, sonlandırma kuralları ve yetenek kestirim yöntemlerini içeren aynı zamanda madde kullanım sıklığı gibi kontrolleri de barındıran tepki setleri üretilmesine olanak sağlamaktadır (Magis & Raiche, 2012).

Araştırmada "catR" kütüphanesi fonksiyonları kullanılarak simülasyon algoritması araştırmacı tarafından yazılmıştır. Her farklı durum için 2 parametreli lojistik model kullanılarak daha önceden kestirilmiş madde parametreleri doğrultusunda 994 kişilik yetenek parametresi örüntüsü üretilmiştir.

Verilerin Analizi

Bu bölümde öncelikli olarak araştırma kapsamında oluşturulan 100 maddelik *İngilizce Kelime Testi*'nin MTK varsayımlarını karşılayıp karşılamadığı incelenmiştir. MTK'nın varsayımlarından tek boyutluluğun sınanmasında açımlayıcı faktör analizi ve doğrulayıcı faktör analizi kullanılmıştır. Açımlayıcı faktör analizinin 1–0 şeklinde puanlanan kategorik verilerde yapılabilmesi için öncelikli olarak STATISTICA programında tetrakorik korelasyon matrisi üretilmiştir.

Oluşturulan tetrakorik korelasyon matrisinden sonra SPSS ve STATISTICA programları aracılığıyla Açımlayıcı Faktör Analizi (AFA) yapılmıştır. AFA ile elde edilen yapının doğruluğunu test etmek amacıyla LISREL programı aracılığıyla, var olan örtük yapının ilgili veri seti ile doğrulanıp doğrulanmadığını test etmede kullanılan Doğrulayıcı Faktör Analizi (DFA) yapılmıştır (Tabachnick & Fidel, 2007). Alanyazında da DFA'nın daha çok klasik faktör analizi çalışmalarından sonra uygulanan bir yöntem olduğu görülmektedir

(Bollen & Long, 1993; Maruyama, 1998). Yapının (modelin) doğruluğu analiz sonucu elde edilen uyum iyiliği istatistiklerine bağlıdır (Schumacker & Lomax, 2004; Hair, Anderson, Babin, Black & Tahtam, 2006;). Yapının doğruluğu, Ki-kare (χ^2) istatistiği, Yaklaşık Hataların Ortalama Karekökü (Root Mean Square Error of Approximation - RMSEA), Uyum İyiliği İndeksi (Goodness of Fit Index - GFI), Karşılaştırmalı Uyum İndeksi (Comparative Fit Index - CFI), Normlaştırılmış Uyum İndeksi (Normed Fit Index - NFI), Normlaştırılmamış Uyum İndeksi (Non-Normed Fit Index – NNFI), Düzeltilmiş İyilik Uyum İndeksi (Adjusted Goodness of Fit Index - AGFI), Artmalı Uyum İndeksi (Incremantel Fit Index - IFI) gibi indeksler ile değerlendirilmektedir (Byrne, 1994; Kline, 2000; Hair vd., 2006; Tabachnick & Fidel, 2007).

Yerel bağımsızlık varsayımının sınanması için AFA sonuçlarıyla birlikte artık korelasyon matrisi (residual correlation matrix) oluşturulmuş ve incelenmiştir.

MTK doğrultusunda model veri uyumunun testi için maddelere ilişkin madde parametreleri ve bireylere ilişkin yetenek parametreleri BILOG-MG programında kestirilmiştir. Model veri uyumunu sınamada -2loglikelihood uyum istatistiği ve ki-kare istatistiğinden yararlanılmıştır.

Madde ve yetenek parametrelerinin değişmezliğini ortaya koymak amacıyla farklı madde gruplarında ve farklı yetenek gruplarında parametreler BILOG- MG programı ile kestirilmiş ve aralarındaki ilişkiye "Pearson Momentler Çarpımı Korelâsyon Katsayısı" kullanılarak SPSS programı aracılığı ile bakılmıştır.

Kullanılan teste ilişkin güvenirliğin belirlenmesinde, iç tutarlılığının göstergesi olan Kuder-Richardson 20 (KR–20) güvenirlik katsayısı kullanılmış ve katsayı EXCEL Programı ile hesaplanmıştır.

Araştırma sorularına yanıt aranırken, kestirilen yetenek parametreleri arasındaki ilişkilere "Pearson Momentler Çarpımı Korelâsyon Katsayısı" ve "Sınıf İçi Korelasyon Katsayısı" ile bakılmıştır. Ayrıca yetenek kestirim yöntemleri arasındaki farklılığı değerlendirmek amacıyla "Farklılıkların Ortalama Karekökü (Root Mean Squared Difference- RMSD)" değerinden yararlanılmıştır.

Kullanılan teste ilişkin betimsel istatistikler ve verilerin MTK'ya uygunluğunun sınanmasına ilişkin analizler aşağıda sıra ile verilmiştir.

Testteki maddelere ait madde güçlük indeksleri ve ayırt edicilik indeksleri KTK'ya göre ITEMAN programı kullanılarak incelenmiş ve analiz sonucunda madde güçlük indeksleri ise 0.05 ile 0.89 arasında bulunmuştur. Madde güçlük indekslerinin ortalaması 0.32; standart sapmaları ise 0.23 olarak elde edilmiştir. Madde ayırt edicilik indeksleri 0.21 ile 0.77 arasında (81 madde için 0.40 üzeri) değişmektedir. Maddelere ait ayırtedicilik katsayılarının ortalaması 0.51; standart sapması 0.13 olarak saptanmıştır.

Tek boyutluluk

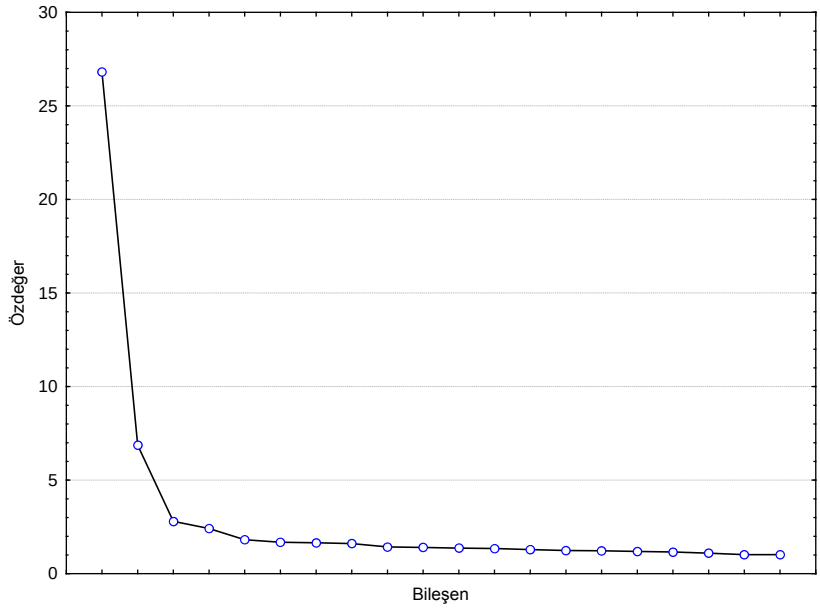
MTK'nın varsayımlarından biri olarak belirli bir maddeye doğru cevap verme olasılığının, kişinin tek bir karakteristiği veya yetenek düzeyiyle belirlendiğini açıklayan tek boyutluluk varsayımına bu araştırmada "Lumsden Yöntemi" olarak da bilinen açımlayıcı faktör analizi aracılığıyla bakılmıştır (Hambleton, Swaminathan & Rogers, 1991). Açımlayıcı faktör analizi sonuçlarını doğrulamak üzere de yine tek boyutluluk varsayımı için doğrulayıcı faktör analizi kullanılmıştır.

Açımlayıcı faktör analizi

Açımlayıcı faktör analizi ile tek boyutluluk incelenirken özdeğer, açıklanan varyans ve maddelere ilişkin faktör yük değerleri incelenmiştir. Alanyazında AFA sonucunda başat bir faktörün elde edilmesi tek boyutluluk göstergesi olarak kabul edilmektedir (Crocker &Algina, 1986; Hambleton, Swaminathan & Rogers, 1991). AFA'da, başat bir faktöre ait açıklanan varyans değerinin %30'dan büyük olması, yamaç birikinti grafiğinde bileşenlerin ivmelerine göre farkların anlamsız hale gelmesi (Hutcheson & Sofroniou, 1999), özdeğerler arasındaki farkın 1/3 oranından büyük olması gibi kriterler dikkate alınarak tek boyutluluğa karar verilmiştir. Araştırma kapsamında ele alınan teste ilişkin açıklanan varyanslar Tablo 2'de, yamaç birikinti grafiği Şekil 2'te verilmiştir.

Tablo 2. *Bileşenlere ilişkin özdeğer ve varyanslar*

	Başlangıç Özdeğerleri			
Bileşenler	*Özdeğer*	*Açıklanan Varyans (%)*	*Özdeğer (Kümülâtif) (%)*	*Açıklanan Varyans (Kümülâtif) (%)*
1	26.809	33.512	26.809	33.512
2	6.860	8.574	33.669	42.086
3	2.796	3.495	36.465	45.581
4	2.411	3.014	38.876	48.595
5	1.816	2.270	40.692	50.865
6	1.680	2.100	42.372	52.965
…	…	…	…	…



Şekil 2. *Yamaç birikinti grafiği*

Tablo 2 ve Şekil 2 incelendiğinde birinci faktöre ait özdeğerin 26.809 olduğu ve diğer faktörlere ait özdeğerlerden ayrıldığı göze çarpmaktadır. Birinci faktör ve ikinci faktöre ait özdeğerler arasındaki farkın neredeyse ikinci faktöre ait özdeğerin üç katı olduğu görülmektedir. Tek faktörün başat bir şekilde toplam varyansın % 33.512'sini açıkladığı da göz önüne alınarak testin tek boyutlu olduğu saptanmıştır. Her bir maddenin tek boyut altındaki faktör yük değerleri belirlenerek maddelerin ölçülen özelliğe ait varyansın ne

kadarını açıkladıkları incelenmiştir. Faktör yük değerleri 0.40'ın altında olan 20 madde testten çıkartılmıştır. 80 maddeye ilişkin faktör yük değerleri Tablo 3'te verilmiştir.

