ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2015; 6(2); 218-237

Bilgisayar Okuryazarlığı Testinin Bilgisayar Ortamında Bireye Uyarlanmış Test Olarak Geliştirilmesi*

Development of Computer Literacy Test as Computerized Adaptive Testing

Durmuş ÖZBAŞI ** Nükhet DEMİRTAŞLI ***

Öz
Bu araştırmanın amacı, Ankara Üniversitesi'nde tüm fakültelerde birinci sınıf öğrencilerine uygulanmakta olan Bilgi ve İletişim Teknolojileri dersi muafiyet sınavı testinin bilgisayar ortamında bireye uyarlanmış test (BOBUT) olarak uygulanabilirliğini araştırmaktır. Araştırma iki temel aşamada gerçekleştirilmiştir. İlk aşamada, hazırlanan maddeleri denemek ve simülatif BOBUT uygulamasında üzerinde çalışılacak verileri elde etmek üzere 1366 üniversite öğrencisiyle çalışılmıştır. İkinci aşamada ise, 142 üniversite birinci sınıf öğrencisiyle canlı (live) BOBUT ve kâğıt-kalem testi uygulaması gerçekleştirilmiştir. Araştırmada veri toplama aracı olarak Bilgisayar Okuryazarlık (BİLOKUR) testi kullanılmıştır. Araştırmada toplanan verilerin madde tepki kuramı'nın (MTK) 3 parametreli lojistik modeline uyum sağladığı tespit edilen 136 maddelik soru havuzu ile canlı BOBUT uygulaması yapılmıştır. Araştırmanın bulgularına göre, simülatif BOBUT uygulamasında kestirilen yeterlik kestirimlerine ilişkin en yüksek güvenirlik ölçüsü, sabit madde sayısına göre test sonlandırma koşulunda bulunmuştur. Ayrıca kullanılan madde sayısı bakımından, en az madde kullanımı, test sonlandırma koşulunun SH<0.50 olduğu durumda gerçekleşmiştir. En yüksek olabilirlik yöntemine (EYOY) göre yeterlik kestiriminin uygulandığı BOBUT testinde öğrencilerin yeterlik ölçüleri, kâğıt-kalem testinden elde edilenlere göre, özellikle uç değerlerde daha güvenilir ölçme sonuç vermiş, standart hata değeri açısından da BOBUT uygulamasıyla daha düşük hata kestirimleri elde edilmiştir. Canlı BOBUT uygulamasından elde edilen ortalama güvenirlik (test bilgi değeri), kâğıt-kalem testinden elde edilen güvenirlik değerinden daha yüksek bulunmuştur. Bu araştırmanın sonuçlarına göre, SH<0.30 test sonlandırma kuralı kullanıldığında EYOY; SH<0.50 ve sabit madde (30) test durdurma kuralı kullanıldığında ise, beklenen sonsal dağılıma (BSD) dayalı yeterlik kestirim değerlerinin daha güvenilir olduğu bulunmuştur. Ayrıca canlı BOBUT uygulamasında elde edilen test bilgi miktarı, kâğıt-kalem testinden elde edilen güvenirlikten anlamlı düzeyde yüksek bulunmuştur.

Anahtar Kelimeler: bilgisayar ortamında bireye uyarlanmış test, madde tepki kuramı, bilgisayar yeterlik sınavı.

Abstract
The purpose of this study is to investigate the applicability of the Computer Adaptive Testing (CAT) of the exemption exam of Information and Communication Technology in computer environment (as CAT) given to the first year students at every faculty of Ankara University each year. The research was carried out in two basic stages. In the first stage, the researchers studied with 1366 university students to obtain the data to study on with CAT application and and to test the prepared items. In the second stage, a paper and pencil test was given to 142 university first year students with live CAT. The test of computer literacy was used as an instrument of data collection. It was also tested in the study if the collected data met the hypothesis of Item

* Bu araştırma, Prof. Dr. Nükhet Demirtaşlı danışmanlığında Durmuş Özbaşı tarafından hazırlanan doktora tezinin bir bölümünden hazırlanmıştır..

**Arş. Gör. Dr. Durmuş ÖZBAŞI Çanakkale Onsekiz Mart Üniversitesi, Eğitim Fakültesi, Çanakkale-Türkiye, dozbasi@gmail.com
***Prof. Dr. Nükhet DEMİRTAŞLI Ankara Üniversitesi, Eğitim Bilimleri Fakültesi, Ankara-Türkiye, rnukhet@yahoo.com

Geliş Tarihi: 06.09.2015
Kabul Tarihi: 10.11.2015

Response Theory (IRT). With this regard, a live CAT application was carried out with 136-item pool which was found to comply with the three-parameter logistic model. According to the findings of the study, the highest reliability estimate found in simulative CAT application was found in test termination condition depending on the fixed item number (with 30 items). Besides, with regards to the number of used item, the least item use happened when test termination condition is Standard Error (SE)<0.50. In the CAT test, students' ability estimates in CAT in which proficiency estimate is done depending on Maximum Likelihood Estimation (MLE) has come up with more reliable results in extreme values compared to those obtained in paper-pencil test, and lower standard error estimates were obtained with the use of CAT application with regards to standard error value. The average reliability obtained from live CAT application was found to be higher that of paper-pencil test. According to the findings of this study, when the SE<0.30 test termination rule is applied, MLE was found to be more reliable, when SE<0.50 and fixed item (30) test termination rule was applied, the proficiency estimate value based on Expected A Posteriori Method (EAP) was found to be more reliable. Besides, the test information amount obtained in live CAT application was significantly higher than that of paper and pencil test.

Key Words: computerized adaptive testing (cat), item response theory (irt), computer literacy test, paper-pencil test

GİRİŞ

Günümüzde bilgisayar teknolojisinin gelişimine paralel olarak testler ve testlerin uygulanma yöntemleri de gelişmektedir. Özellikle, bilgisayar teknolojisinin gelişmesi ile birlikte son yirmi yıldır, eğitimde çeşitli amaçlarla (seçme, yerleştirme, teşhis, vb.) testler *bilgisayar ortamında bireye uyarlanan testler* (Computerized Adaptive Test) olarak kullanılmaktadır. Bilgisayar ortamında bireye uyarlanan test (BOBUT), psikometrik özellikleri daha önceden kestirilmiş bir madde havuzundaki maddeler arasından uygun seçimlerle, yanıtlayıcıların yeterlik (proficiency) düzeylerine uygun maddeler seçilerek, her birey için tüm maddelerin aynı olmadığı test olarak tanımlanabilir (Weiss, 2004). Bunu başarabilmek için de, tüm yanıtlayıcılara aynı güçlük dağılımında aynı maddeleri vermek yerine, aşağı-yukarı yöntemi olarak da bilinen; yanıtlayıcı doğru yanıt verirse daha zor, yanlış yanıt verirse daha kolay bir maddenin sorulmasına dayanan bir yöntem kullanılır (Rudner, 1998). Bu nedenle BOBUT uygulamalarında bireyin karşısına yeterlik düzeyine en yakın madde getirildiğinden uygulanan madde sayısında önemli miktarda bir azalma da sağlanmış olmaktadır. Böylece daha az madde ile daha güvenilir ölçme sonuçlarının elde edilmesi de mümkün olabilmektedir. (Çıkrıkçı-Demirtaşlı, 1999; Embretson ve Reise, 2000; Kalender, 2009; Mcglohen ve Chang, 2008).

BOBUT uygulamalarında önemli bir nokta, uygulama süresince yapılacak yeterlik kestirimlerinde hangi kuramın dikkate alındığı ve güvenilir sonuçlara nasıl ulaşıldığıdır. Birçok BOBUT uygulamasının psikometrik temeli Madde Tepki Kuramı (MTK)'na dayalı olarak düzenlenmektedir. Bazı BOBUT uygulamaları Klasik Test Kuramına (KTK) dayalı olarak yapılsa da (Frick, 1992; Rudner, 2002), BOBUT uygulamasında MTK'nın da sağladığı bazı avantajlardan (bireye özgü yeterlik ve hata kestirimi yapılabilmesi, madde ve test parametrelerinin değişmezlik özelliği gibi) yararlanarak, gerek testi alan yanıtlayıcıya gerekse testi uygulayana birtakım kolaylıklar sağlaması, BOBUT uygulamalarında KTK yerine MTK'nın tercih edilmesine neden olmuştur. MTK, bireylerin testle ölçülen yeterlik düzeyi ile testteki herhangi bir maddeyi yanıtlama davranışı arasında bir ilişki olduğunu belirten ve bu ilişkiyi olasılıklı bir modelle açıklayan bir kuramdır (Embretson ve Reise, 2000; Hambleton, Swaminathan ve Rogers, 1991; Wainer ve diğerleri, 1990). Ayrıca test geliştirme, madde analizi ve puanlama gibi bazı avantajlara sahip güçlü bir psikometrik paradigma olması da (Thompson ve Weiss, 2011), BOBUT uygulamalarında MTK'nın tercih edilmesini arttırmıştır.

MTK'ya dayalı BOBUT uygulamasında aşağıdaki sıra izlenir (Lord ve Stocking, 1988) :

- Belli bir yöntemle bireyin yeterlik parametresi bir kestirimi elde edilir.

- Madde parametreleri daha önceden kestirilmiş maddelerden oluşan bir havuzdan, bireyin yeterliğini en iyi kestirecek maddeler seçilir.
- Seçilen maddeler uygulanır ve bir sonraki madde seçilmeden önce bireyin son yeterlik düzeyi kestirilerek, havuzdan bireyin son yeterlik düzeyine en uygun madde seçilir.
- Seçilen sonlandırma kuralına göre test uygulaması sonlandırılır.

BOBUT uygulamalarının önemli bir boyutu da test uygulamasını başlatma, sürdürme ve sonlandırmada kullanılan ölçütlerdir. BOBUT uygulamalarındaki test başlatma ve test sürdürmede yeterlik kestirim yöntemlerini karşılaştıran çeşitli araştırmalar (Barrada, Olea, Ponsoda, Abad, 2010; Eggen, 2004; Keller, 2000; Kingsbury ve Zara, 1989; van Rijn, Eggen, Hemker ve Sanders, 2002) yapılmıştır. Bu araştırmalarda, test başlatma ve testi sürdürmede güvenilir kestirim veren yöntemin *En Yüksek Olabilirlik* -EYOY (Maximum Likelihood Estimation-MLE) Yöntemi olduğu bulunmuştur.

Yurt içinde ve yurt dışında yapılan araştırmaların çoğunda (Bulut ve Kan, 2012; Chae, Kang, Jeon ve Linarce, 2000; Frick, 1992; İşeri, 2002; Kalender, 2011; Kaptan, 1993; Keller, 2000; Kezer, 2013; Koşan-Aytuğ, 2013; McDonald, 2002; Miller, 2003; Mills ve Stocking, 1996; Mills ve Steffen, 2000; Öztuna, 2008; Rudner ve Guo, 2011; Scfhaer, Steffen, Golub-Smith, Mills ve Durso, 1995; Tian, Miao, Zhu ve Gong, 2007; Weiss ve Betz, 1973; Wainer, Dorans, Flaugher, Green, Mislevy ve Steinberg, 1990; Zitny, Halama, Jelinek, ve Kveton, 2012) simülatif BOBUT uygulamaları farklı uygulama stratejileri altında karşılaştırılarak, BOBUT uygulamasının kâğıt-kalem testine göre göstereceği psikometrik farklılıklar tespit edilmeye çalışılmıştır. BOBUT uygulaması ile kâğıt-kalem testi uygulamalarından kestirilen yeterlik /başarı düzeylerini karşılaştıran araştırmalar incelendiğinde, BOBUT uygulamalarının yeterlik kestirimlerinin güvenirliğini (yeterlik düzeyinde kestirilen standart hata değerinin düşük olması) artırdığı ve kullanılan madde sayısında önemli ölçüde tasarruf sağlandığı ulaşılan ortak bulgular arasındadır.

Türkiye'de merkezi olarak uygulanan geniş ölçekli sınav uygulamalarında (Yabancı Dil Sınavı, Temel Eğitime ve Orta Öğretime Geçiş Sınavı, Yükseköğretime Geçiş Sınavı, Lisans Yerleştirme Sınavı, vb.) çeşitli testler uygulanmakta ve bu uygulanan testlerin sonuçlarına dayanarak bireyler hakkında önemli kararlar alınmaktadır. Ancak bu testlerde, yanıtlayıcının kendi yeterlik düzeyine denk olan ve olmayan tüm soruları yanıtlaması beklenmektedir. Yanıtlayıcının tüm soruları yanıtlaması daha uzun zaman harcanmasına neden olmakta ayrıca yanıtlayıcının kendi yeterlik düzeyinin üstündeki çok zor ve altında kalan çok kolay birçok soruyu da yanıtlamasını gerektirmektedir. Bunların dışında bu tür sınavlarda test gizliliği konusunda da önemli sınırlılıklar bulunmaktadır. (Davis ve Dodd, 2005; French ve Thompson, 2003). Tüm bu sınırlılıkların önemli bir kısmı bu testler BOBUT uygulaması olarak geliştirildiğinde giderilebilir.

Türkiye'de BOBUT uygulamaları ile ilgili sınırlı sayıda görgül çalışma bulunmaktadır. İlk çalışmalar başarı testlerinin BOBUT olarak uygulanabilirliğini (Kaptan, 1993; Köklü, 1990; Yaşar, 1999) sonraki çalışmalar ortaöğretim ve yükseköğretim okullarına öğrenci seçmede kullanılan testlerle, farklı konularda yeterlik belirleme amaçlı testlerin BOBUT olarak uygulamasının geleneksel (kâğıt kalem testi) uygulamayla karşılaştırmasını konu edinmiştir (Aytuğ-Koşan, 2013; Bulut ve Kan, 2012; İşeri, 2002; Kalender, 2011; Kezer, 2013). Bir kısım çalışma ise, tıp alanında hasta beyanına dayalı teşhis araçlarının BOBUT olarak geliştirilmesiyle ilgilidir (Öztuna, 2008).

Bu çalışmada, üniversitelerde zorunlu temel derslerden biri olan "Temel bilgisayar" dersinden muaf tutulacak öğrencilere karar vermede kullanılan "Bilgisayar Okuryazarlığı (BİLOKUR) Muafiyet" testinin BOBUT olarak uygulanabilirliği araştırılmıştır. Her yıl uygulanmakta olan Bilgisayar muafiyet sınavı ile çok sayıda öğrencinin bu dersi alıp almayacağına ilişkin karar verilmektedir. Bu sınavlara yönelik olarak, her yıl birçok soru hazırlanmaktadır. Bu sınavlarda uygulanan testlerde soruların tamamını, yeterlik düzeyi ne olursa olsun tüm öğrenciler yanıtlamak durumundadır. Bir başka ifadeyle, öğrenciler kendi yeterliklerinin üstünde ve altında kalan gereğinden fazla sayıda soruyu da yanıtlamak zorunda kalmaktadırlar. Bu durum testlerin kullanışlılığının zayıflamasına ve test maliyetinin artmasına neden olmaktadır. BİLOKUR testinin bu sınırlıkları gidermek üzere,

kâğıt-kalem testi olarak olarak uygulanan BİLOKUR testinin BOBUT olarak uygulananabilirliğini sınamak bu araştırmanın problemini oluşturmaktadır.

Araştırmanın Amacı

Bu çalışmanın amacı, Bilgisayar Okuryazarlığı Testi (BİLOKUR)'nin BOBUT olarak uygulanabilirliğini farklı koşullar için araştırmaktır. Bu amaçla testin BOBUT uygulaması ile kâğıt-kalem uygulamasından elde edilen psikometrik nitelikleri karşılaştırılarak, en uygun BOBUT uygulaması stratejileri (testi sürdürme ve sonlandırma stratejileri) saptanmaya çalışılmıştır.

Bu genel amaç doğrultusunda BİLOKUR Testinin BOBUT olarak uygulanabilirliği aşağıdaki iki soru kapsamında sınanmıştır;

1) Simülatif BOBUT uygulamasında farklı yeterlik kestirim yöntemleri (En yüksek olabilirlik Yöntemi - EYOY ve Beklenen sonsal dağılım- BSD) ile farklı test sonlandırma kuralları (sabit test uzunluğu (k=30) ve ölçmenin standart hatası (SH<0.30 ve SH<0.50) kapsamında elde edilen birey yeterlik parametreleri ve test bilgi değeri (güvenirlik) arasında anlamlı bir fark var mıdır?
2) Canlı BOBUT uygulaması için sonlandırma kuralı olarak standart hata değeri ve sabit madde koşulu uygulandığında, birey yeterlik parametreleri, test güvenirliği ve madde sayıları kâğıt-kalem uygulamasından elde edilen değerlerden anlamlı bir farklılık göstermekte midir?

YÖNTEM

"Araştırmada, BİLOKUR testinin BOBUT olarak uygulanabilirliğini test etmek üzere, farklı yeterlik kestirim yöntemleri ve farklı sonlandırma kuralları karşılaştırılmıştır. Çalışmada, post-hoc simülasyon yöntemi ile farklı yeterlik kestirim yöntemleri EYOY, BSD ve test sonlandırma kurallarına (SH <0.30 ve SH<0.50), dayalı koşullara bağlı olarak yeterlik kestirimleri yapılmıştır. Bu amaçla simülatif BOBUT programı olan SimulCAT (Han, 2010) yazılımından yararlanılmıştır. Bu yönüyle araştırma mevcut kuramsal bilginin gelişmesine ve genişlemesine katkıda bulunan, aynı zamanda uygulamaya da katkı getiren temel araştırma modelindedir. Temel araştırmalar Karasar (2011)'e göre, kuramlara dayalı olarak teorilerin gelişmesine katkıda bulunmak, varsayımlar geliştirerek ve bunları test ederek, sonuçlarını bilimsel olarak yorumlayarak bilgilerin genişlemesini ve gelişmesini amaçlayan araştırmalardır. Daha sonra ise, canlı (live) BOBUT ve kağıt kalem testinden elde edilen yeterliklerin psikometrik özellikleri bir grup öğrencinin katıldığı uygulamada karşılaştırılmıştır. Bu amaçla, Kalender (2011) tarafından geliştirilmiş olan BOBUT uygulama yazılımı kullanılmıştır. Bu yönüyle de araştırma var olan bir uygulamayı geliştirme amacını taşıyan uygulamalı araştırma niteliğindedir.

Çalışma Grubu

Araştırmanın amacına yönelik olarak BİLOKUR testine ait veriler, Ankara Üniversitesi'nin çeşitli fakültelerinde 2012-2013 ve 2013-2014 eğitim öğrenim yılında birinci sınıfta öğrenim görmekte olan üniversite öğrencilerine uygulanarak elde edilmiştir. Çalışma grubunu, bu grupta yer alan farklı öğrencilerin oluşturmuş ve çalışmanın iki aşamasında yer almışlardır. İlk aşamada, soru bankası oluşturmak için, 2012-2013 eğitim öğretim yılında aynı bilgisayar dersinin okutulduğu birinci sınıf öğrencilerine uygulanmıştır. Uygulama; Eğitim, Hukuk, Mühendislik, Eczacılık, Dil-Tarih Coğrafya, Tıp fakületesi birinci sınıfta okuyan toplam 1452 üniversite öğrencisini kapsamıştır. Geliştirilen BİLOKUR testi maddeleri, 6 farklı madde grubuna bölünerek araştırma kapsamındaki öğrencilere uygulanmıştır. Bunun nedeni, aynı öğrencilerin bir oturumda 191 soruya yorgunluk, sıkılma faktörleri yüzünden güvenilir bir şekilde yanıt verememe olasılığıdır. Ancak ilk aşamada, öğrencilerin bazıları teste hiç yanıt vermediği veya birkaç maddeye yanıt verdiği için çalışma grubundan çıkarılmış ve sonuç olarak 1366 üniversite birinci sınıf öğrencisiyle pilot uygulamaya

ilişkin veriler toplanmıştır. İkinci aşamada, bu veriler simülatif BOBUT uygulamasında kullanılmıştır. Uygulamaya katılan öğrencilerin dağılımı Tablo 1'de verilmiştir.

Tablo 1. Soru Bankasının Pilot Uygulamasına Katılan Öğrencilerin Cinsiyetlerine Göre Dağılımı

Cinsiyet	F	%
Erkek	679	49
Kız	687	51
Toplam	1366	100

Araştırmanın son aşamasında, parametreleri kestirilen ve MTK'ya uygunluğu tespit edilen BİLOKUR testi maddeleri canlı BOBUT ve kağıt-kalem testi olarak, Ankara Üniversitesi Eğitim Bilimleri Fakültesi'nde farklı bölümlerde (sınıf öğretmenliği, sosyal bilgiler öğretmenliği, rehberlik ve psikolojik danışmanlık, bilgisayar eğitimi ve teknolojileri öğretmenliği ve okul öncesi öğretmenliği) öğrenim görmekte olan 2013-2014 eğitim öğretim yılında 142 üniversite birinci sınıf öğrencisine uygulanmıştır. Bu uygulamaya katılan öğrencilerin cinsiyetlerine ilişkin dağılım Tablo 2'de verilmiştir.

Tablo 2. Kağıt-Kalem Testi ile Canlı (Live) BOBUT Uygulamasına Katılan Öğrencilerin Cinsiyetlerine Göre Dağılımı

Cinsiyet	f	%
Erkek	32	23
Kadın	110	77
Toplam	142	100

Veri Toplama Araçları

Araştırmada veri toplama aracı olarak, BİLOKUR testi kullanılmıştır. Bu test aynı öğrencilere hem kağıt-kalem ortamında hem de BOBUT olarak uygulanmıştır.