Tablo 3. *Maddelere ilişkin faktör yük değerleri*

Madde	*Faktör Yük değeri*	*Madde*	*Faktör Yük değeri*	*Madde*	*Faktör Yük değeri*	*Madde*	*Faktör Yük değeri*
M1	0.445	*M26*	0.518	*M50*	0.440	*M70*	0.687
M4	0.533	*M27*	0.559	*M51*	0.752	*M71*	0.640
M5	0.480	*M29*	0.684	*M52*	0.559	*M72*	0.726
M6	0.491	*M30*	0.542	*M53*	0.649	*M73*	0.666
M7	0.572	*M31*	0.494	*M54*	0.605	*M74*	0.421
M9	0.605	*M32*	0.497	*M55*	0.779	*M75*	0.629
M10	0.497	*M33*	0.531	*M56*	0.499	*M76*	0.443
M11	0.436	*M34*	0.696	*M57*	0.515	*M77*	0.513
M12	0.517	*M35*	0.737	*M58*	0.537	*M81*	0.434
M13	0.634	*M36*	0.672	*M59*	0.637	*M84*	0.564
M14	0.657	*M37*	0.641	*M60*	0.633	*M86*	0.567
M15	0.643	*M39*	0.424	*M61*	0.541	*M87*	0.474
M17	0.625	*M40*	0.552	*M62*	0.629	*M88*	0.490
M19	0.571	*M42*	0.631	*M63*	0.458	*M89*	0.537
M20	0.532	*M43*	0.584	*M64*	0.712	*M90*	0.425
M21	0.618	*M44*	0.700	*M65*	0.655	*M92*	0.454
M22	0.500	*M45*	0.723	*M66*	0.649	*M94*	0.566
M23	0.528	*M47*	0.637	*M67*	0.620	*M97*	0.733
M24	0.527	*M48*	0.716	*M68*	0.419	*M99*	0.404
M25	0.507	*M49*	0.471	*M69*	0.576	*M100*	0.536
Açıklanan Varyans= 33.512							

Tablo 3'teki faktör yük değerleri incelendiğinde 59 maddenin faktör yük değerinin 0.50'nin üzerinde olduğu görülmektedir. 80 maddelik test için madde ayırtedicilik indeksleri yeniden hesaplanmış ve madde ayırtedicilik indekslerinin 0.36 ile 0.77 arasında değiştiği saptanmıştır.

Doğrulayıcı faktör analizi

Açımlayıcı faktör analizi sonucu elde edilen tek faktörlü yapının doğruluğunu test etmek amacıyla doğrulayıcı faktör analizi yapılmıştır. Yapının doğruluğunun göstergelerinden biri olan Ki-Kare değeri 8732.30 bulunmuştur ($p<.01$). Ki-kare değerinin model uyumu için önemli bir ölçüt olan serbestlik derecesine (sd) oranı da sıklıkla değerlendirilmektedir. Kelloway (1996), χ^2/sd oranının 5'ten küçük olmasını iyi uyumun göstergesi olarak yorumlamaktadır (Haşlaman, 2005). Analiz sonucunda χ^2/sd oranı (8732.3/3080) 2.84 olarak tespit edilmiştir. Diğer model uyum indekslerine bakıldığında 80 maddenin tek faktör altında toplandığı yapıda Ortalama Hataların Karekökü (RMSEA) 0.043 olarak tespit edilmiştir. Gizil değişkenler arasında ilişkinin olmadığını öngören modelin ürettiği kovaryans matrisi ile önerilen modelin ürettiği kovaryans matrisini karşılaştıran ve modelin değerlendirilmesinde örneklem büyüklüğünü ve modeldeki serbestlik derecesini dikkate alan uyum testi olan Karşılaştırmalı Uyum İndeksi (CFI) ve CFI gibi örneklem büyüklüğünü ve modeldeki serbestlik derecesini dikkate alarak değer üreten bir başka uyum indeksi olan Artmalı Uyum İndeksi (IFI) 0.96 bulunmuştur. Bentler tarafından CFI'ya benzer

bir mantıkla geliştirilmiş olan Normlaştırılmış Uyum İndeksi (NFI) 0.92; modelin karmaşıklığını dikkate alarak değer üreten ve serbestlik derecesini dikkate alan Normlaştırılmamış Uyum İndeksi (NNFI) 0.96 olarak bulunmuştur. 0.90 ve üzeri iyi, 0.95 üzeri mükemmel uyum olarak kabul edilir. Bu çalışma kapsamında elde edilen NFI değeri Schermelleh-Engel ve Moosbrugger'in (2003) belirlediği ölçütler doğrultusunda mükemmel olarak kabul edilebilir. NNFI, CFI ve IFI indekslerinin 0. 95'in üzerinde olması mükemmel uyuma karşılık gelmektedir (Schermelleh-Engel & Moosbrugger, 2003). Çalışmada ortaya konan model için bulunan 0.96 değerleri modelin uygunluğu için 'mükemmel uyum' olarak değerlendirilebilir. Ki-Kare uyum testine alternatif olarak gösterilebilen (Sümer, 2000) Uyum İyiliği İndeksi (GFI) 0.82 olarak tespit edilmiştir. Örneklem genişliği dikkate alınarak düzeltilmiş bir GFI değeri veren Düzeltilmiş İyilik Uyum İndeksi (AGFI) değeri 0.81 olarak bulunmuştur. Tek boyutlu yapının doğruluğuna ilişkin elde edilen uyum indeksleri incelendiğinde değerlerin kabul edilebilir düzeyde olduğu görülmektedir (Jöroskog & Sörbom, 2001; Schermelleh-Engel & Moosbrugger, 2003).

Yerel bağımsızlık

Test maddelerinin birbirinden bağımsız olmasını ifade eden yerel bağımsızlık varsayımı, araştırmada tek boyutluluğun sağlanmasıyla ilişkilendirilerek değerlendirilmiştir. Alanyazında sıkça belirtildiği gibi yerel bağımsızlık varsayımı, tek boyutluluk ile paraleldir (Hambleton & Swaminathan, 1989; Hambleton, Swaminathan & Rogers, 1991). Tek boyutluluğun, yerel bağımsızlığa ilişkin bir kanıt olarak kabul edilebilmesinin yanında, artık korelasyon matrisi'nin incelenmesi ve tüm ikili çiftlere ait korelasyon katsayılarının düşük bulunması da yerel bağımsızlığın bir göstergesi olarak kabul edilmektedir (Embretson & Reise, 2000). Faktör analizlerinden sonra elde edilen 80 maddeye ilişkin artık korelasyon matrisi oluşturulmuş ve ikililere ait korelasyon katsayıları -0.19 ile 0.13 arasında bulunmuştur (İkili çiftlerin 491'inde (% 15) korelasyon katsayısı mutlak olarak 0.05'ten yüksek çıkmıştır). Artık değerlere ilişkin korelasyon katsayılarının sıfıra yakın çıkması ve tek boyutluluk varsayımının sağlanmasından dolayı yerel bağımsızlık varsayımının da karşılandığı kabul edilmiştir.

Model veri uyumu

Model veri uyumu için, her model bazında ayrı ayrı maddelerin ki-kare değerleri hesaplanmış ve modele uyum gösteren maddeler belirlenmiştir. Ki-kare değerleri AFA ve DFA sonunda belirlenen 80 madde ile BILOG-MG programı kullanılarak hesaplanmıştır.

Maddelere ilişkin ki-kare hesaplaması sonucunda; bir parametreli model için 39 maddenin modele uyum göstermediği, iki parametreli model için 8 maddenin modele uyum göstermediği, üç parametreli model için 18 maddenin modele uyum göstermediği saptanmıştır. Ki-Kare değerlerine bakılarak en çok veri uyumunu sağlayan modelin iki parametreli model olduğu anlaşılmaktadır. Bir parametreli modelde maddelerin %51.25'i; iki parametreli modelde maddelerin %90.00'ı; üç parametreli modelde maddelerin %77.50'si model veri uyumunu sağlamıştır. Ayrıca -2 Log Likelihood değerlerine bakıldığında en küçük -2 Log Likelihood değerine sahip olan üç parametreli model ile iki parametreli model arasındaki farkın çok fazla olmadığı (1PL için -2 Log Likelihood= 73878.647; 2PL için -2 Log Likelihood=73162.405; 3PL için -2 Log Likelihood=72976.339) dikkate alınarak iki parametreli model tercih edilmiştir (Hambleton, Swaminathan & Rogers, 1991).

İki parametreli modele uyum gösteren kalan 72 madde için ile madde parametreleri kestirilmiş ve maddelere ait madde güçlük parametresi (b) -1.86 ile 3.99 arasında; madde

ayırtedicilik parametresi (a) ise 0.39 ile 1.10 arasında değişmiştir. Madde güçlük parametreleri ortalaması 0.90; madde ayırtedicilik parametreleri ortalaması ise 0.68 olarak bulunmuştur. 72 maddeye ilişkin -2 Log Likelihood değeri 65668.769 olarak elde edilmiştir.

İki parametreli model ile uyum veren 72 madde doğrultusunda uygulama yapılan 994 öğrenciye ait yetenek parametreleri kestirilmiş ve yetenek parametreleri -2.530 ile 3.978arasında bulunmuştur. Yetenek parametrelerine ilişkin standart hatalar, 0.038 ile 0.456 arasında saptanmıştır.

Madde parametrelerinin değişmezliği

Madde parametrelerinin değişmezliği, aynı maddelerin farklı bireylere uygulanmasıyla benzer değerler elde edilmesidir. Madde parametrelerinin değişmezliğini ortaya koymak için iki farklı çalışma yapılmıştır. Değişmezliğin kanıtlanmasında yapılan ilk çalışma, yetenek düzeyi yüksek ve yetenek düzeyi düşük olan öğrencilerden madde parametreleri elde edilmesi şeklinde olmuştur. Yetenekleri kestirilen 994 öğrenci yetenek düzeylerine göre yüksekten düşüğe sıralanmış ve 497 öğrenci üst yetenek grubunda, 497 öğrenci alt yetenek grubunda kalacak şekilde ikiye bölüme ayrılmıştır. Üst ve alt yetenek düzeyindeki gruplardan 360'ar öğrenci alınarak iki farklı yetenek düzeyindeki grup için madde parametreleri kestirilmiştir.

İki farklı gruba ilişkin kestirilen madde parametrelerinin arasındaki ilişkiye gruplara ait çarpıklık ve basıklık katsayıları incelenerek "Pearson Momentler Çarpımı Korelâsyon Katsayısı" belirlenerek bakılmıştır. Alt ve üst gruba ait madde ayırtedicilik parametreleri arasındaki korelasyon 0.55 ($p<0.01$) olarak elde edilmiştir. Madde güçlük parametreleri arasındaki korelasyon ise 0.82 ($p<0.01$) olarak saptanmıştır.

Madde parametrelerinin değişmezliğinin incelenmesinde bir başka çalışma tesadüfî olarak oluşturulmuş 300'er kişilik üç grup ile yapılmıştır. Çalışma grubunu oluşturan 994 öğrenci birbiri ile kesişmeyecek şekilde tesadüfî olarak üçe ayrılmıştır (300'er kişi olmak kaydıyla). Grupları oluşturmak için EXCEL programı ile tesadüfî sayılar üretilmiş ve öğrenciler oluşturulan tesadüfî sayılar aracılığıyla üç gruba -bir öğrenci sadece bir grupta yer alacak şekilde- atanmışlardır. Bu üç grup için madde parametreleri kestirilmiş ve yine aynı şekilde aralarındaki korelâsyona "Pearson Momentler Çarpımı Korelâsyon Katsayısı" ile bakılmıştır. Madde parametreleri arasındaki korelâsyonlar Tablo 4'te verilmiştir.