Bilgisayar Okuryazarlığı Testi

Çalışmada kullanılan ölçme aracı, bilgisayar okuryazarlığıyla ilgili temel bilgi ve becerileri ölçmeyi amaçlayan BİLOKUR testidir. Bu test, her yıl üniversiteye yeni başlayan tüm öğrencilere Ankara Üniversitesi Enformatik Bölümü tarafından uygulanarak, bu dersten muaf olup olmayacak öğrenciler saptanır. Bu test, Avrupa Bilgisayar Yetkinlik Belgesi (European Computer Driving Licence-ECDL) programında tanımlanan bilgi ve becerileri ölçmeyi amaçlar. Bilgisayar okuryazarlığı testinin kapsamı "Bilgi ve İletişim Teknolojisi Kavramları" modülündeki yeterliklerle ile sınırlıdır. Bu modül; bir kişisel bilgisayarın fiziksel yapısı, veri saklama, bellek, toplumda çok kullanılan yazılım uygulamaları ve bilgisayar ağlarının kullanımı ile ilgili temel kavramların bilinmesini içermektedir. Aday ayrıca, bilgi teknolojilerinin günlük kullanımı ve bilgisayarların insan sağlığına etkileri ile bilgisayarlarla ilgili bazı güvenlik ve hukuk konuları hakkında bilgi sahibi olmalıdır (http://enformatik.ankara.edu.tr/?page_id=174).

Bilgi ve İletişim Teknolojisi Kavramları modülü temel alınarak hazırlanan BİLOKUR testi kapsamında adaylarda yoklanan beceriler sıralanmıştır (ECDL, 2007):

- Donanımı ve bilgisayar performansını nasıl etkilediğini bilir,

- Yazılımın ve uygulama yazılımlarının ne olduğunu anlar ve örneklendirir,
- Bilgisayarlar arasındaki ağın nasıl olduğunu anlar ve internete bağlanmanın değişik yollarını bilir,
- Bilgi ve İletişim Teknolojisi Kavramlarını anlar günlük yaşamımızda kullanımını örneklendirir,
- Bilgisayar kullanımında güvenlik ve sağlık faktörlerini anlar,
- Telif hakları, veri koruma ve yasal kullanım hakkındaki temel bilgileri bilir.

BİLOKUR testi, teknoloji okuryazarlığı alanındaki taksonomilere (Tomei, 2005) uygun olarak ve bilgisayar okuryazarlığı ile ilgili temel bilgi ve becerileri başat faktör olarak ölçmek üzere hazırlanmıştır. Bu süreçte hazırlanan maddeler havuzu belirtke tablosunda belirtilen becerilere dayalı olarak, iki ölçme ve değerlendirme uzmanı ile bir bilgisayar eğitimi ve teknolojileri alanında uzman olan üç kişilik bir uzman grup tarafından i) maddenin beceriyi temsil durumu ii) maddenin çoktan seçmeli madde tekniğine uygun yazılma durumu iii) bilimsel doğruluk iv) dil ve anlatım bakımından uygunluk ölçütleri bakımından; "uygun", "uygun değil" ve "düzeltilmeli" kategorileri kapsamında incelenmiştir. Bu ölçütler bakımından onaylanan ve düzeltme önerisi verilen maddeler (191 madde) gözden geçirilerek testin deneme uygulamasına alınmıştır.

BİLOKUR testini oluşturan maddeler, 2012-2013 eğitim-öğretim yılında toplam 1366 öğrenciye yaklaşık iki ay gibi bir sürede uygulamıştır. Madde havuzunun geniş olmasından dolayı, maddeler 6 farklı soru grubu/test olarak uygulanmıştır. Uygulama sonrası madde istatistikleri KTK'ya dayalı olarak hesaplanmış ve özet sonuçlar Tablo 3'te verilmiştir.

Tablo 3. BİLOKUR Testi'nde Kullanılan Maddelere İlişkin Betimsel İstatistikler

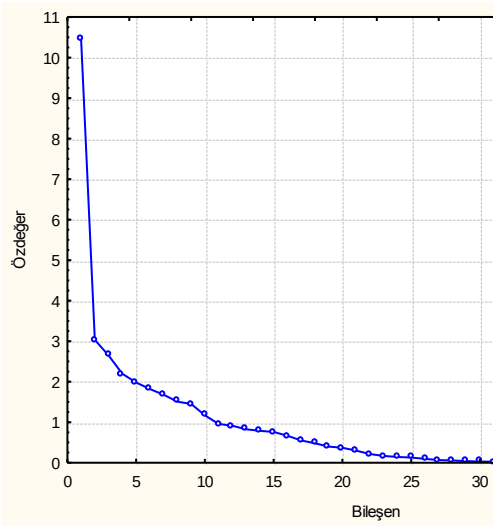
	Madde grupları (k: madde sayısı)					
Test İstatistikleri	Grup1 (k= 50)	Grup2 (k=50)	Grup3 (k=46)	Grup4 (k=15)	Grup5 (k=20)	Grup6 (k=10)
Ortalama	29.62	21.32	23.03	7.70	17.43	6.26
Ortanca	30	20	21	8	18	6
Ortalama $\bar{p}$	0.59	0.43	0.50	0.51	0.87	0.62
Ortalama ayırt edicilik	0.57	0.41	0.38	0.33	0.69	0.57
Tepe Değeri	31	16	21	8	20	7
En Küçük	11	6	7	0	0	0
En Büyük	44	43	45	14	20	10
Standart Sapma	6.21	6.77	7.86	2.99	3.25	1.82
Varyans	38.52	45.85	61.82	8.95	10.59	3.33
Basıklık	0.43	0.30	-0.12	-0.34	8.49	0.08
Çarpıklık	-0.62	0.68	0.47	-0.34	-0.92	-0.58
Çarpıklığın Standart Hatası	0.14	0.16	0.17	0.15	0.07	0.08

Tablo 3 incelendiğinde, maddelerin çoğunun ortalama güçlük düzeyinde olduğu, az sayıda maddenin kolay madde olduğu saptanmıştır. Maddelerin ortalama madde ayırt edicilik değerleri incelendiğinde ise, en küçük 0.33 en yüksek 0.69 arasında değiştiği bulunmuştur.

BOBUT uygulamasında MTK'ya göre ölçeklenmiş maddelerin kullanılması, MTK'nın birey ve madde parametrelerinde değişmez (invariant) kestirimler vermesini sağlar. Ancak bu avantajların elde edilmesi, büyük ölçüde kullanılan verilerin model ile uyumlu olmasına (Fan, 1998; Hambleton ve Swaminathan, 1989) ve ölçmeye konu olan psikolojik özelliğin, eldeki veriler tarafından ölçülebildiğinin kanıtlarının ortaya konmasına (Stark ve Chernyshenko, 2001) bağlıdır. Bu amaçla, soru bankasında kalan maddeler MTK'ya dayalı ön analizlere tabi tutulmuştur. Bu analizlerde MTK'nın tek boyutluluk, yerel bağımsızlık varsayımlarının karşılanma durumu, hız testi olup olmadığının kontrolü ile madde ve birey yeterlik parametrelerinin değişmezliği sınanmıştır.

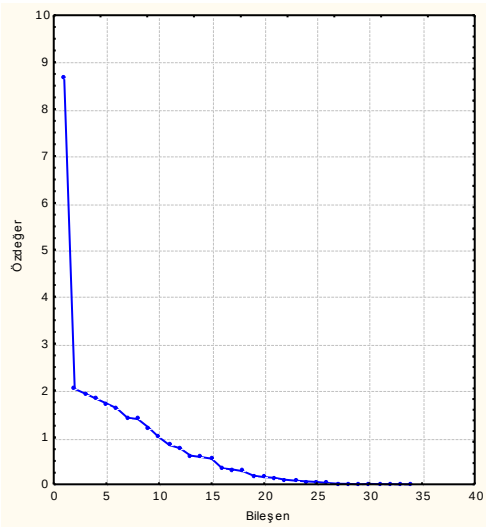
MTK'nın en önemli varsayımlarından biri olan tek boyutluluk, testi oluşturan maddelerin başat bir özelliği ölçmesi ve madde faktör yüklerinin bu boyut altında yük değerleri (faktör yükü> 0.30) vermesi olarak tanımlanmaktadır (Embretson ve Reise, 2000; Hambleton ve Swaminathan, 1985). Tek boyutluluğun sınanmasında açımlayıcı faktör analizi (AFA) kullanılmıştır. AFA değişken azaltma ve ortaya çıkan faktörleri isimlendirmenin dışında, davranışın anlaşılmasına olanak veren kuramsal yapı (gözlenemeyen gizil/örtük değişkenler) ile benzer olup olmadığını ortaya koyar (Kline, 2000). AFA'nın 1-0 şeklinde kategorik olarak puanlanan verilerde uygulanabilmesi için tetrakorik korelasyon matrisinin oluşturulması gerekmektedir (Baykul, 2010; Sheskin, 2004). Bu çalışmada her madde grubu ayrı yanıtlayıcı gruplarında uygulandığından, her grupta uygulanan madde grubunda maddeler-arası "tetrakorik korelasyon" matrisine dayalı temel eksenler analizine göre madde gruplarının (testin) kendi içerisinde tek boyutlu olup olmadığı incelenmiştir. Buna göre, tek boyutluluk varsayımının sınanmasında boyut sayısına karar verilirken kullanılabilecek en pratik yol olarak tetrakorik maddeler-arası korelasyon matrisinin örtük kökleri (λ_r) dikkate alınmıştır (Lord, 1980).

Buna göre, birinci özdeğer ikinci özdeğere göre büyükse ve ikinci öz değer kendinden sonra gelen özdeğerle arasındaki fark büyük değilse, maddelerin tek boyutlu bir yapıyı temsil ettiğinden söz edilebilir. Buna ilişkin madde gruplarına ait yamaç birinti grafikleri ve açıklanan varyans değerleri aşağıda verilmiştir.



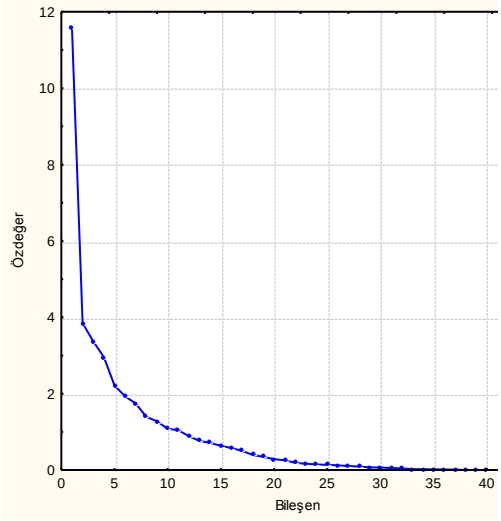
Şekil 1. Grup 1 maddeleri Yamaç Birikinti grafiği

(1. Boyutun açıkladığı toplam varyans: %28.97)

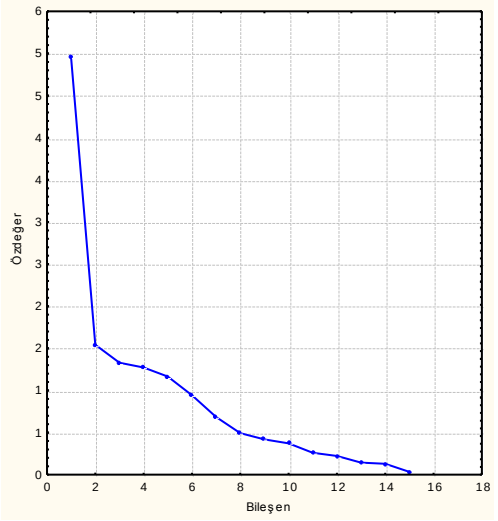


Şekil 2. Grup 2 maddeleri yamaç birikinti grafiği

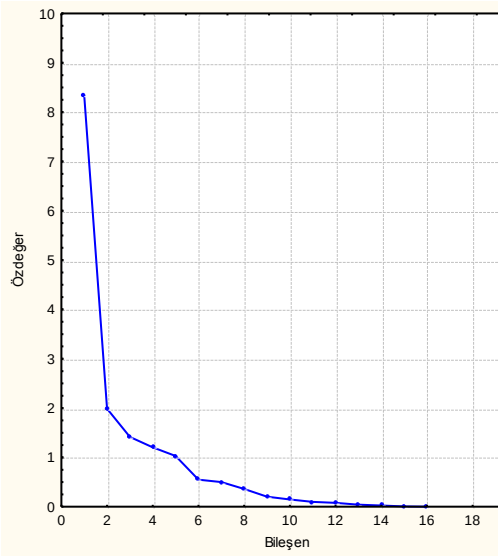
(1.Boyutun açıkladığı toplam varyans: %25.53)



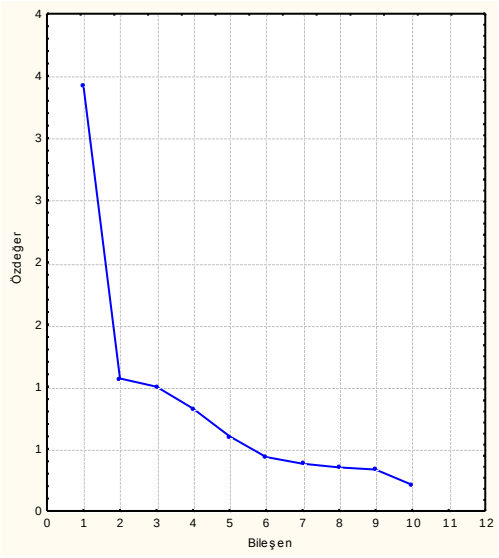
Şekil 3**.** Grup 3 maddeleri yamaç birikinti grafiği (1.Boyutun açıkladığı toplam varyans:% 37.78)



Şekil 4. Grup 4 maddeleri yamaç birikinti grafiği (1.Boyutun açıkladığı toplam varyans: 33.10)



Şekil 5. Grup 5 maddeleri yamaç birikinti grafiği (1.Boyutun açıkladığı toplam varyans: %52.16)



Şekil 6. Grup 6 maddeleri yamaç birikinti grafiği (1.Boyutun açıkladığı toplam varyans: %35.50)

MTK modellerinin avantajlarından yararlanmak için, maddelerin yerel bağımsızlık göstermesi de beklenir. Yanıtlayıcının yeterliği sabit olmak üzere test maddelerinden birine verdiği yanıtın, testin diğer maddelerine verdiği yanıtı etkilememesi olarak tanımlanan yerel bağımsızlık varsayımında; belli bir yeterlik düzeyindeki kişilerin maddelere vermiş oldukları yanıtlar arasındaki korelasyonun sıfır olduğu kabul edilir (Lord, 1980). Alan yazında birçok araştırmacı yerel bağımsızlık varsayımını, tek boyutluluk ile ilişkilendirerek açıklamıştır (Emretson ve Reise, 2000; Hambleton, Swaminathan ve Rogers, 1991). Bu kapsamda, araştırmada kullanılan tüm maddelerin uygulandığı gruplarda başat tek boyutluluğu sağladığı ölçüde yerel bağımsızlık varsayımını da karşıladığı sonucuna ulaşılmıştır. Bu bakımdan, bu araştırmada kullanılan BİLOKUR testi maddeleri, yerel bağımsızlık özelliğini göstermektedir.

MTK'da test edilmesi gereken önemli varsayımlardan biri de testin hız testi olup olmadığıdır. Çünkü MTK'daki modeller, hız sınırlaması olmadan uygulanan testlerdeki modellere uygundur.

Yanıtlayıcılar zamanı yetişemediği için maddeye yanlış yanıt vermemiş veya maddeyi boş bırakmamış olmalıdır (Hambleton ve Swaminathan, 1989). Bu bağlamda, araştırmada kullanılan testin hız testi gibi çalışıp çalışmadığı MTK modellerine uyum sağlaması açısından oldukça önemlidir. Hambleton, Swaminathan ve Rogers (1991), hız testi kontrolü için önerdikleri ölçüt, maddelerin hemen hemen tümünü tamamlayan öğrenci sayısının tüm öğrenci sayısına yakın olmasıdır. Bu yöndeki incelemeler sonucunda, her madde grubunda "son maddeye erişen öğrenci" oranının %95 ile %100 arasında değiştiği saptanmıştır. Bu sonuç, her grupta alınan testlerin bir hız tesi gibi çalışmadığını kanıtlar niteliktedir.

MTK ile ölçme aracından elde edilen verilerin, yorumlanması ve raporlanmasında sağladığı avantajların elde edilmesi, büyük ölçüde kullanılan veriler (maddeler) ve MTK modeli arasındaki uyuma bağlıdır (Fan, 1998; Hambleton ve Swaminathan, 1989). MTK'da model-veri uyumunun olup olmadığının kontrol edilmesi için öncelikle MTK'nın temel varsayımlarının karşılanıp karşılanmadığının test edilmesi gerekir. Bu bağlamda, araştırmada MTK varsayımlarının karşılanmasına ilişkin yukarıda açıklanan analizlerden elde edilen bulgular, verilerin model-veri uyumu için kanıt niteliğindedir. Ayrıca -2likelihood değerleri, modele ilişkin test düzeyinde gözlenen ve beklenen madde yanıt örüntülerine bağlı doğru ve yanlış yanıtlanma olasılıkları farklı yeterlik aralıklarında bulunan bireyler için karşılaştırılır. (Reise, 1990; Zimowski, Muraki, Mislevy ve Bock 2003). Araştırma kapsamında BİLOKUR testine ilişkin -2likelihood değerleri, modele ilişkin doğru karar verebilmek amacıyla her madde grubu için incelenmiş ve Tablo 4'te verilmiştir.

Tablo 4. Maddelerin model veri uyumuna ilişkin MTK modellerinden hesaplanan -2Loglikelihood değerleri

Madde Grubu	1PLM	2PLM	3PLM
Grup 1	10052.90	9974.30	9968.58
Grup 2	4219.67	4172.18	3837.42
Grup 3	4379.50	4256.12	4050.86
Grup 4	2241.53	2263.78	2233.88
Grup 5	7486.56	7238.04	5553.42
Grup 6	4586.15	4566.18	4537.55

Tablo 4'te verilen -2Loglikelihood değerleri incelendiğinde, 3PLM ile hesaplanan -2Loglikelihood değerleri diğer modellerden daha düşük olması, bu modele verilerin daha iyi uyum verdiğini göstermiştir. Ayrıca kullanılan testin maddelerinin çoktan seçmeli olması, şans başarısı faktörünü gözeten bir model (Crocker ve Algina, 1986; Hambleton, Swaminathan ve Rogers, 1991) olduğu için kestirimlerde 3PL model tercih edilmiştir.

3PL'ye göre kestirimleri yapılan bu maddelerin madde ayırıcılık gücü indeksi, a, ortalama değeri 1.132; ortalama madde güçlük indeksi, b, 0,542 ve ortalama şans parametresi c, değeri 0.363 olarak hesaplanmıştır. Model-veri uyumu testi sonucunda 15 madde model veri uyumu göstermediği için soru bankasından çıkarılmıştır.

BİLOKUR testi soru bankasından tek boyutluluk, model veri uyumu göstermeme gibi nedenlerden dolayı toplamda 55 madde çıkarılmış ve nihai olarak soru bankasında, bilgisayar okuryazarlığı becerilerini ölçebilecek 136 madde kalmıştır.

Bu sınamanın ardından, madde ve birey parametrelerinin değişmezliği incelenmiştir. MTK modeline göre kestirilen madde parametrelerinin değişmezliği, madde parametrelerinin testin uygulandığı gruptan bağımsız olarak kestirilebilmesidir (Hambleton, Swaminathan ve Rogers,1991). Bunu sınamak için, her madde grubunun uygulandığı öğrenci grubu tesadüfi olarak iki gruba ayrılarak her alt örneklemden madde parametreleri kestirilmiş ve kestirilen madde parametreleri arasındaki korelasyon "Pearson Momentler Çarpımı Korelasyon Katsayısı" ile incelenmiştir. Ayrıca, parametre değişmezliği incelemek için oluşturulan farklı öğrenci gruplarından kestirilen madde parametrelerinin büyüklük sırasının benzer olup olmadığının incelemek amacıyla da "Spearman Brown Sıra Farklı Korelasyonu" yapılarak incelenmiş ve sonuçlar Tablo 5'te verilmiştir.

Tablo 5. Madde Parametrelerinin Değişmezliğine ilişkin korelasyon değerleri

Madde grubu	Yeterlik Grubu		a	b	c
Grup 1 (k=33)	Grup1-Grup2	Pearson	0.36**	0.91**	0.68**
		Spearman	0.34**	0.90**	0.52**
Grup 2 (k=31)	Grup1-Grup2	Pearson	0.05	0.64**	0.68**
		Spearman	0.06	0.66**	0.66**
Grup 3 (k=38)	Grup1-Grup2	Pearson	-0.12	0.69**	0.76**
		Spearman	-0.07	0.62**	0.69**
Grup 4 (k=12)	Grup1-Grup2	Pearson	0.48**	0.75**	0.85**
		Spearman	0.59*	0.79**	0.68**
Grup 5 (k=14)	Grup1-Grup2	Pearson	0.62**	0.97**	0.76**
		Spearman	0.56**	0.94**	0.70**
Grup 6 (k=8)	Grup1-Grup2	Pearson	-0.05	0.63*	-0.23
		Spearman	-0.06	0.91**	-0.32

* p<0.05; **p<0.01; k: madde sayısı

Tablo 5'te görüldüğü gibi, her madde grubunun uygulandığı gruplarda, tesadüfi olarak iki gruba ayrılan bireylerden kestirilen a, b ve c parametreleri arasındaki korelasyonlarda, genellikle b ve c parametrelerinde orta ve iyi düzeyde manidar ilişki bulunmuştur. Ancak madde ayırt edicilik parametrelerine ilişkin korelasyonlar madde güçlük ve şans parametresine göre daha düşük düzeyde çıkmıştır. Bu durum, alan yazında yapılan çalışmalarla (Çıkrıkçı-Demirtaşlı, 2002; Fan, 1998; Kalender, 2011; Kelecioğlu, 2001; Kezer, 2013) paralellik göstermektedir. Stocking (1990) madde parametrelerinin kestirim yapıldığı grubun homojen olmasının, madde parametrelerinin değişmezliğini azalttığını belirtmiştir.

MTK'nın değişmezlikle ilgili diğer sayıltılarından birisi de yeterlik parametrelerinin değişmezliğidir. Araştırmada yanıtlayıcılara ait yeterlik parametrelerinin madde örnekleminden bağımsız kestirilip kestirilmediğini saptamak için maddeler tesadüfi olarak iki gruba ayrılmıştır. Buna gore, BİLOKUR testinden model-veri uyumunu sağlayan maddeler tesadüfi olarak iki gruba ayrılmış, iki madde seti oluşturulmuştur. Bu madde setlerinden kestirilen birey yeterlik parametreleri arasındaki korelasyonlar hesaplanmış ve sonuçlar Tablo 6'da özetlenmiştir.