Tablo 4. *Tesadüfi üç gruba ait madde parametreleri arasındaki korelasyonlar*

Gruplar	*a parametresi*	*a parametresine ait standart hata*	*b parametresi*	*b parametresine ait standart hata*
Grup1 – Grup2	0.74*	0.68*	0.96*	0.76*
Grup1 – Grup3	0.70*	0.56*	0.96*	0.80*
Grup2 – Grup3	0.60*	0.64*	0.96*	0.74*

*p<0.01

Tablo 4'te görüldüğü gibi, tesadüfi olarak oluşturulan üç gruptan kestirilen madde ayırtedicilik parametreleri arasındaki korelasyonlar 0.60 ile 0.74 arasında değişmektedir. Üç gruba ilişkin madde güçlük parametreleri arasındaki korelasyonlar ise 0.96 ve 0.96 olarak elde edilmiştir.

Yetenek düzeyi farklı gruplardan kestirilen madde parametrelerine ait korelasyonlar incelendiğinde; gerek yetenek düzeyi düşük ve yüksek alt-üst gruptan, gerekse tesadüfi olarak oluşturulmuş üç gruptan elde edilen korelasyonların kabul edilebilir düzeyde

oldukları görülmektedir. Kestirilen madde güçlük parametrelerine ait elde edilen yüksek korelasyonlar, farklı yetenek gruplarındaki bireyler için madde güçlük parametrelerinin oldukça benzer olduğunu göstermektedir. Madde ayırtedicilik parametrelerine ilişkin korelasyonların madde güçlük parametrelerine göre düşük çıkması; puan dağılımının normalliğinden madde güçlük parametresinin, madde ayırtedicilik parametresine göre daha az etkilenmesiyle açıklanabilir (Kelecioğlu, 2001). Alanyazında madde güçlük ve ayırtedicilik parametrelerine ilişkin korelasyon incelemelerinde benzer bulgulara rastlanmaktadır (Fan, 1998; Gelbal, 1994; Somer 1998). Bununla birlikte madde ayırtedicilik parametreleri arasında 0.01 düzeyinde manidar çıkan korelâsyonlar kestirilen parametreler arasında orta düzey bir ilişki olduğunu göstermektedir. Hesaplanan korelasyonlar veri setinin, MTK'nın bir özelliği olan madde değişmezliğini taşıdığını göstermektedir.

Yetenek parametrelerinin değişmezliği

Bireylere ait yetenek parametrelerinin madde setinden bağımsızlığını incelemek için tesadüfi olarak iki farklı madde seti oluşturulmuştur. Çalışmada kullanılan 72 maddeden tesadüfi olarak 30'ar maddelik iki farklı madde seti seçilmiştir. Birinci madde setinde 1-7-12-23-26-35-37-39-44-47-51-52-56-57-58-59-62-63-64-65-67- 69-70-72-73-74-75-81-88-94 numaralı maddeler; ikinci madde setinde 4-6-9-10-11- 14-15-17-19-21-24-27-29-30-31-33-45-48-49-50-53-54-55-61-66-68-71-84-87-90 numaralı maddeler yer almaktadır. İki farklı madde seti ile 994 öğrenciye ait yetenek parametreleri iki parametreli lojistik modele göre yeniden kestirilmiş ve testin tamamı da dâhil olmak üzere kestirilen yetenek parametreleri arasındaki korelâsyonlara bakılmıştır. Kestirilen yetenek parametreleri arasındaki ilişkiye "Pearson Momentler Çarpımı Korelâsyon Katsayısı" ile bakılmıştır. Hesaplanan korelasyon katsayıları Tablo 5'te verilmiştir.

Tablo 5. *Farklı madde setlerine ait yetenek kestirimleri arasındaki korelasyonlar*

r	*Madde Seti1*	*Madde Seti2*	*Testin Tamamı*
Madde Seti1	1.00		
Madde Seti2	0.81*	1.00	
Testin Tamamı	0.92*	0.95*	1.00

* $p < 0.01$

Tablo 5'te görüldüğü gibi oluşturulmuş iki farklı madde seti ile kestirilen yetenek parametreleri arasında manidar pozitif yönde yüksek bir ilişki bulunmuştur ($r=0.81$; $p<0.01$). Bununla birlikte ek olarak 72 maddeden oluşan testin tamamı ile de iki farklı madde setinden elde edilmiş parametreler arasındaki korelâsyonlara bakılmış ve korelasyonlar 0.92 ile 0.95 bulunmuştur ($p<0.01$). Elde edilen yüksek korelasyonlar bireylere ait yetenek parametrelerinin madde setinden bağımsız olarak kestirilebildiğini göstermektedir.

Testin güvenirliği

Araştırmanın amacı doğrultusunda kâğıt kalem testi ve BOBUT olarak uygulaması yapılacak olan İngilizce Kelime Testi'nin KTK çerçevesindeki güvenirliği için iç tutarlılığının göstergesi olan Kuder-Richardson 20 (KR-20) güvenirlik katsayısı hesaplanmıştır. MTK varsayımları ve model veri uyumunu sağlayan, değişmezlik özelliğini taşıdığı saptanan 72 madde için hesaplanan KR-20 güvenirlik katsayısı 0.93 olarak bulunmuştur. BOBUT uygulamasında her madde için hesaplanan madde bilgi fonksiyonu ve sonrasında test bilgi fonksiyonu, uygulama esnasında ilgili yetenek düzeyi için en çok bilgiyi veren maddenin

kullanılmasını sağladığı için MTK çerçevesinde de teste ilişkin güvenirliğin kanıtı olmuştur.

BULGULAR

Bilgisayar Ortamında Bireye Uyarlanmış Test Uygulamasında Madde Sayısı Dağılımı Nasıldır?

Oluşturulan çevrim içi ortam kullanılarak yapılan BOBUT uygulamasına katılan öğrencilerin teste ilişkin yanıtları Tablo 6'da verilmiştir.

Tablo 6. *Öğrencilere ait BOBUT yanıtları*

Öğrenci No	*Yanıtlanan Soru Sayısı*	*Doğru Yanıt Sayısı*	*Yanlış Yanıt Sayısı*	*Öğrenci No*	*Yanıtlanan Soru Sayısı*	*Doğru Yanıt Sayısı*	*Yanlış Yanıt Sayısı*
1	12	6	6	*35*	10	4	6
2	14	9	5	*36*	19	2	17
3	19	16	3	*37*	10	6	4
4	20	13	7	*38*	10	4	6
5	17	11	6	*39*	17	9	8
6	17	11	6	*40*	10	4	6
7	12	8	4	*41*	15	9	6
8	23	17	6	*42*	14	9	5
9	13	7	6	*43*	11	4	7
10	12	7	5	*44*	16	11	5
11	18	12	6	*45*	9	5	4
12	18	12	6	*46*	12	4	8
13	22	15	7	*47*	10	4	6
14	23	16	7	*48*	12	5	7
15	13	8	5	*49*	12	7	5
16	15	4	11	*50*	10	4	6
17	21	13	8	*51*	14	5	9
18	11	4	7	*52*	21	13	8
19	15	9	6	*53*	18	12	6
20	22	15	7	*54*	22	15	7
21	21	11	10	*55*	18	12	6
22	10	4	6	*56*	20	14	6
23	10	3	7	*57*	13	9	4
24	20	14	6	*58*	22	20	2
25	23	17	6	*59*	17	10	7
26	14	9	5	*60*	11	9	2
27	20	16	4	*61*	13	8	5
28	19	15	4	*62*	11	5	6
29	18	11	7	*63*	13	7	6
30	16	8	8	*64*	18	12	6
31	24	16	8	*65*	21	15	6
32	21	13	8	*66*	13	9	4
33	18	12	6	*67*	15	9	6
34	13	8	5				

Tablo 6'da görüldüğü gibi öğrenciler, madde havuzunda bulunan 72 maddeden farklı sayıda madde ile karşılaşmışlardır. Yapılan BOBUT uygulamasında en az madde ile testi bitiren öğrenci toplam 9 madde ile karşılaşmıştır. En çok madde ile testi bitiren öğrenci ise

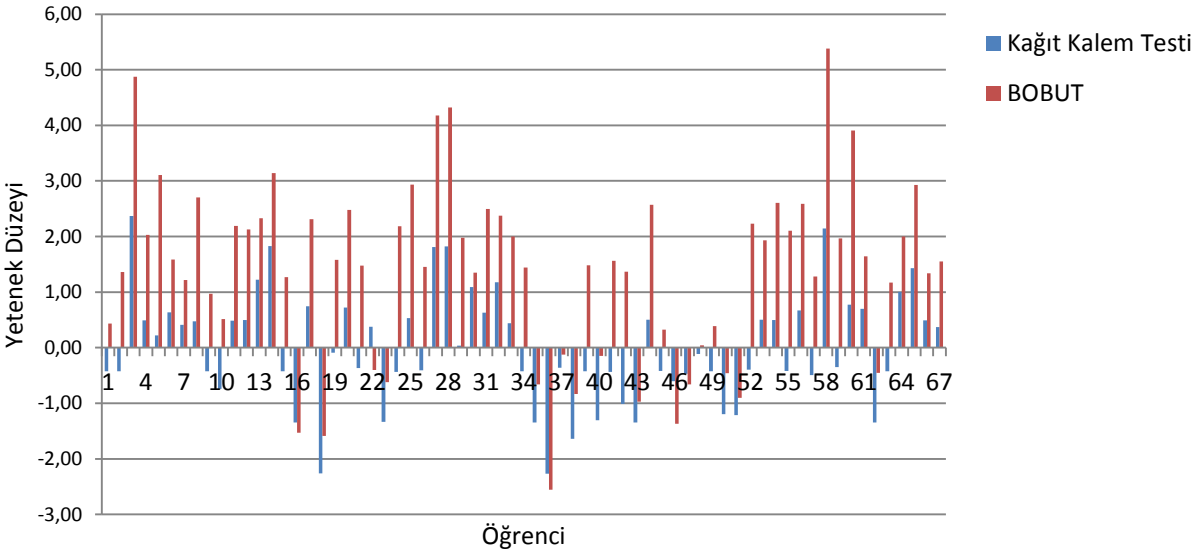
toplam 24 maddeye yanıt vermiştir. Elde edilen bulgular, yetenek kestirimi için 72 maddeden oluşan kâğıt kalem testine göre madde sayılarında % 66.67 ile % 87.50 arasında tasarruf sağlandığını göstermektedir. BOBUT uygulamasında öğrenciler, ortalama 16 (15.836) madde ile karşılaşmışlardır. Bu da testin ortalama % 78.01 oranında kısaldığını göstermektedir.

Öğrencilerin, araştırmada kullanılan BOBUT algoritması dolayısıyla en az 7 madde yanıtlamak zorunda kalmalarının madde sayılarında artışa sebebiyet verebileceği ihtimali göz önünde bulundurulmalıdır. İlk sorudan itibaren yeteneğin kestirildiği ve aynı sonlandırma kurallarının korunduğu bir durumda öğrencilerin karşılaştıkları madde sayılarının daha da düşeceği düşünülebilir. BOBUT uygulamalarında bireylerin yanıtladığı madde sayısının; sonlandırma kuralının katılığına, sabit madde kullanılıp kullanılmadığına bağlı olarak değişebileceği unutulmamalıdır.