Tablo 6**.** Farklı Madde Setlerine Ait Yeterlik Ölçüleri Arasındaki Korelasyonlar

Madde grubu	Madde Seti	Madde Seti 1	Madde Seti 2	Testin Tümü
Grup 1	Madde Seti 1	1.00		
	Madde Seti 2	0.69*	1.00	
	Testin Tümü	0.89*	0.93*	1.00
Grup 2	Madde Seti 1	1.00		
	Madde Seti 2	0.75*	1.00	
	Testin Tümü	0.92*	0.92*	1.00
Grup 3	Madde Seti 1	1.00		
	Madde Seti 2	0.78*	1.00	
	Testin Tümü	0.93*	0.94*	1.00
Grup 4	Madde Seti 1	1.00		
	Madde Seti 2	0.52*	1.00	
	Testin Tümü	0.88*	0.84*	1.00

Tablo 6 Devamı

Madde grubu	Madde Seti	Madde Seti 1	Madde Seti 2	Testin Tümü
Grup 5	Madde Seti 1	1.00		
	Madde Seti 2	0.54*	1.00	
	Testin Tümü	0.95*	0.77*	1.00
Grup 6	Madde Seti 1	1.00		
	Madde Seti 2	0.34*	1.00	
	Testin Tümü	0.86*	0.64*	1.00

**p<0.01

Her madde grubunda, alt madde setlerinden kestirilen yeterlik parametreleri arasındaki ilişkilerin pozitif yönde orta ve yüksek düzeyde olmak üzere anlamlı ilişkiler gösterdiği bulunmuştur. Buna göre BİLOKUR testini oluşturan soru bankasında kestirilen yeterlik parametrelerinin büyüklerine ilişkin sıraların madde alt setlerine göre tutarlı olduğu, değişmediği sonucuna ulaşılmıştır.

Verilerin Analizi

Yukarıda özetlenen ön analizlerden sonra, araştırma soruları çerçevesinde simülatif BOBUT uygulaması ile farklı yeterlik kestirim yöntemleri ve test sonlandırma koşullarında kestirimler yapılmıştır. Ardından simülatif BOBUT uygulamasından çıkan sonuçlara göre, kâğıt-kalem testinden elde edilen yeterlik kestirimlerine ilişkin karşılaştırmalar yapmak üzere canlı BOBUT uygulaması yapılmıştır. Farklı test sonlandırma (sabit madde ve sabit hata) ve farklı yeterlik kestirimi koşullarında tekrarlanan BOBUT uygulamalarından kestirilen yeterlik ölçülerinin farklılığı tek yönlü varyans analizi kullanılarak incelenmiştir. Varyans analizinde, varyansların homojen olmaması durumunda gruplar arası farklar için Dunnet C testi kullanılmıştır. Ayrıca simülatif ve canlı BOBUT uygulamalarından elde edilen test bilgi miktarları arasında anlamlı bir fark olup olmadığını incelemek için bağımsız örneklemler için t-testi kullanılmıştır.

Simülatif BOBUT Uygulaması

Simülatif BOBUT uygulaması, SimulCAT (Han, 2010) programının yardımıyla yapılmıştır. Bu uygulama için gereken yeterlik parametreleri, madde havuzunun kağıt-kalem testi olarak uygulanmasından elde edilen veriler 3PLM göre ve yetenek kestirim yöntemi olarak, EYOY kullanılarak elde edilmiştir. SimulCAT programının varsayılan ayarları; madde kullanım sıklığının kontrol edilmemesi, kapsam dengelemesinde (content balancing) herhangi bir kontrol yapılmaması ve başlangıç maddesi için yeterlik düzeyi -0.5 ile 0.5 arasında herhangi bir madde ile başlaması gibi özellikleri değiştirilmemiştir. Simülatif BOBUT uygulamasında, EYOY, Bayes ve farklı sonlandırma kuralları (sabit soru sayısı ve sabit standart hata değeri) kullanılarak yeterlik kestirimleri yapılmış, elde edilen yeterlik kestirimleri ve standart hata değerleri incelenmiştir. Güvenirlik ve SH arasındaki ilişki (Wang, Hanson ve Lau, 1999):

$$r^2 = 1 - SH(\theta)^2 \quad (1)$$

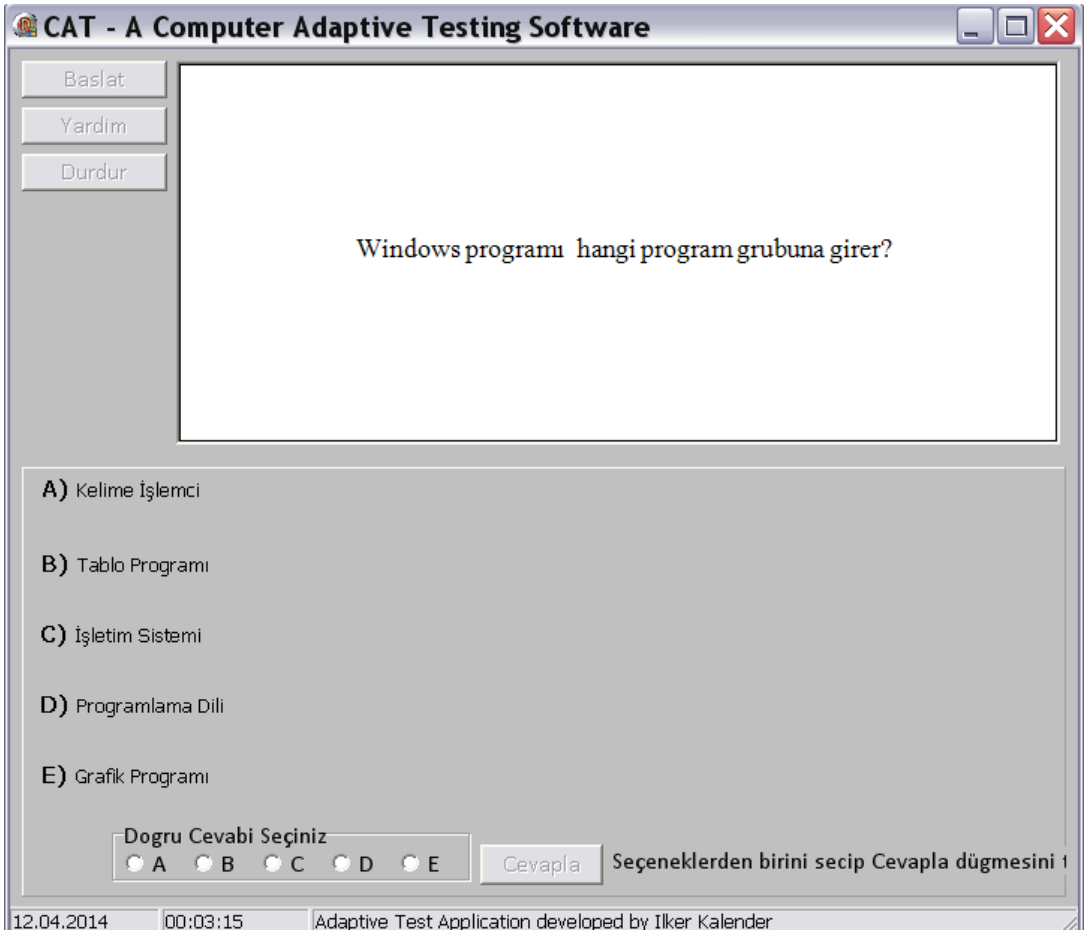
şeklinde tanımlanmaktadır. Babcock ve Weiss (2002) çalışmalarında test durdurma koşulunu belirlerken, güvenirliğin karesini göz önünde bulundurmuşlardır. Bu araştırma kapsamında, klasik test kuramındaki r^2= 0.91 güvenirlik değerine karşılık gelen standart hata değeri 0.30 ve r^2= 0.75 güvenirlik değerine karşılık gelen 0.50 standart hata, test sonlandırma koşullarında kesme değeri olarak alınmıştır. Simülasyon çalışmasında karşılaştırılan BOBUT stratejileri Tablo 7'de özetlenmiştir.

Tablo 7. Araştırmada Uygulanan Simülatif BOBUT Stratejileri

Teste Başlama Kuralı	Yeterlik Kestirim Yöntemi	Sonlandırma Kuralı
-0.5<θ<0.5	EYOY	Sabit test uzunluğu (k=30 madde) Ölçmenin Standart Hatası<0.30 Ölçmenin Standart Hatası<0.50
-0.5<θ<0.5	BSD	Sabit test uzunluğu (k=30 madde) Ölçmenin Standart Hatası<0.30 Ölçmenin Standart Hatası<0.50

Canlı BOBUT Uygulaması

Bilgisayar ortamında bireye uyarlanmış testin gerçek ortamda uygulanabilmesi için Kalender (2011) tarafından geliştirilen BOBUT yazılımı kullanılmıştır. Kağıt-kalem testi araştırmacı tarafından geliştirilen soru bankasına bağlı olarak, belirtke tablosundaki kazanımların dağılımlarına eşit oranda toplamda 30 soru olacak şekilde oluşturulmuştur. Soru yükleme işlemi bittikten sonra bilgisayar ortamında öğrencilerin soruları yanıtlamaları için uygulama yazılımı kullanılmıştır. Kullanılan yazlımın ekran görüntüsü Şekil 7'de verilmiştir.



Şekil 7. Canlı BOBUT Uygulamasında Bir Maddenin Ekran Görüntüsü

Öğrencinin sorulara verdiği bir doğru ve bir yanlış yanıttan sonra yeterlik düzeyi ve standart ölçme hatası hesaplanmaya başlamaktadır. Testin başlangıcında madde havuzunun en kolay beş maddesi arasından tesadüfi olarak biri başlangıç maddesi olarak ekrana gelmektedir. Sonlandırma kuralı ise, sabit soru ve sabit hata koşullarında çalışılmıştır. Sabit hata koşuluna göre testi durdurmada, yeterlik düzeyine bağlı olarak hesaplanan standart hata değeri 0.30 ve bir diğer koşul olarak da 0.50 olarak belirlenmiştir. Eğer 30 madde boyunca standart hata değeri 0.30/0.50 değerine ulaşmazsa oturum sona ermektedir. Yeterlik kestirimleri EYOY yöntemi kullanılarak yapılmıştır.

BULGULAR

Araştırmanın ilk sorusu olan, BOBUT uygulamalarında kullanılan yeterlik kestirim yöntemi ve test sonlandırma kurallarına göre yeterlik kestirimleri yapılmış ve Tablo 8'de bu kestirimlere ait betimsel istatistikler verilmiştir.

Tablo 8. Simülatif BOBUT Uygulamasında Farklı Sonlandırma Kuralı (ölçme hatası ve sabit madde sayısı) ve Yeterlik Kestirim Yöntemlerinden Elde Edilen Yeterlik Kestirimlerine ve Madde Sayılarına Ait Betimsel İstatistikler

Sonlandırma Kuralı	Yeterlik Kestirim Yöntemi	N	Ortalama Yeterlik	Standart sapma	Güvenirlik Ölçüleri	Ortalama Madde Sayısı
SMS (k=30)	BSD	1366	0.01	0.93	14.94	SMS
	EYOY	1366	-0.18	1.05	13.97	SMS
SH<0.30	BSD	1366	-0.00	0.95	12.91	59.8
	EYOY	1366	0.10	0.62	17.83	56.22
SH<0.50	BSD	1366	-0.55	1.03	5.73	11.11
	EYOY	1366	0.01	0.89	7.50	12.21

SH: Standart Hata SMS: Sabit Madde Sayısı

Tablo 8'de görüldüğü gibi, simülatif BOBUT'a dayalı karşılaştırmalar yapmak üzere, ilk olarak madde sayısı 30 olacak şekilde sabitlenmiş ve yeterlik kestirimleri EYOY ve BSD ile ayrı ayrı tekrarlanmıştır. Daha sonra SH<0.30 ve SH<0.50 olarak sabitlendiği koşulda EYOY ve BSD yöntemlerine göre yeterlik kestirimleri ayrı ayrı hesaplanmıştır. Simülatif BOBUT uygulamasında kestirilen yeterlik ölçülerine ilişkin güvenirlik ölçüleri, sabit madde test sonlandırma koşulunda, diğer test sonlandırma koşullarına göre daha yüksek bulunmuştur. Ayrıca kullanılan madde sayısı bakımından, en az madde kullanımı, test sonlandırma koşulunun SH<0.50 olduğu durumda gerçekleşmiştir (simülatif BOBUT uygulaması yeterlik kestiriminde kullanılan madde sayısı EYOY ve SH<50 koşulunda en fazla 101, en az 10; diğer tüm simülatif BOBUT yeterlik kestirim koşullarında kullanılan madde sayısı en az 10, en fazla 136 olarak gerçekleşmiştir).

Simülatif BOBUT uygulamasında elde edilen yeterlik kestirimleri için oluşturulan farklı durumlar sonucunda elde edilen yeterlik ölçüleri arasında anlamlı bir farklılığın olup olmadığını incelemek amacıyla gruplar arasında tek yönlü varyans analizi yapılmış ve Tablo 9'da verilmiştir.

Tablo 9. Farklı Test Sonlandırma Koşullarında (Sabit Madde ve Sabit Hata) Farklı Yeterlik Kestirimi koşullarında (BSD ve EYOY) Tekrarlanan BOBUT Uygulamalarından Kestirilen Yeterlik Ölçülerinin Farklılığına İlişkin Anova Sonuçları

Varyansın Kaynağı	Kareler Toplamı	sd	Kareler Ortalaması	F	p	Anlamlı Fark
Gruplararası	396.57	5	79.31	92.76	.00	1-2; 1-4; 2-3; 2-4; 2-5; 2-6; 3-5; 4-5; 5-6
Gruplariçi	7002.11	8190	0.85			
Toplam	7398.69	8195				

1: BSD ve sabit madde; 2: EYOY ve sabit madde; 3: BSD ve Sabit hata (SH<0.30); 4: EYOY ve sabit hata (SH<0.30); 5: EYOY ve sabit hata (SH<0.50); 6: BSD ve sabit hata (0.50)

Tablo 9'a göre, test sonlandırma kuralı olarak sabit madde ve sabit hata koşullarında farklı kestirim yöntemleriyle elde edilen yeterlik ölçüleri arasında anlamlı bir fark bulunmuştur [F(5-8190)=92.76; p<0.01]. Başka bir ifadeyle, kestirilen yeterlik parametreleri kullanılan BOBUT stratejisine göre değişmektedir. BOBUT uygulama koşullarına göre oluşan bu farklılığın, hangi yeterlik kestirim yönteminin lehine olduğunu tespit etmek amacıyla yeterlik ölçülerine Dunnet C testi uygulanmıştır. Buna göre, BSD ve sabit madde koşulunda uygulanan BOBUT uygulamasında kestirilen yeterlik puanları ortalaması ($\bar{X} = 0.01$), EYOY ve sabit madde stratejine göre kestirilen yeterlik puanları ortalamasından ($\bar{X} = -0.18$) daha yüksek bulunmuştur. Yine BSD ve sabit madde koşulu ile kestirilen yeterlik puanları ortalaması ile EYOY ve sabit hata (SH<0.30) koşuluna göre kestirilen ortalama yeterlik ($\bar{X} = 0.10$) değerine göre daha küçük olduğu bulunmuştur. Yapılan bazı araştırmalarda da (Bulut ve Kan, 2012; Wang, Kuo, Tsai ve Laio, 2012) BSD stratejisine göre kestirilen yeterlik puanlarının, kağıt-kalem testinden elde edilen yeterlik puanları ile arasındaki ilişkiler incelenmiş ve oldukça yüksek korelasyonlar elde edilmiştir. Bu çalışmada da BSD stratejisi ile kestirilen yeterlik puanları ortalamasının kağıt-kalem testi puanları ortalamasına ($\bar{X} = 0.00$) yakın olması, bu bulgunun yapılan çalışmalar (Bock ve Misley, 1982; Raiche ve Blais, 2002) ile paralellik gösterdiğini desteklemektedir.

Araştırmanın ikinci alt amacı doğrultusunda, simülatif BOBUT uygulamasına ilişkin farklı koşullarda elde edilen kestirimler sonucunda, EYOY ve SH<0.30 koşulunda elde edilen kestirimlerin güvenirlik ölçülerinin daha yüksek olmasından hareketle, canlı BOBUT uygulamasında test sonlandırma kuralı olarak ölçmenin standart hatasına (SH<0.30) koşulu belirlenmiştir. Ayrıca madde havuzunda yeteri kadar madde bulunmaması durumunda bireye en fazla 30 madde uygulama koşulu da canlı BOBUT uygulama yazılımında sabitlenmiştir. Yeterlik kestirimi olarak da, EYOY seçilmiştir. Böylece 142 öğrenci BİLOKUR testini hem kağıt-kalem hem de BOBUT olarak almıştır. Katılımcıların her iki uygulamadan teste verdikleri yanıtlardan kestirilen birey yeterlik parametreleri karşılaştırıldığında Tablo 10' da özetlenen bulgular elde edilmiştir.

Tablo 10. BOBUT Uygulaması ve Kağıt-Kalem Testinden Elde Edilen Yeterlik Ölçülerine Ait Betimsel Sonuçlar

	Kağıt-Kalem				Canlı BOBUT			
	Ortalama	Ss	En küçük	En yüksek	Ortalama	Ss	En küçük	En yüksek
Yeterlik ölçüleri	-0.06	0.73	-1.76	1.83	-0.42	1.32	-2.85	2.14
Hata	0.61	0.08	0.39	0.94	0.32	0.09	0.14	0.63
Güvenirlik	2.79	0.83	1.13	6.58	12.39	8.54	2.52	51.02

Tablo 10 incelendiğinde, her iki test uygulamasındaki standart hatalar ve güvenirlik değerleri karşılaştırıldığında, BOBUT uygulamasından elde edilen yeterlik kestirimlerine ait ortalama standart hata değeri ile kağıt-kalem ortamında uygulanan testen elde edilen yeterlik değerlerine ilişkin ortalama standart hata değerleri arasında oldukça büyük bir farkın oluştuğu Tablo 10'dan anlaşılmaktadır. Bu durumda, EYOY kestirim yönteminin uygulandığı BOBUT testinde öğrencilerin yeterlik ölçüleri, kâğıt-kalem testinden elde edilenlere göre, özellikle uç değerlerde daha iyi ölçme sonucu vermiş, bir diğer ifadeyle, standart hata değeri açısından da BOBUT uygulamasıyla daha güvenilir yeterlik kestirimleri elde edilmiştir

Canlı BOBUT ve kağıt-kalem testinden elde edilen yeterlik ölçülerinin sıraları arasındaki tutarlılık "Spearman Brown Sıra Farkları" korelasyonu incelenmiştir. Buna göre, her iki uygulamadan elde edilen yeterlik ölçülerinin sıraları arasında pozitif yönde iyi düzeyde (r= 0.70; p<.01) bir korelasyon bulunmuştur. Bu bulguya göre, canlı BOBUT uygulaması ile kağıt-kalem testi uygulamasından elde edilen yeterlik ölçüleri arasında daha yüksek tutarlılığın bulunmamasının nedeni öğrencilerin test

alma davranışlarından (sınav olma motivasyonlarının gerçek sınav ortamındaki kadar yüksek olmaması vb.) kaynaklanmış olabilir. Buna göre, canlı BOBUT uygulaması ile kağıt-kalem testinden elde edilen yeterlik ölçüleri arasında çok büyük bir değişiklik olmamakla birlikte, BOBUT uygulaması ile daha az sayıda madde ile daha güvenilir yeterlik parametreleri elde etmek mümkündür. Araştırmanın bu bulgusu, alan yazında yapılan diğer araştırmalarla (Bulut ve Kan, 2012; Chae ve Diğerleri, 2000; Frick, 1992; İşeri, 2002; Kalender, 2011; Kaptan, 1993; Kezer, 2013; Koşan-Aytuğ, 2013; McDonald, 2002; Miller, 2003; Mills ve Stocking, 1996; Mills ve Steffen, 2000; Öztuna, 2008; Rudner ve Guo, 2011; Scfhaer ve Diğerleri, 1995; Tian ve Diğerleri, 2007; Zitny ve Diğerleri, 2012) tutarlık göstermektedir.

Canlı BOBUT uygulaması ile kağıt-kalem testlerine ilişkin güvenirlik ölçülerini karşılaştırmak için her iki uygulamada da elde yeterlik ölçülerine ilişkin test bilgi (test information) miktarları hesaplanmıştır. Test bilgi miktarları arasında anlamlı bir fark olup olmadığını bulmak amacıyla ilişkili örneklemler için t-testi kullanılmıştır. Yapılan t-testi sonuçları Tablo 11'de verilmiştir.

Tablo 11. Canlı BOBUT ve Kağıt-Kalem Testlerine İlişkin Test Bilgi Miktarlarına Ait t-testi Sonuçları

	N	$\bar{X}$	Ss	Sd	t	p	En Küçük	En Yüksek
Kağıt-Kalem Testi	142	2.79	0.839	143.72	13.31	0.000	1.13	6.58
Canlı BOBUT	142	12.39	8.55				2.52	51.02

Tablo 11'e göre, test bilgi miktarları ortalamalarına ilişkin olarak, Canlı BOBUT uygulaması ile kağıt-kalem testinden elde edilen değerler arasında anlamlı bir farklılık göstermektedir [$t_{(143.72)}$=13.31; p<.05]. Canlı BOBUT uygulamasından elde edilen ortalama test bilgi miktarı ($\bar{X}$=12.39), kağıt-kalem testinden elde edilen ortalama test bilgi miktarından ($\bar{X}$=2.79) oldukça yüksek bulunmuştur. Buna göre, Canlı BOBUT uygulaması sonuçlarından elde edilen test bilgi miktarları, kağıt-kalem testinden elde edilen değerlerden daha yüksektir.