Bilgisayar Ortamında Bireye Uyarlanmış Test İle Kâğıt Kalem Testi Uygulamalarında Kestirilen Yetenek Parametreleri Arasında Manidar Bir İlişki Var Mıdır?

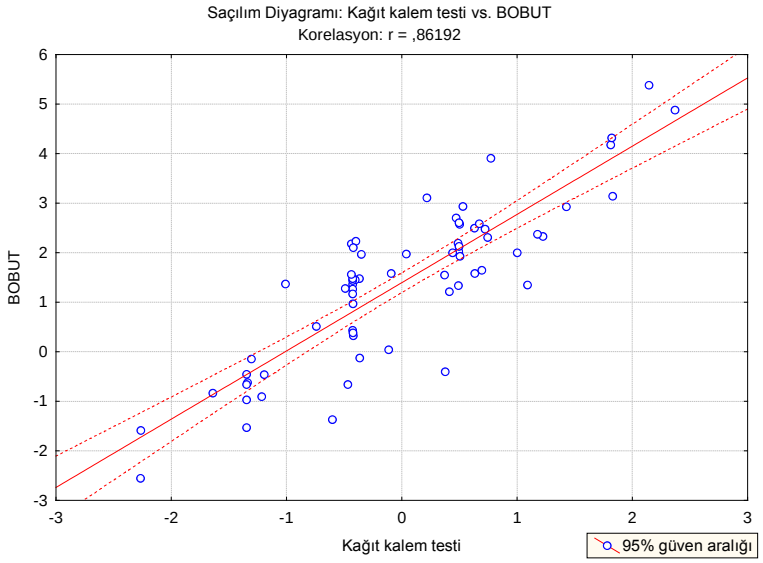
Öğrencilerin BOBUT uygulaması ve geleneksel kâğıt kalem testi uygulaması ile kestirilen yetenek parametreleri arasındaki ilişkiye bakmak için iki test farklı zamanlarda uygulanmıştır. Öncelikle kâğıt kalem testini alan öğrenciler daha sonra bilgisayar laboratuarlarında araştırma kapsamında oluşturulan çevrimiçi ortam aracılığı ile bilgisayar ortamında testi almışlardır. Kâğıt kalem testi uygulamasından elde edilen puanlar iki parametreli modele göre kalibre edilmiştir. Kâğıt kalem test uygulaması sonucu öğrencilerin kestirilen yetenek parametrelerinin -2.27 ile 2.37 arasında değiştiği görülmektedir. Yetenek kestirimlerine ilişkin aritmetik ortalama 0.00, standart sapma ise 0.99 olarak elde edilmiştir. Kestirimlere ilişkin standart hata puanları ise 0.005 ile 0.470 arasında değişmektedir.

Çevrimiçi ortam kullanılarak yapılan BOBUT uygulamasına ilişkinelde edilen bulgular incelendiğinde öğrencilerin en az 9, en fazla 24 madde ile karşılaştığı görülmektedir. BOBUT algoritmasında sonlandırma kuralı olarak belirlenen yetenek kestirimine ilişkin standart hatanın 0.50'den küçük olması ya da son iki standart hata arasındaki farkın 0.01'den küçük olması durumu doğrultusunda standart hatalar 0.42 ile 0.95 arasında değişmiştir. Öğrencilere ilişkin kestirilen yetenek parametreleri -2.55 ile 5.38 arasında değişmektedir. Yetenek parametrelerine ilişkin aritmetik ortalama 1.40, standart sapma 1.60 olarak saptanmıştır. İki uygulamayı da alan öğrencilerin yetenek parametreleri karşılaştırmalı olarak Şekil 3'te verilmiştir.



Şekil 3. *Kâğıt kalem testi ve BOBUT uygulaması sonuçları*

Şekil 3 incelendiğinde öğrencilerin BOBUT uygulamasındaki kestirilen yetenek parametrelerinin kâğıt kalem testi uygulamasına göre biraz daha yüksek olduğu görülmektedir. Ancak kestirimlerin birbirine benzer olduğu, çoğunlukla öğrencilerin yetenek parametrelerinin aynı paralellikte kestirildiği göze çarpmaktadır. Araştırmanın cevaplamaya çalıştığı ikinci soru çerçevesinde kâğıt kalem testi ve BOBUT uygulamalarında kestirilen yetenek parametreleri arasında manidar bir ilişki olup olmadığına "Pearson Momentler Çarpımı Korelâsyon Katsayısı" ile bakılmıştır. İki uygulamadaki yetenek parametrelerine ilişkin saçılım diyagramı Şekil 4'te verilmiştir.



Şekil 4. *Kâğıt kalem testi ve BOBUT uygulamalarından elde edilen yetenek parametrelerine ilişkin saçılım diyagramı*

Kağıt kalem testi ve BOBUT uygulamalarında kestirilen yetenek parametreleri arasında 0.86 düzeyinde pozitif yüksek korelasyon elde edilmiştir ($p<0.01$). İki uygulamaya

ait yetenek parametreleri arasındaki manidar yüksek korelasyon, uygulamaların benzer yetenek kestiriminde bulunduğunu göstermektedir.

Ayrıca kestirilen yetenek parametreleri arasında "Sınıf İçi Korelasyon Katsayısı (SKK)"na bakılmış ve iki uygulama sonucunda elde edilen yetenek parametreleri arasında korelasyon katsayısı 0.77 (%95 güven aralığında 0.66<SKK<0.85) olarak bulunmuştur ($p<0.01$).

Bilgisayar Ortamında Bireye Uyarlanmış Test Uygulamasında Simülatif Veriler Doğrultusunda Farklı Başlatma Ve Sonlandırma Kuralları İle Yetenek Kestirim Yöntemlerinin Farklılığına Göre Yetenek Parametreleri Arasında Manidar Bir İlişki Var Mıdır?

Araştırmanın alt amacı doğrultusunda alanyazında belirtilen farklı BOBUT stratejilerinde kestirilen yetenek parametrelerinin arasında bir ilişki olup olmadığı merak edilmiş ve bu doğrultuda BOBUT uygulamalarında kullanılan farklı stratejiler göz önüne alınarak farklı başlangıç kuralları, yetenek kestirim yöntemleri ve sonlandırma kuralları dâhilinde kestirilen yetenek parametreleri arasında ilişki olup olmadığına simülatif veriler kullanılarak bakılmıştır. Araştırmanın daha önceki bölümlerinde belirtildiği gibi, bireylere ait başlangıç yetenek düzeyleri 0 (sıfır) ve daha önceden kestirilmiş yetenek düzeyleri olacak şekilde iki farklı durum ele alınmıştır. Üç farklı yetenek kestirimi ile sabit uzunluk, SE<0.50 ve SE<0.30 sonlandırma kuralları ile birlikte 2x3x3 şeklinde toplam 18 farklı durum oluşturulmuştur (Tablo 7).

Tablo 7. *Simülatif BOBUT Stratejileri*

Durum	*Başlangıç Yenetek Düzeyi*	*Yetenek Kestirim Yöntemi*	*Sonlandırma Kuralı*
1.$\theta_{B=0,\ ML,\ SU}$	0	ML	Sabit Uzunluk
2.$\theta_{B=0,\ ML,\ SE<0.50}$	0	ML	SE<0.50
3.$\theta_{B=0,\ ML,\ SE<0.30}$	0	ML	SE<0.30
4.$\theta_{B=0,\ EAP,\ SU}$	0	EAP	Sabit Uzunluk
5.$\theta_{B=0,\ EAP,\ SE<0.50}$	0	EAP	SE<0.50
6.$\theta_{B=0,\ EAP,\ SE<0.30}$	0	EAP	SE<0.30
7.$\theta_{B=0,\ MAP,\ SU}$	0	MAP	Sabit Uzunluk
8.$\theta_{B=0,\ MAP,\ SE<0.50}$	0	MAP	SE<0.50
9.$\theta_{B=0,\ MAP,\ SE<0.30}$	0	MAP	SE<0.30
10.$\theta_{B=Kes,\ ML,\ SU}$	Önceden kestirilen	ML	Sabit Uzunluk
11.$\theta_{B=Kes,\ ML,\ SE<0.50}$	Önceden kestirilen	ML	SE<0.50
12.$\theta_{B=Kes,\ ML,\ SE<0.30}$	Önceden kestirilen	ML	SE<0.30
13.$\theta_{B=Kes,\ EAP,\ SU}$	Önceden kestirilen	EAP	Sabit Uzunluk
14.$\theta_{B=Kes,\ EAP,\ SE<0.50}$	Önceden kestirilen	EAP	SE<0.50
15.$\theta_{B=Kes,\ EAP,\ SE<0.30}$	Önceden kestirilen	EAP	SE<0.30
16.$\theta_{B=Kes,\ MAP,\ SU}$	Önceden kestirilen	MAP	Sabit Uzunluk
17.$\theta_{B=Kes,\ MAP,\ SE<0.50}$	Önceden kestirilen	MAP	SE<0.50
18.$\theta_{B=Kes,\ MAP,\ SE<0.30}$	Önceden kestirilen	MAP	SE<0.30
θ_{FULL}	Kâğıt Kalem Testi		

Her farklı durum için 2 parametreli lojistik model kullanılarak R İstatistik programında daha önceden kestirilmiş madde parametreleri doğrultusunda 994 kişilik simülatif yetenek parametresi örüntüsü üretilmiştir.