SONUÇLAR ve TARTIŞMA

Bu araştırmada, Ankara Üniversitesi Bilgi ve İletişim Teknolojileri dersinden muaf tutulacakları belirlemek üzere uygulanan BİLOKUR testinin Bilgi ve İletişim Teknolojisi Kavramları modülünün BOBUT olarak uygulanabilirliği çeşitli koşullarda test edilmiştir. Araştırmada bu amaçla, gerçek veriler kullanılarak kurgulanan sabit soru, sabit hata durdurma kurallarına ile iki farklı yeterlik kestirim yöntemine (EYOY ve BSD) göre, post hoc simülasyon çalışmaları gerçekleştirilmiştir.

Araştırmada elde edilen bulgulardan şu sonuçlara ulaşılmıştır: (1)Simülatif BOBUT uygulamasında farklı test sonlandırma koşullarında (sabit madde, SH<0.30 ve SH<0.50) farklı yeterlik kestirim yöntemlerinden (EYOY ve BSD) elde edilen yeterlik ölçüleri karşılaştırıldığında, sabit madde (30) ve SH<0.50 test sonlandırma koşulunda kestirilen yeterlik ölçülerinde BSD daha büyük değerler alırken; SH<0.30 test sonlandırma koşulunda elde edilen yeterlik ölçülerinden EYOY ile kestirilen yeterlik ölçülerinin daha güvenilir olduğu bulunmuştur. Bu bağlamda, farklı test sonlandırma koşullarında EYOY ve BSD yaklaşımlarının farklı güvenirlikte yeterlik kestirimlerinde bulunduğu söylenebilir. Kezer (2014) tarafından yapılan bir araştırmada ise, EYOY ve BSD yaklaşımları arasında bir fark bulmazken, Wang (1997) tarafından yapılan bir araştırmada BSD'nin EYOY'a göre nispeten daha yüksek standart hata barındırabileceği belirtilmiştir. Eggen (2004) ise test başlatma ve sürdürme işleminde en iyi yöntemin EYOY olduğu belirtmiştir. Bu durumun daha çok kullanılan soru bankasının büyüklüğü ve soru bankasındaki sorularının kalitesi ile ilgili olabileceği düşünülmektedir.

Araştırmanın ikinci bölümünde, ilk bölümden elde edilen BOBUT uygulama stratejilerine dayalı olarak, canlı BOBUT uygulaması ile kağıt kalem testinden elde edilen yeterlik ölçümleri karşılaştırılmıştır. Bu iki uygulama formatı arasındaki farklar incelendiğinde ise, özellikle testle ölçülen yeterlik düzeyi $\theta<0$ olduğu durumlarda bu farkların büyüdüğü, $0<\theta<2$ arasında bu farkın azaldığı bulunmuştur.(2) BİLOKUR testi kağıt-kalem testi olarak uygulanması ile BOBUT olarak uygulanması arasında yeterlik kestirimi açısından çok büyük farkın olmadığı sonucuna ulaşılmıştır. (3)Ayrıca, aynı testi canlı BOBUT ve kağıt-kalem alan katılımcıların iki uygulamadan elde edilen yeterlik düzeyleri arasında büyük ölçüde tutarlık bulunmuştur. Bu sonuç, canlı BOBUT uygulaması ile kağıt-kalem testinin kestirimlerinin birbirine yakın olduğunu, öğrencilerin yeterlik ölçümlerindeki sıralarının her iki uygulamada da çok değişmediğini göstermiştir. Bir başka deyişle, BİLOKUR testinin ilgili modülünün BOBUT olarak uygulanabileceğine ilişkin önemli bir kanıt elde edilmiştir. Benzer sonuçlar Kalender (2011) ve Kezer (2014) tarafından yapılan araştırmalarda da bulunmuştur. Adı geçen araştırmalar üniversite öğrencilerinin katılımıyla gerçekleşmiş; kağıt-kalem testi ile canlı BOBUT uygulamalarına ilişkin test puanları arasında arasında pozitif yönde ilişkiler bulmuşlardır. (4) Canlı BOBUT uygulaması ile kağıt kalem testinden elde edilen yeterlik parametrelerine ait test bilgi miktarları karşılaştırıldığında ise, canlı BOBUT uygulamasında, kağıt-kalem testinden elde edilen test bilgi miktarının anlamlı düzeyde yüksek olduğu ve bu nedenle de daha güvenilir yeterlik kestirimleri yaptığı sonucuna ulaşılmıştır. Yurt içi ve yurt dışında BOBUT ile ilgili yapılan gerek simülatif gerekse canlı BOBUT uygulamlarının birçoğunda (Bulut ve Kan, 2012; Chae ve Diğerleri, 2000; Frick, 1992; Kalender, 2011; Kaptan, 1993; Kezer, 2013; Koşan-Aytuğ, 2013; McDonald, 2002; Miller, 2003; Mills ve Steffen, 2000; Öztuna, 2008; Rudner ve Guo, 2011; Thompson & Weiss, 2011;Weiss ve Betz, 1973) benzer sonuçlar elde edilmiştir. Buna göre, BOBUT uygulaması ile elde edilen kestirimlerin kağıt-kalem testi kestirimlerine göre daha güvenilir olduğu söylenebilir.

Bu araştırma elde edilen sonuçlar doğrultusunda yapılacak araştırmalar için çeşitli önerilerde bulunulabilinir: Soru havuzunun büyük olmadığı durumlarda, yeterlik kestirim yöntemi olarak EYOY'un tercih edilmesi daha uygun olacaktır. Çünkü BSD yeterlik kestirim yöntemi, EYOY'a göre daha fazla madde ile kestirim yapmaktadır. Güvenilir yeterlik kestirimi yapmak için, test sonlandırma kuralı olarak standart hata ve yeterlik kestirim yöntemi olarak EYOY'un tercih edilebilir. Araştırmada canlı BOBUT uygulamasında BİLOKUR testinin sadece bir modülü kullanılmıştır. Daha sonra yapılacak çalışmalarda, bilgi ve iletişim teknolojilerine ilişkin modüller için de BOBUT çalışmaları yapılarak, kâğıt-kalem uygulaması ile sonuçları karşılaştırılabilir. Bu araştırmada maddenin kullanım sıklığı (item exposure) kontrol altına alınmamıştır. Yapılacak araştırmalarda daha geniş bir madde havuzu ile bu faktörün kontrol edildiği koşullarda uygulamanın psikometrik özellikleri test edilebilir.

KAYNAKÇA

Babcock, B., & Weiss, J. (2009, June 2). *Termination criteria in computerized adaptive tests: variable-length cats are not biased.* Kasım 12, 2011 tarihinde Realities of CAT Paper Session: http://www.psych.umn.edu/psylabs/catcentral/ adresinden alındı

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), 438-452.

Baykul, Y. (2010). *Eğitimde ve psikolojide ölçme: klasik test teorisi ve uygulaması* (2 b.). Ankara: Pegem Akademi yayınevi.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement, 6, 431–444.

Bulut, O., & Kan, A. (2012). Application of computerized adaptive testing to Entrance Examination for Graduate Studies in Turkey. *Eurasian Journal of Educational Research*(49), 61-80.

Chae, S., Kang, U., Jeon, E., & Linarce, J. M. (2000). *Development of computerized middle school achievement test.* Seuol, South Korea: Komesa Press.

Çıkrıkçı- Demirtaşlı, N. (2002). A study of Raven Standard Progressive Matrices Test's item measures under classical and item response models: an empirical comparison. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 35*(1-2), 71-79.

Çıkrıkçı-Demirtaşlı, N. (1999). Psikometride yeni ufuklar: Bilgisayar ortamında bireye uyarlanmış test. *Türk Psikoloji Bülteni, 5*(13), 31-36.

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

Davis, L.L. & Dodd, B.G. (2005). *Strategies for controlling item exposure in computerized adaptive testing with partial credit model.* Pearson Education Measurement.

De Ayala, R. (2009). *The Theory and practise of item response theory.* New York Kondon: The Guilford Press.

ECDL. (2007). *European Computer Driving Licence 5.0 Müfredat Versiyonu.* Ocak 1, 2012 tarihinde European Computer Driving Licence Foundation: www.ecdl.com adresinden alındı

Eggen, , T. J. (2004). *Contributions to the theory and practice of computerized adaptive testing.* Arnhem, NL.: Omslag: Roel Ottow / Harold Kainama, Druk: Print Partners Ipskamp B.V., Enschede, Citogroep.

Embretson, E., & Reise, P. (2000). *Item response theory for psychologists.* New Jersey: Lawrence ErlbaumAssociates, Publishers Mahwah.

Fan, X. (1998). Item response theory and classical test theory: An Empirical comparison of their ıtem-person statistics. *Educational and Pscyhological Measurement, 58*(3), 357-382.

Frick, T. (1992). Computerized adaptive mastery tests as a expert systems. *Journal of Educational Computing Research, 8*(2), 182-213.

French, B.F. & Thompson, T.D. (2003). The Evaluation of exposure control prosedures for an operational CAT. The annual meeting of the American educational Research Assıciation, Chiccago.

Frick, T. (1992). Computerized adaptive mastery tests as a expert systems. Journal of Educational Computing Research, 8(2), 182-213.

Hambleton, K., & Swaminathan, H. (1985). *Item response theory : Principles and applications.* Hingham, MA, U.S.A. : Distributors for North America, Kluwer Boston.: Kluwer-Nijhoff Pub.

Hambleton, K., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory* (volume 2 b.). Newbury, London.

Han, K. T. (2010, Mart 20). *SimulCAT: Simulation software for computerized adaptive testing [computer program].* Eylül 17, 2013 tarihinde SimulCat: http://www.hantest.net/ adresinden alındı

İşeri, İ. (2002). *Assessment of students' mathematics achievement through computer adaptive testing procedures.* Yayınlanmamış Doktora Tezi. Ankara: ODTÜ Ortaöğretim Fen ve Matematik Alanları Eğitimi Bölümü.

Kalender, İ. (2009). Başarı ve yetenek kestirimlerinde yeni bir yaklaşım: Bilgisayar ortamında bireyselleştirilmiş testler (computerized adaptive tests-CAT). *CITO Egitim Kuram ve Uygulama*(5), 39-48.

Kalender, İ. (2011). *Effects of different computerized adaptive testing strategies on recovery of ability.* Yayınlanmamış Doktora Tezi. Ankara: ODTU.

Kaptan, F. (1993). *Yetenek Kestiriminde adaptive (bireyselleştirilmiş) test uygulaması ile geleneksel kağıt-kalem testi uygulamasının karşılaştırılması.* Yayınlamamış Doktora Tezi. Ankara: Hacettepe Üniversitesi.

Karasar, N. (2011). *Bilimsel araştırma yöntemi* (22 b.). Ankara: Nobel yayın Dağıtım.

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.

Keller, A. (2000). *Ability estimation procedures in computerized adaptive testing.* USA: American Institute of Certified Public Accountants-AICPA Research Concortium-Examination Teams.

Kezer, F. (2013). *Bilgisayar ortamında bireye uyarlanmış test stratejilerinin karşılaştırılması.* Yayınlamamış Doktora Tezi. Ankara: Ankara Üniversitesi Eğitim Bilimleri Enstitüsü.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items forcomputerized adaptive tests. *Applied Measurement in Education,, 2*(4), 359-375.

Kline, P. (2000). *An easy guide to factor anaylsis.* Routledge, Taylor and Franchis groups.