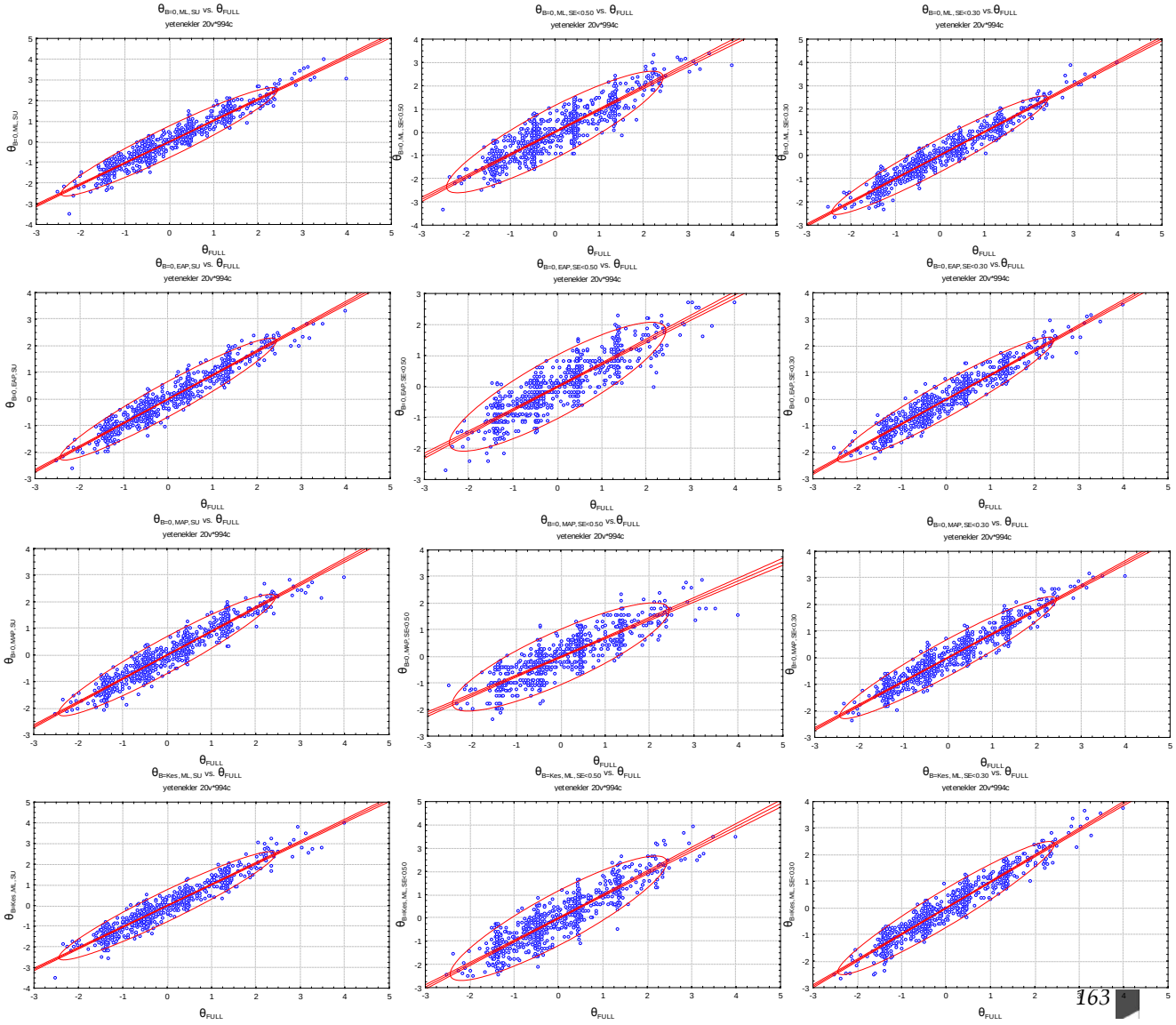
Farklı stratejiler ile kestirilen yetenek parametreleri kağıt kalem testinden kestirilen yetenek parametreleri arasında manidar bir ilişki olup olmadığına "Pearson Momentler Çarpımı Korelâsyon Katsayısı" ile bakılmış ve 18 farklı strateji ile kağıt-kalem testi sonucunda kestirilen yetenek parametreleri arasında 0.01 düzeyinde manidar korelasyon katsayıları elde edilmiştir (Tablo 8).

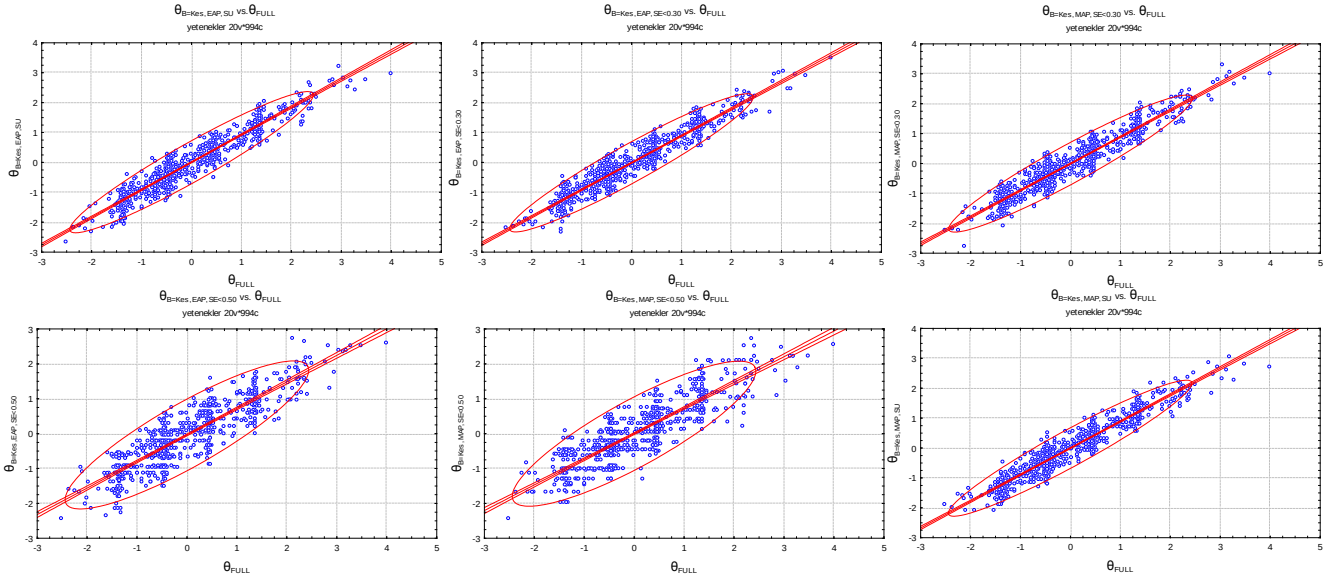
Tablo 8. *Farklı stratejiler kullanılarak elde edilmiş yetenek parametreleri ve kâğıt kalem testinden elde edilen yetenek parametreleri arasındaki pearson momentler çarpımı korelâsyon katsayıları*

Strateji	*r (_,θFULL)*	*Strateji*	*r (_,θFULL)*
$\theta_{B=0,\ ML,\ SU}$	0.96*	$\theta_{B=Kes,\ ML,\ SU}$	0.95*
$\theta_{B=0,\ ML,\ SE<0.50}$	0.90*	$\theta_{B=Kes,\ ML,\ SE<0.50}$	0.90*
$\theta_{B=0,\ ML,\ SE<0.30}$	0.96*	$\theta_{B=Kes,\ ML,\ SE<0.30}$	0.96*
$\theta_{B=0,\ EAP,\ SU}$	0.96*	$\theta_{B=Kes,\ EAP,\ SU}$	0.95*
$\theta_{B=0,\ EAP,\ SE<0.50}$	0.87*	$\theta_{B=Kes,\ EAP,\ SE<0.50}$	0.87*
$\theta_{B=0,\ EAP,\ SE<0.30}$	0.95*	$\theta_{B=Kes,\ EAP,\ SE<0.30}$	0.95*
$\theta_{B=0,\ MAP,\ SU}$	0.95*	$\theta_{B=Kes,\ MAP,\ SU}$	0.95*
$\theta_{B=0,\ MAP,\ SE<0.50}$	0.86*	$\theta_{B=Kes,\ MAP,\ SE<0.50}$	0.86*
$\theta_{B=0,\ MAP,\ SE<0.30}$	0.95*	$\theta_{B=Kes,\ MAP,\ SE<0.30}$	0.95*

*p<0.01

Farklı stratejiler ile kestirilmiş yetenek parametreleri ile kâğıt kalem testinden elde edilen yetenek parametrelerine ait saçılım diyagramları Şekil 5'te sunulmuştur.





Şekil 5. F*arklı stratejiler kullanılarak elde edilen yetenek parametreleri ve kâğıt kalem testinden elde edilen yetenek parametrelerine ait saçılım diyagramları*

Analiz sonucunda BOBUT uygulamasına ilişkin oluşturulan 18 farklı strateji ile kağıt kalem testinden elde edilen yetenek parametreleri arasındaki korelasyon katsayılarının 0. 86 ile 0.96 arasında değiştiği görülmektedir ($p<0.01$) (Tablo 8). Tüm stratejiler için pozitif yüksek düzeyde manidar korelasyonlar elde edilmiştir. En düşük korelasyon (0.858), başlangıç yetenek düzeyinin 0, sonlandırma kuralının SE<0.50 olarak alındığı MAP yetenek kestirimi yöntemine (B=0, MAP, SE<0.50) ait olduğu; en yüksek korelasyonun (0.959) ise, başlangıç yetenek düzeyinin 0, sonlandırma kuralının SE<0.30 olarak alındığı ML yetenek kestirimi yöntemine (B=0, ML, SE<0.30) ait olduğu saptanmıştır. Elde edilen korelasyon katsayıları, farklı BOBUT stratejileri kullanılarak kestirilen yetenek parametrelerinin öğrencilerin kağıt kalem testinden kestirilen yetenek parametrelerine çok benzer olduğunu göstermektedir.

Kağıt kalem testi ve 18 farklı stratejiden elde edilen yetenek parametreleri arasında ayrıca "Sınıf İçi Korelasyon Katsayısı"na bakılmış ve elde edilen korelasyon katsayıları Tablo 9'da verilmiştir.

Tablo 9'dagörüldüğü gibi hesaplanan sınıf içi korelasyon katsayılarında; en düşük korelasyon 0.84, en yüksek korelasyon ise 0.96 olarak elde edilmiştir. En düşük ve en yüksek korelâsyonun Pearson Momentler Çarpımı Korelâsyonun da olduğu gibi sırayla; başlangıç yetenek düzeyinin 0, sonlandırma kuralının SE<0.50 olarak alındığı MAP yetenek kestirimi yöntemine (0.844) ve başlangıç yetenek düzeyinin 0, sonlandırma kuralının SE<0.30 olarak alındığı ML yetenek kestirimi yöntemine (0.958) ait olduğu saptanmıştır.

Tablo 9. *Farklı stratejiler kullanılarak elde edilmiş yetenek parametreleri ve kâğıt kalem testinden elde edilmiş yetenek parametreleri arasındaki sınıf içi korelâsyon katsayıları*

Yetenek Parametreleri	*Sınıf İçi Korelâsyon Katsayısı*	*% 95 Güven Aralığı*	
		Alt Sınır	*Üst Sınır*
θ_{FULL} - $\theta_{B=0, ML, SU}$	0.95*	0.948	0.959
θ_{FULL} - $\theta_{B=0, ML, SE<0.50}$	0.90*	0.886	0.910
θ_{FULL} - $\theta_{B=0, ML, SE<0.30}$	0.96*	0.953	0.963
θ_{FULL} - $\theta_{B=0, EAP, SU}$	0.95*	0.949	0.960
θ_{FULL} - $\theta_{B=0, EAP, SE<0.50}$	0.86*	0.842	0.874
θ_{FULL} - $\theta_{B=0, EAP, SE<0.30}$	0.95*	0.948	0.959
θ_{FULL} - $\theta_{B=0, MAP, SU}$	0.95*	0.944	0.956
θ_{FULL} - $\theta_{B=0, MAP, SE<0.50}$	0.84*	0.826	0.861
θ_{FULL} - $\theta_{B=0, MAP, SE<0.30}$	0.95*	0.944	0.956
θ_{FULL} - $\theta_{B=Kes, ML, SU}$	0.95*	0.947	0.958
θ_{FULL} - $\theta_{B=Kes, ML, SE<0.50}$	0.89*	0.881	0.906
θ_{FULL} - $\theta_{B=Kes, ML, SE<0.30}$	0.96*	0.950	0.961
θ_{FULL} - $\theta_{B=Kes, EAP, SU}$	0.95*	0.946	0.957
θ_{FULL} - $\theta_{B=Kes, EAP, SE<0.50}$	0.87*	0.850	0.881
θ_{FULL} - $\theta_{B=Kes, EAP, SE<0.30}$	0.95*	0.944	0.956
θ_{FULL} - $\theta_{B=Kes, MAP, SU}$	0.95*	0.946	0.958
θ_{FULL} - $\theta_{B=Kes, MAP, SE<0.50}$	0.85*	0.833	0.867
θ_{FULL} - $\theta_{B=Kes, MAP, SE<0.30}$	0.95*	0.942	0.955

*p<0.01

Farklı BOBUT stratejileri ile kestirilen yetenek parametrelerinin kendi aralarındaki korelasyonlara bakılmış ve elde edilen korelasyon katsayıları Tablo 10'da verilmiştir.

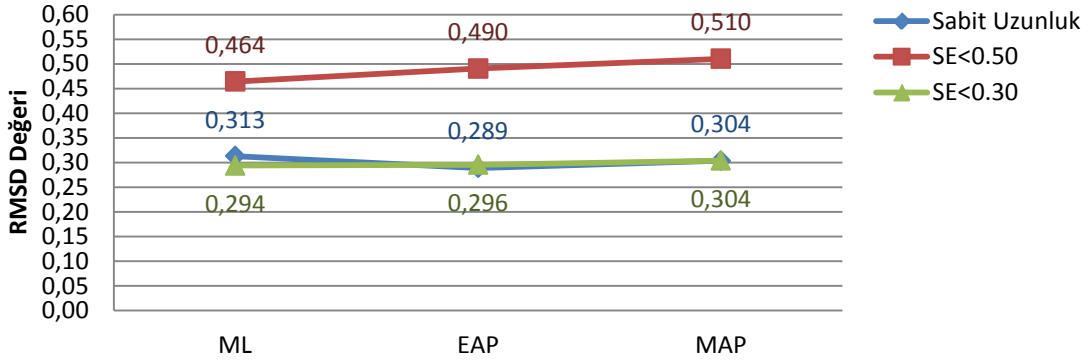
Tablo 10. *Farklı stratejiler kullanılarak kestirilen yetenek parametreleri arasındaki korelasyonlar*

*r**	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*
1	-																	
2	0.8	-																
3	0.9	0.8	-															
4	0.9	0.8	0.9	-														
5	0.8	0.7	0.8	0.8	-													
6	0.9	0.8	0.9	0.9	0.8	-												
7	0.9	0.8	0.9	0.9	0.8	0.9	-											
8	0.8	0.7	0.8	0.8	0.7	0.8	0.8	-										
9	0.9	0.8	0.9	0.9	0.8	0.9	0.9	0.8	-									
10	0.9	0.8	0.9	0.9	0.8	0.9	0.9	0.8	0.9	-								
11	0.8	0.8	0.8	0.8	0.7	0.8	0.8	0.7	0.8	0.8	-							
12	0.9	0.8	0.9	0.9	0.8	0.9	0.9	0.8	0.9	0.9	0.8	-						
13	0.9	0.8	0.9	0.9	0.8	0.9	0.9	0.8	0.9	0.9	0.8	0.9	-					
14	0.8	0.7	0.8	0.8	0.7	0.8	0.8	0.7	0.8	0.8	0.7	0.8	0.8	-				
15	0.9	0.8	0.9	0.9	0.8	0.9	0.9	0.8	0.9	0.9	0.8	0.9	0.9	0.8	-			
16	0.9	0.8	0.9	0.9	0.8	0.9	0.9	0.8	0.9	0.9	0.8	0.9	0.9	0.8	0.9	-		
17	0.8	0.7	0.8	0.8	0.7	0.8	0.8	0.7	0.8	0.8	0.7	0.8	0.8	0.7	0.8	0.8	-	
18	0.9	0.8	0.9	0.9	0.8	0.9	0.9	0.8	0.9	0.9	0.8	0.9	0.9	0.8	0.9	0.9	0.8	-

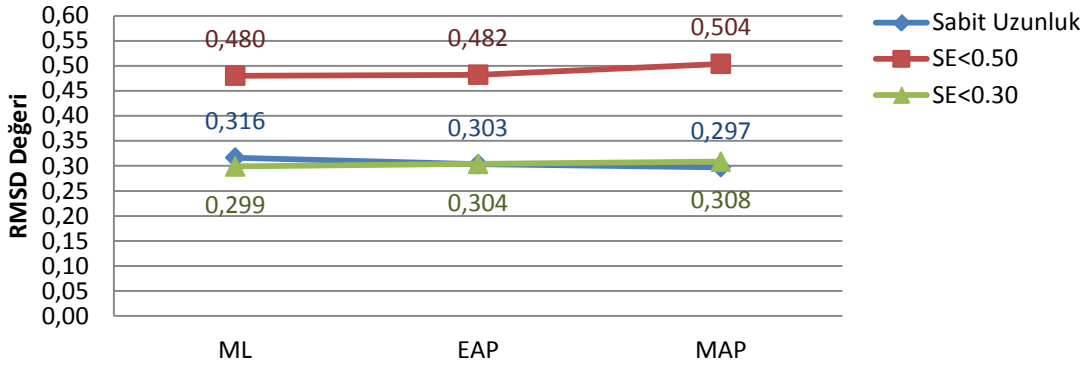
* p<0.01

Farklı BOBUT stratejileri ile kestirilen yetenek parametreleri arasında pozitif yönde yüksek düzeyde manidar korelasyon olduğu ve korelasyon katsayılarının 0.74 ile 0.92 arasında değiştiği saptanmıştır ($p<0.01$). Tablo 10 incelendiğinde görüleceği gibi, korelasyon katsayısının 0.80'nin altına düştüğü tüm durumlar sonlandırma kuralı olarak SE<0.50 alınması durumudur. Yetenek kestirimlerine ilişkin standart hatanın 0.50'nin altında olma kuralı toplam 153 farklı ikiliden 15'inde mevcuttur. 15 ikiliye ait korelasyon katsayılarının 0.74 ile 0.81 arasında değiştiği ve 14 ikilinin korelasyon katsayısı 0.80'nin altında olduğu görülmektedir.

En düşük korelasyon katsayılarına SE<0.50 sonlandırma kuralına sahip stratejilerin sahip olduğu bulgusu standart hata üst sınırının 0.50 olmasından dolayı gerçek puan aralığının geniş olmasına ve 0.30 durumuna göre daha düşük güvenirlikte kestirim yapılmasına dayandırılabilir. Benzer bir durum 0.50 standart hatayı temel alan stratejilerin kâğıt kalem testinden elde edilen yetenek parametreleri ile korelâsyonlarında da ortaya çıkmıştır. Tablo 10'da görüleceği gibi diğerlerine göre görece düşük korelasyon katsayıları SE<0.50 sonlandırma kuralına sahip stratejilerde hesaplanmıştır. Bu bulgular sonlandırma kuralının esnek olması halinde yetenek kestirimlerinin, katı kurallara göre bir ölçüde daha az keskinlikte belirlendiğini göstermektedir. Böylelikle geniş gerçek puan aralığına sahip kestirimlerin kendi aralarındaki korelasyon katsayıları da diğerlerine göre düşük çıkma eğilimi göstermiştir. Yetenek parametreleri arasındaki korelasyona ek olarak üç farklı yetenek kestirim yöntemi arasında fark olup olmadığına bakmak için yöntemlere ilişkin RMSD değerleri hesaplanmış ve Şekil 6 ile Şekil 7'de verilmiştir.



Şekil 6. *Başlangıç yetenek düzeyinin "0" alındığı stratejilere ilişkin RMSD değerleri*



Şekil 7. *Başlangıç yetenek düzeyinin daha önceden kestirildiği stratejilere ilişkin RMSD değerleri*

Kestirilen yetenek parametrelerine ilişkin farklılıkların karşılaştırılmasında kullanılan RMSD değerlerine bakıldığında, başlangıç yetenek düzeyi için 0 alındığı durumda RMSD değerleri farklarının 0.01 ile 0.046 arasında; başlangıç yetenek düzeyi için daha önceden kestirilmiş puanlar alındığında 0.002 ile 0.024 arasında değiştiği görülmektedir. Aradaki farkların küçüklüğü ML, EAP ve MAP yetenek kestirim yöntemleri için kestirilen yetenek parametrelerinde farklılık olmadığını göstermektedir.

SONUÇ TARTIŞMA VE ÖNERİLER

Bu araştırmada, madde tepki kuramının bir uygulaması olan bilgisayar ortamında bireye uyarlanmış test yöntemi ile kâğıt kalem test yönteminin karşılaştırılması ve bilgisayar ortamında bireye uyarlanmış test yöntemine ilişkin farklı stratejilerin karşılaştırılması amaçlanmıştır. Araştırma sonucunda, oluşturulan çevrimiçi ortam kullanılarak yapılan BOBUT uygulamasından elde edilen yetenek kestirimlerinde kâğıt kalem testine göre madde sayılarında büyük oranda (ortalama % 78) tasarruf sağlandığı görülmüştür. Bu sonuç BOBUT uygulamalarında daha az madde ile daha kısa bir test elde edildiğini göstermiştir. Test uzunluğundaki tasarruf, test süresini de paralel bir şekilde etkilemiş ve öğrenciler uygulamaya ilişkin ilgili testi daha kısa sürede bitirmişlerdir. Alanyazındaki ilgili araştırmalar incelendiğinde, madde sayısına ilişkin benzer sonuçlar elde edildiği görülmektedir. Yapılan araştırmalarda BOBUT uygulamalarında madde sayısında genel olarak % 50 ile % 70 arasında tasarruf sağlandığı görülmüştür (McBride & Martin, 1983; Olsen, Maynes, Slavvson & Ho, 1989; Kaptan, 1993; İşeri, 2002; McDonald, 2002; Scullard, 2007; Cömert, 2008; Öztuna, 2008; Kalender, 2011; Smits, Cuijper & Straten, 2011; Bulut & Kan, 2012). Daha az madde ile yetenek kestiriminde bulunabilen BOBUT uygulaması ile kağıt kalem testi uygulamasından elde edilen yetenek parmetreleri arasında pozitif yönde yüksek korelasyon katsayıları elde edilmiştir. Elde edilen yüksek korelasyon katsayıları BOBUT uygulaması sonucunda kağıt kalem testine benzer yetenek parametresi kestirimleri yapıldığını göstermiştir. Alanyazındaki BOBUT çalışmalarında, BOBUT yönteminin kâğıt kalem testleri yerine kullanılabilirliğini ortaya koymak adına, genellikle geleneksel kâğıt kalem testi ile BOBUT uygulamalarından elde edilen yetenek parametreleri arasındaki ilişkiye bakılmış ve yüksek korelasyonlar elde edilmiştir. Alanyazındaki çalışmalar incelendiğinde genellikle 0.74 ile 0.98 arasında yüksek korelasyon katsayılarının elde edildiği görülmektedir (Kaptan, 1993; İşeri, 2002; McDonald, 2002; Scullard, 2007; Cömert, 2008; Evans, 2010; Kalender, 2011; Kaskatı, 2011; Smits, Cuijper & Straten, 2011; Bulut & Kan, 2012; Wang vd., 2012; Zitny vd., 2012). Yapılan araştırmada iki uygulamadan elde edilen yetenek parametreleri arasındaki 0.86 düzeyindeki korelasyon katsayısı, alanyazındaki bulgulara paralel şekilde geleneksel kağıt kalem testi ve BOBUT uygulamalarının benzer sonuçlar verdiğini desteklemektedir.