Köklü, N. (1990). *Klasik test teorisinie gore gelistirilen tailored test ile grup testi arasında bir karşılaştırma.* Yayınlanmamış Doktora Tezi. Ankara: Hacettepe Üniversitesi.

Koşan-Aytuğ, M. (2013). *Tıp eğitiminde gelişim sınavı soru bankası oluşturulması ve benzetim verileri ile bilgisayar uyarlamalı test uygulaması.* Yayınlanmamış Doktora Tezi. Ankara: Ankara Üniversitesi Eğitim Bilimleri Enstitüsü.

Lord, F. (1980). *Applications of Item Response Theory to Practical Testing: Problems.* New Jersey: Lawrence Erlbaum.

Lord, F., & Stocking, M. (1988). Item response theory. J. Keeves içinde, *Educational Research, Methodology, And Measurement: An International Handbook* (s. 269-272). NewYork: Pergamon press.

Mcdonald, P. (2002). *Computer adaptive test for measureing personality factors using item response theory.* Unpublished Doctoral Dissertaion. London: The University Western of Ontario.

Mcglohen, M., & Chang, H. H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods, 40*(3), 808-821.

Miller, D. (2003). *Assessment of student achievement: A comparative study of student achievement using paper and pencil assessment and computerized adaptive testing (CAT).* Unpublished Dissertation. Mich: Graduate School of Wayne State University.

Mills, N., & Steffen, M. (2000). The GRE computer adaptive test: operational issues. V. d. Glass içinde, *Computerized adaptive testing: Theory and practice.* Netherlands: Kluwer Academic Publishers.

Mills, N., & Stocking, L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education, 9*(4), 287-304.

Öztuna, D. (2008). *Kas-İskelet sistemi sorunlarının özürlülük değerlendiriminde bilgisayar uyarlamalı test yönteminin uygulanması.* Yayınlanmamış Doktora Tezi. Ankara: Ankara Üniversitesi Sağlık Bilimleri Ensititüsü.

Raîche, G. & Blais, J.-G. (2002). Practical Considerations about expected a posteriori estimation in adaptive testing: Adaptive a priori, adaptive correction for bias, and adaptive integration interval. Communication at the 11th Biennal intenational objective measurement workshop. New Orleans, LO: IOMW. [ERIC DOCUMENT NO ED 464 110]

Reise, P. (1990). A Comparison of item and person fit methods of assessing model data fit in IRT. *Applied Psychological Measurement, 14*(2), 127-137.

Rudner, L. (2002). An Examination of decision-theory adaptive testing procedures. *Paper presented at the annual meeting of the American Educational Research Association, April 1-5.* New Orleans, LA.

Rudner, L. (1998, Kasım). *An On-line, Interactive, Computer Adaptive Testing Tutorial*. Ocak 12, 2012 tarihinde Measurement Resources from EdRes.org: http://edres.org/scripts/cat adresinden alındı

Rudner, L., & Guo, F. (2011). *Computerized adaptive testing for small scale programs and instructional systems.* Graduate Management Admission Council-GMAC.

Scfhaer, A., Steffen, M., Golub-Smith, L., Mills, N., & Durso, R. (1995). *The introduction and comparability of the computer adaptive GRE general Test.* GRE Board Professional Report No 88-08a.

Sheskin, D. (2004). *Handbook of parametric and nonparametric statistical Procedures.* Ewyork, Chapman&Hall/CRC.

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

Stark, S., & Chernyshenko, O. S. (2001). Examining model-data fit using graphical and statistical methods. *16th annual conference for the Society of Industrial and Organizational Psychology.* San Diego, CA.

Stocking, M. (1990). Specifying optimum examinees for item parameter estimation in item response theory. *Psychometrika*(55), 461-475.

Thompson, A., & Weiss, D. (2011). A Framework for the Development of Computerized Adaptive Tests. *Practical Assessment Research & Evaluation, 16*(1), 1-9.

Tian, J.-Q., Miao, D.-M., Zhu, X., & Gong, J.-J. (2007). An ıntroduction to the computerized adaptive testing. *US-China Education Review, 4*(1).

Tomei, A. (2005). *Taxonomy For The Technology Domain.* Hersbe: Information Science Publishing: ISBN 1-59140-526-2.

van Rijn, P., Eggen, T. J., Hemker, B. T., & Sanders, P. F. (2002). Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Applied Psychological Measurement*(26), 393-411.

Wainer, H., Dorans, J., Flaugher, R., Green, F. B., Mislevy, R., Steinberg, L., et al. (1990). *Computerized adaptive testing: A Primer.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Wang, T., Hanson, B., & Lau, C. (1999). Reducing bias in cat trait estimation: A Comparison of approaches. *Applied Psychological Measurement, 23*(3), 263-278.

Weis, J. D., & Betz, E. N. (1973). *Ability measurement: Conventional or adaptive.* Minnesota, USA: Psychometric Methods Program Department of Psychology University of Minnesota.

Weiss, D. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37*(2), 70-84.

Yaşar, M. (1999). *Bireyselleştirilmiş testler üzerine bir çalışma.* Yayınlanmamış Doktora Tezi. Ankara: Hacettepe Üniversitesi.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG (Version 3.0) [computer program]*. Mooresville, IN: Scientific Software.

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

EXTENDED ABSTRACT

Introduction

In parallel with the developments in computer Technologies in today's World, the tests and methods used in test application also develop accordingly. Thanks to the advancements in the World of computer technology, tests have been used as Computer Adaptive Testing (CAT) in education for various purposes (Selection, Placement, diagnosis and etc.) in last 20 years. CAT can be defined as tests which offer different test items for test-takers, which are prepared appropriately choosing items from pre-prepared test pool considering test takers' proficiency levels (Weiss, 2004). To achieve that, a new method is used rather than giving the same items with the same difficulty level to everybody, and in this method, if test-takers find the correct response, the next item gets more difficult, if test taker is wrong, the next item gets easier, which is known as more or less method (Rudner, 1998). Therefore, as the most appropriate item is directed to test takers considering their proficiency levels in CAT applications, a significant amount of decrease in the number of directed items is obtained. Thus, it becomes possible to have more reliable measuring results using less test items (Embretson & Reise, 2000; Mcglohen & Chang, 2008). There is very limited number of empirical studies in literature in Turkey. Relevant preliminary studies investigated the applicability of achievement test as CAT application (Kaptan, 1993; Köklü, 1990), and the following studies focused on a comparison between traditional applications (paper pencil test) and the application of the tests used to select students for higher education institutions and the test aiming to measure proficiency levels in CAT form (Aytuğ-Koşan, 2013; Bulut & Kan, 2012; İşeri, 2002; Kalender, 2011; Kezer, 2013).The purpose of the study is to investigate how applicable computer literacy test is in CAT form under various conditions. To reach that aim, psychometric qualifications of paper-pencil test form and CAT form will be compared, the most appropriate CAT application strategies (maintenance and termination of test) will be found out.

Method

Different proficiency estimate methods and different termination rules were compared to test how applicable computer literacy test was in CAT form. In this study, proficiency estimates were done considering post-hoc simulation method and different proficiency estimate methods (maximum likelihood estimation method-MLE) and expected A Posteriori Method- EAP and test termination rules (Standard error-SE) is equal to SE<0.30 and SE<0.50). For this purpose, SimulCAT software which is a CAT program developed by Han (2010), was used in the study. This research contributes to improvement and development of already existing hypothetic knowledge as well as contributing to application as a basic research model. According to Karasar (2011), fundamental research studies are the ones which contribute to the development of theories, hypothesis formation, test them and discuss the findings of the test. Then, the psychometric characteristics of proficiency levels obtained from live CAT and paper-pencil test were compared to those of a group of students. For this purpose, CAT application software which was developed by Kalender (2011) was used. With this regard, this study seems to be an applied research which aims to improve an already existing study.

The data of computer literacy test was collected from the first year students at various faculties of Ankara University in 2012-2013 and 2013-2014 education years. The study group participated in the research in two phases. In the first phase, 1452 university first year students were given a computer literacy test developed for this study dividing all the items in 6 groups in 2012-2013 academic years. In the last phase of the study, the items of the computer literacy test whose appropriacy to IRT was confirmed and whose parameters were estimated, were implemented on 142 university first year students at Ankara University, Faculty of Educational Science in 2013-2014 academic year in the form of paper-pencil and live CAT test. 142 students participated in the real CAT application, 32 (23%) of them were male, and 110 (77%) of them were female.

The items making up the computer literacy test were administered on 1366 students in about 2-month period, which is quite short in 2012-2013 academic year. As the item pool is very big, the

items were administered in 6 different item/test groups. After the application, item statistics were calculated based on CTT. According to that calculation, most of the items were found to be at average difficulty, few items were found to be easy. When the average item discrimination values of the items were examined, what was found was that it changed between the lowest was 0.33 and the highest was 0.69. In CAT application, using the items scaled based on IRT provides invariant estimates for the individual and item parameters of IRT. However, having these advantages is largely dependent on the compatibility of the used data with the model (Fan, 1998; Hambleton & Swaminathan, 1989) and revealing the evidence of the measurability of psychological feature by the available data (Stark & Chernyshenko, 2001). For this aim, the items left in the item bank were subject to preliminary analysis based on IRT. With this analysis, the following issues were tested, such as unidimensionality, meeting the local independence assumptions, the control of speed test and constancy of item and individual proficiency parameters.

Results and Discussion

According to research findings, a significant difference was found among the proficiency measurements obtained through various estimation methods under the condition of fixed item and fixed error as test termination rule. According to that, the means of proficiency scores estimated in CAT application applied under EAP and fixed item condition was found to be higher than that of proficiency score means estimated depending on MLE and fixed item strategy. The means of proficiency score average estimated under the condition of EAP and fixed item and average proficiency estimated under the condition of MLE and fixed error (SH<0.30) was found to be lower than ($\overline{X} = 0.10$). Some studies investigated the relationship between the proficiency score obtained from paper-pencil test and the proficiency score estimated according to EAP strategy, and quite high correlations were obtained (Bulut & Kan, 2012; Wang, Kuo, Tsai & Laio, 2012). In this study, the fact that the average proficiency score estimated through EAP strategy was close to that of paper-pencil test ($\overline{X} = 0.00$), demonstrates a parallelism with the findings of the studies carried out in the field.

When comparing the item quantities in competency prediction of the real CAT application and paper-and-pencil test, real CAT application provided huge savings. Also, when comparing the reliability values of real CAT and paper-and-pencil test, the reliability values gained from the real CAT application was found significantly higher. Therefore, it was concluded that Computer Literacy test carried out with the CAT application can predict the competency measures more reliably when compared with paper-and-pencil tests. This findings of the study seems to be consistant with those of the other studies in the literature (Bulut & Kan, 2012; Chae, Kang, Jeon and Linarce, 2000; Frick, 1992; İşeri, 2002; Kalender, 2011; Kaptan, 1993; Kezer, 2013; Koşan-Aytuğ, 2013; McDonald, 2002; Miller, 2003; Mills & Steffen, 2000; Mills & Stocking, 1996; Öztuna, 2008; Rudner & Guo, 2011; Scfhaer, Steffen, Golub-Smith, Mills & Durso, 1995; Tian, Miao, Zhu & Gong, 2007).