BOBUT yönteminin sahip olduğu farklı başlatma kuralları, yetenek kestirim yöntemleri ve sonlandırma kuralları dikkate alınarak 18 farklı durum oluşturulmuş ve simülasyon çalışmaları ile oluşturulan bu farklı BOBUT stratejilerinden elde edilen yetenek parametresi kestirimleriyle kağıt kalem testi uygulamasından elde edilen yetenek parametreleri arasında bakılan ilişkide; farklı stratejiler ve kağıt kalem testinden elde edilen yetenek parametreleri arasında pozitif yönde yüksek korelasyon katsayıları bulunmuştur. Bu sonuç, BOBUT yöntemine ait -araştırma kapsamında ele alınan- iki başlangıç yetenek düzeyi, üç yetenek kestirim yöntemi ve üç sonlandırma kuralı kullanılarak oluşturulan durumlar ile, kâğıt kalem testinden kestirilen yetenek parametrelerine çok benzer yetenek parametrelerinin kestirildiğini göstermektedir. Yine benzer şekilde, oluşturulan 18 farklı

BOBUT stratejisi ile kestirilen yetenek parametreleri arasında pozitif yönde yüksek korelasyon katsayıları elde edilmiştir. Bu bulgu, farklı başlatma ve sonlandırma kullanılarak farklı yetenek kestirim yöntemi ile yetenek parametresi kestirildiğinde benzer sonuçlar alınabileceğini göstermektedir. Alanyazında simülatif karşılaştırmaların BOBUT çalışmalarında sıkça yapıldığı görülmektedir. McDonald (2002), Scullard (2007), Barrada vd. (2010), Evans (2010), Kalender (2011), Smits, Cuijper ve Straten (2011), Wang vd. (2012) gibi araştırmacılar yetenek parametrelerinin kestirimlerine ilişkin farklı durumlar için karşılaştırmalar yapmışlar ve araştırmalarında kestirilen yetenek parametreleri arasında yüksek korelâsyonlar bulmuşlardır.

Sonlandırma kuralı olarak belirlenen sabit uzunluk, yetenek kestirimine ilişkin standart hatanın 0.50'den küçük olması ve standart hatanın 0.30'dan küçük olması durumları göz önüne alındığında farklı BOBUT stratejilerinden elde edilen yetenek parametreleri arasındaki ilişkilerden en düşük ilişkilerin, sonlandırma kuralının standart hatanın 0.50'den küçük olması durumunda elde edildiği bulunmuştur. Gerek kağıt kalem testinden elde edilen yetenek parametreleri ve farklı BOBUT stratejilerden elde edilen yetenek parametreleri arasındaki ilişkilerde, gerekse farklı BOBUT stratejilerinden elde edilen yetenek parametrelerinin kendi aralarındaki ilişkilerinde aynı sonlandırma kuralı ile düşük ilişki elde edildiği gözlenmiştir. Bu düşük ilişki, sonlandırma kuralının esnek olmasından dolayı daha düşük güvenirlikte ve dolayısıyla daha geniş gerçek puan aralıklarında yetenek kestirimlerinin yapıldığına bağlanmıştır. Sonlandırma kuralının esnekliğinin yetenek kestiriminin keskinliğinde etkili olduğu sonucuna ulaşılmıştır. McDonald (2002) da, yaptığı bir araştırmada esnek sonlandırma kuralı ile daha yanlı ve daha büyük hataya sahip sonuçlar elde edildiğini saptamıştır.

Farklı BOBUT stratejileri kullanılarak kestirilen yetenek parametrelerine ilişkin yetenek kestirim yönteminden kaynaklı farklılık olup olmadığına RMSD değeri ile bakılmış ve ML, EAP ve MAP yetenek kestirim yöntemleri için kestirilen yetenek parametrelerinde farklılık olmadığı görülmüştür. Alanyazında benzer araştırmalarda; Barrada ve ark. (2010) simülasyon çalışmalarının sonunda yöntemler arasında fark bulmazken, Bulut ve Kan (2012) ile Wang ve ark. (2012) araştırmaları için söz konusu testlerde RMSD bulguları sonucunda EAP yöntemi lehine saptama yapmışlardır.

Araştırma sonuçları, BOBUT yöntemi ile elde edilen yetenek kestirimlerinin geleneksel kağıt kalem testinden elde edilen yetenek kestirimlerine çok benzer olduğunu ortaya koymuştur. Daha az madde ile daha kısa sürede, her birey için elde edilen hata kestirimi ve her madde için hesaplanan madde bilgi fonksiyonu ile daha geçerli ve güvenilir bir şekilde bireylerin yetenek düzeylerinin belirlenebildiği bu yöntemin kullanım alanlarının arttırılması ve ülke genelinde geniş ölçekli testlerde uygulanabilirliği düşünülmelidir. Bu doğrultuda yine çoklu puanlanan veriler için de BOBUT stratejilerinin daha iyi anlaşılması adına uygulamalı ve kuramsal araştırmalar yapılabilir.

KAYNAKÇA

Baker, F. B.& Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. New York: Marcel Bekker Inc.

Barrada, J. R., Olea, J., Ponsoda, V. & Abad, F. J. (2010). A method for the comparison of item selection rules in computerized adaptive testing. *Applied Psychological Measurement*, 34(6), 438–452.

Bejar, I. I. (1983). *Achievement testing: Recent advances*. London: Sage Publications.

Bollen, K. A. & Long, J. S. (1993). *Testing structural equation models.* London: Sage Publications.

Bulut, O. & Kan, A. (2012) Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Egitim Arastirmalari-Eurasian Journal of Educational Research*, 49, 61-80.

Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/Windows: Basic concepts, applications and programming.* California: Sage Publications, Inc.

Choi, S. W., Grady, M. W. & Dodd, B. G. (2011). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement*, *71* (1), 37-53.

Cömert, M. (2008). Bireye uyarlanmış bilgisayar destekli ölçme ve değerlendirme yazılımı geliştirilmesi. *Yayınlanmamış Yüksek Lisans Tezi.* Bahçeşehir Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Crocker, L. & Algina, J. (1986). *Introduction classical and modern test theory.* New York: Harcourt Brace Javonovich College Publishers.

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

Evans, J. J. (2010). Comparability of examinee proficiency scories on computer adaptive tests using real and simulated data. *Unpublished Doctoral Dissertation.* The State University of New Jersey, USA.

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

Folk, W. G. & Smith, R. L. (2002). Models for delivery of CBTs. Mills, C.N., Potenza, M.T., Fremer, J.J., and Ward, W.C. (Ed.). *Computer-based testing: Building the foundation for future assessments.*New Jersey: Lawrence Erlbaum Associates.

Gelbal, S. (1994). p madde güçlük indeksi ile Rasch modelinin b parametresi ve bunlara dayalı yetenek ölçülerine üzerine bir karşılaştırma. *Eğitim Fakültesi Dergisi*, 10, 85-94.

Georgiadou, E., Triantafillou, E. & Economides, A.A. (2006). Evaluation parameters for computer adaptive testing. *British Journal of Educational Techonology, 37* (2), 261-278.

Hair, J., Anderson, R., Babin, B., Black, W. & Tahtam, R. (2006). *Multivariate data analysis.* New Jersey, Prentice Hall, Inc.

Hambleton, R. K. & Swaminathan, H. (1989). *Item response teory: Principles and applications*. USA: Kluwer Nijhoff Publishing.

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

Haşlaman, T. (2005). Programlama dersi ile ilgili özdüzenleyici öğrenme stratejileri ile başarı arasındaki ilişkilerin incelenmesi: Bir yapısal eşitlik modeli. *Yayınlanmamış Yüksek Lisans Tezi.* Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

Hutcheson, G. & Sofroniou, N. (1999). *The multivariate social scientist.* California: SAGE Pub.

Iseri, A. I. (2002). Assessment of students' mathematics achievement through computer adaptive testing procedures. *Unpublished Doctoral Dissertation.* Middle East Technical University, Ankara.

Jöreskog, K. G. & Sörbom, D.(2001). *Lisrel 8: User's reference guide.* Chicago, IL: Scientific Software International, Inc.

Kalender, İ. (2011). Effects of different computerized adaptive testing strategies on recovery of ability. *Yayınlanmamış Doktora Tezi.* Middle East Technical University, Ankara.

Kaptan, F. (1993). Yetenek kestiriminde adaptive (bireyselleştirilmiş) test uygulaması ile geleneksel kâğıt-kalem testi uygulamasının karşılaştırılması. *Yayınlanmamış Doktora Tezi.* Hacettepe Üniversitesi, Ankara.

Kaskatı, O. T. (2011). Rasch modelleri kullanarak romatoid artirit hastaları özürlülük değerlendirimi için bilgisayar uyarlamalı test yöntemi geliştirilmesi. *Yayınlanmamış Doktora Tezi.* Ankara Üniversitesi Sağlık Bilimleri Enstitüsü, Ankara.

Kelecioğlu, H. (2001). Örtük özellikler teorisindeki b ve a parametreleri ile klasik test teorisindeki p ve r istatistikleri arasındaki ilişki. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 20, 104-110.

Kline, R. B. (2000). *Principles and practice of structural equation modeling.* NewYork: The Guilford Press.

Linden, W. J. & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. Linden, W.J. &Glas, C. A. W. (Ed.). *Elements of adaptive testing*. New York: Springer.

Lord, F. M. & Stocking, M. L. (1988). Item response theory. *Educational research, methodology, and measurement: An international handbook* (Edt: J. P. Keeves). New York: Pergamon Press.

Magis, D. & Raiche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software*, *48* (8) , 1-31.

Maruyama, G. M. (1998). *Basics of structural equation modeling*. California: Sage Publications, Inc.

McBride, J. R. & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military design. *New horizons in testing: Latent trait test theory and computerized adaptive testing* (Edt: D. J. Weiss). New York: Academic Press.

McDonald, P. L. (2002). Computer adaptive test for measuring personality factors using item response theory. *Unpublished Doctoral Dissertation.* The University Western of Ontario, London.

Olsen, J. B., Maynes, D. D., Slavvson, D. & Ho, K. (1989).Comparison of paper administered, computer administered and computerized adaptive achievement tests. *Journal of Educational Computing Research*, *5* (31), 311-326.

Öztuna, D. (2008). Kas-iskelet sistemi sorunlarının özürlülük değerlendiriminde bilgisayar uyarlamalı test yönteminin uygulanmas. *Yayınlanmamış Doktora Tezi.* Ankara Üniversitesi Sağlık Bilimleri Enstitüsü, Ankara.

Patton, J. M., Cheng, Y., Yuan, K. H. & Diao, Q. (2013). The influence of item calibration error on variable-length computerized adaptive testing. *Applied Psychological Measurement*, *37* (1), 24-40.

Rudner, L. M. (1998). *An on-line, Interactive computer adaptive testing mini tutorial.* http://edres.org/scripts/cat/catdemo.htm. Erişim Tarihi: 02.01.2011.

Schermelleh-Engel, K. & Moosbrugger, H. (2003). Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8* (2), 23-74.

Schumacker, R. E. & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling*. New Jersey: Lawrence Erlbaum Associates Publishers.

Scullard, M. G. (2007). Application of item response theory based computerized adaptive testing to the strong interest inventory. *Unpublished Doctoral Dissertation.* University of Minnesota, USA.

Segall, D. O. (2003). *Computerized adaptive testing. Encyclopedia of social measurement.* www.academicpress.com/refer/measure/. Erişim Tarihi: 02.12.2011.

Smits, N., Cuijpers, P. & Straten, A. (2011). Applying computerized adaptive testing to the CES-D Scale: A simulation study. *Psychiatry Research*, 188, 145-155.

Somer, O. (1998). Kişilik testlerinde klasik ve modern test kuramları ile madde analizi. *Türk Psikoloji Dergisi*, 13, 1–15.

Sukamolson, S. (2002). Computerized test/item banking and computerized adaptive testing for teachers and lecturers. http://www.stc.arts.chula.ac. th/ITUA/Papers_for_ITUA_Proceedings/Suphat2.pdf. Erişim Tarihi: 15.11.2011.

Sümer, N. (2000). Yapısal eşitlik modelleri: Temel kavramlar ve örnek uygulamalar. *Türk Psikoloji Yazıları*, *3* (6) 49-74.

Tabachnick, B.G. & Fidell L.S. (2007). *Using multivariate statistics.* USA, Pearson Education.

Tian, J., Miao, D., Zhu, X. & Gong, J. (2007). An introduction to the computerized adaptive testing. *US-China Education Rewiew*, *4* (1), 72-81.

Veldkamp, B. P. & Linden. W. J. (2010). Designing item pools for adaptive testing. *Elements of adaptive testing* (Eds: W. J. Linden. & C.A.W. Glas). New York: Springer.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J. Steinberg, L. & Thissen, D. (2000). *Computerized adaptive testing: a primer*. New Jersey: Lawrence Erlbaum Associates, Publishers.

Wang, H. P., Kuo, B. C., Tsai, Y. H. &Liao, C. H. (2012). A cefr-based computerized testing system for chinese proficiency. *TOJET: The Turkish Journal of Educational Technology*, *11* (4), 1-12.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53 (6),* 774-789.

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, *2* (1), 1-27.

Way, W. D. (2006). *Practical questions in introducing computerized adaptive testing for K-12 assessments*(Research Report). Pearson Educational Measurement. http://education.pearsonassessments.com/NR/rdonlyres/EC965AB8-EE70-46E5-B1A5036BE41AB899/0/RR_05_03.pdf?WT.mc_id=TMRSPractical_Questions_in_Introducing_Computerized. Erişim Tarihi: 08.06.2012.

Zitny, P., Halama, P., Jelinek, M. & Kveton, P. (2012). Validity of cognitive ability tests–comparison of computerized adaptive testing with paper and pencil and computer-based forms of administrations. *Studia Psychologica*, *54* (3), 181-194.

A Comparison of Computerized Adaptive Testing Strategies[4]

Fatih KEZER[5] & Nizamettin KOÇ[6]

Introduction

Computerized Adaptive Testing (CAT) is being used increasingly worldwide because of its numerous advantages. CAT provides a valid and reliable testing process based on its strong algorithm of Item Response Theory (IRT). In particular, detailed information about the quality of the questions through an item-characteristic curve as an outcome of IRT enables the efficient use of items to determine an individual's ability level quickly and accurately. Considering the existing assessment and evaluation approaches that are used, CAT provides a valid and reliable method for estimating the abilities of individuals through the invariance characteristics, eliminating the estimation limitation of ability parameters that depend on items and item parameters based on the group. This advantage of comparability of ability parameters from various tests facilitates test standardization. The basic function of CAT is to offer the most-appropriate items to an individual during a test and thus determine his or her ability level accurately (θ).

CAT has not yet been utilized extensively in Turkey due to limitations such as strong assumption fundamentals in line with its basic IRT, hardware and software requirements, and its need for a large pool of items. Centralized exams are practiced with traditional paper-and-pencil tests based on the current Classical Test Theory. Considering its advantages, the use of CAT should be expanded, particularly for large-scale tests. Therefore, there is a necessity to understand how CAT functions and to conduct research accordingly.

This study aims to compare the CAT method, which is based on individualized testing according to differences in ability levels, to traditional paper-and-pencil testing methods, and to compare various strategies of computerized adaptive testing within the framework of an English vocabulary test. In this regard, this study investigates the item number distribution of CAT applications, significant relationships between CAT and paper-and-pen test ability parameter estimations, significant relationships between ability parameters according to the various ability estimation methods, and different rules for starting and stopping in line with simulative data.

Method

The study is a basic research model, as it covers rules for starting and stopping CAT applications as well as comparisons of ability estimation methods.

Data were collected from 1,166 students studying at the preparatory class of Ankara University School of Foreign Languages during the 2012–2013 academic year. At the beginning of each academic year, the School of Foreign Languages administers a placement test and classifies students according to their proficiency levels as A1, A2, and B1; this study collected data from students in levels A1 and A2. Data were gathered in three different stages. In each stage, different students were examined. In the first stage, a pilot application with 105 students aimed to determine the quality and clarity of items. In the second stage,

[4] Bu makale, Fatih Kezer'in, Ankara Üniversitesi Eğitim Bilimleri Enstitüsü'nde, Prof. Dr. Nizamettin KOÇ danışmanlığında yapmış olduğu "Bilgisayar Ortamında Bireye Uyarlanmış Test Stratejilerinin Karşılaştırılması" (2013) adlı doktora tezinden üretilmiştir.

[5] PhD - Kocaeli University Education Faculty - fatihkezer@yahoo.com
[6] Prof. Dr. - Ankara University Faculty of Educational Sciences - nkoc@ankara.edu.tr

data from 994 students were used to determine the psychometric characteristics of items and to calibrate the IRT estimations of items for the CAT applications. In the third stage, data were collected from 67 students to compare the individualized computer test scores to paper-and-pencil tests using actual practice apart from the simulations.

The data-collection instrument was the English Vocabulary Test developed for the research. In addition, simulative data were produced using the open source statistical program R to compare the different strategies. The researcher created an online environment to conduct the computerized adaptive tests.

Eighteen (18) different situations were created to examine the various starting rules, ability estimation methods, and stopping rules used in the CAT applications. At the beginning of the test, ability scores of the individuals were set as 0 (zero) for one situation and based on previously estimated abilities for the other situation. For the ability estimation, three different current methods of IRT were considered in the form of different situations as ML, MAP, and EAP. The termination rules the study used are the situation of standard error being less than 0.50 as a frequently used termination rule in the CAT applications, the situation of standard error being less than 0.30, and the situation of fixed length.

Item statistics were then analyzed to determine item quality through pilot applications. Test items were changed and revised based upon the calculated item-difficulty index and item-discrimination index. The test that was developed was examined to see whether it meets the IRT assumptions; thus, exploratory and confirmatory factor analysis for unidimensionality and EFA results for local independence were used along with the residual correlation matrix. Item and ability parameters were estimated, and -2log-likelihood accordance statistics and chi-square statistics were utilized to test model data accordance. Parameters from different item groups and different ability groups were obtained to define item and ability parameter invariance, and their relationships were analyzed. Test reliability was achieved using the Kuder-Richardson Formula 20 (K–R 20) reliability coefficiency as a sign of internal reliability. Relationships between the estimated ability parameters were sought using the Pearson Product-Moment Correlation Coefficient and the Intraclass Correlation Coefficient.

Moreover, the value of Root Mean Squared Difference, or RMSD, was referred to in order to evaluate the difference between ability estimation methods.

Results, Discussion, and Suggestions

It was demonstrated that ability estimations derived from the CAT application of the created online environment saved a large percentage (78% average) of items compared to paper-and-pencil tests. This result shows that CAT application leads to developing shorter tests with fewer test items. The advantage in the shorter test length affected the test duration accordingly, and students completed their tests in less time. A highly significant positive correlation coefficiency was found between the CAT application with fewer items and paper-and-pencil tests for the ability estimation. These high correlation coefficiencies showed that ability parameter estimations similar to paper-and-pencil tests were achieved as a result of the CAT application. The correlation coefficiency of 0.86 obtained from the two applications is parallel to the findings in the literature, in that traditional paper-and-pencil tests and CAT applications revealed similar results.

Eighteen (18) different situations were created considering the different starting rules, estimation method, and stopping rules of the CAT method. The relationship between the ability parameter estimations of the CAT strategies developed by simulations and ability

parameters from the paper-and-pencil test applications revealed a positive high correlation coefficiency between ability parameters of different CAT strategies and paper-and-pencil tests. This finding demonstrates that situations based on using the two beginning ability levels, three ability estimation methods, and three termination rules of the CAT method estimate similar ability parameters to those of ability parameters based on paper-and-pencil tests. Similarly, a high positive correlation coefficiency was obtained among 18 different CAT strategies and estimated ability parameters. This finding demonstrates that when ability parameters were estimated through a different ability estimation method using different starting and stopping, similar results could be achieved. Considering the fixed length as a stopping rule, standard error of ability estimation was being less than 0.50, and standard error being less than 0.30, it is seen that the lowest relationship among the ability parameters obtained from different CAT strategies is possible when the termination rule standard error is less than 0.50. The same termination rule leads to a low-level relationship as in that between the ability parameters based on paper-and-pencil tests and the ability parameters from different CAT strategies and the relationships among different CAT strategies. This low level of relationship is linked to low-reliability level of the termination rule due to its flexibility and thus, ability estimations of large-scale actual scores. Ability parameters estimated using different CAT strategies were analyzed to determine the effect of the ability estimation method through the RMSD value, and it was found that there is no difference among the ability estimation parameters for the ability estimation methods of ML, EAP, and MAP.

The findings show that ability estimations of CAT methods are similar to those of traditional paper-and-pencil tests. CAT provides fewer items in a shorter period of time, error estimation for each individual, and information function calculated for each item; therefore, it determines ability levels of individuals in a more valid and reliable manner. In conclusion, its use needs to be expanded, and its applicability for large-scale tests nationwide should be considered. In this regard, practical and theoretical research should be conducted to comprehend the use of CAT strategies for multi-scored data.

Key Words: Computerized adaptive testing, Testing strategies, Item response theory

Atıf için / Please cite as:
Kezer, F. & Koç, N. (2014). Bilgisayar ortamında bireye uyarlanmış test stratejilerinin karşılaştırılması [A comparison of computerized adaptive testing strategies]. *Eğitim Bilimleri Araştırmaları Dergisi - Journal of Educational Sciences Research, 4* (1), 145-174. http://ebad-jesr.com/