ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(2);161-177

Bireyselleştirilmiş Bilgisayarlı Sınıflama Testi Kriterlerinin Test Etkililiği ve Ölçme Kesinliği Açısından Karşılaştırılması*

A Comparison of Computerized Adaptive Classification Test Criteria in Terms of Test Efficiency and Measurement Precision

Ceylan GÜNDEĞER** Nuri DOĞAN***

Öz

Bu çalışmada Bireyselleştirilmiş Bilgisayarlı Sınıflama Testleri'nin (BBST) etkililiğinin sınıflama kriterlerine, madde seçme ve yetenek kestirim yöntemlerine göre nasıl değiştiğinin belirlenmesi amaçlanmıştır. Bu amaçla 3 Parametreli Lojistik Model temel alınmış; belirlenen kesme noktası ve etrafında yüksek bilgi verecek şekilde 500 maddelik bir havuz oluşturulmuş; birey yetenekleri (N(0,1)) 3000 kişi üzerinden türetilmiş ve bireylerin madde cevap örüntüleri R yazılımda rasgele türetilmiştir. Sınıflama kriterlerinden Ardışık Olasılık Oran Testi (AOOT), Genelleştirilmiş Olabilirlik Oranı (GOO) ve Güven Aralığı (GA) yöntemleri; yetenek kestirim yöntemlerinden Beklenen Sonsal Dağılım (BSD) ve Ağırlıklandırılmış Olabilirlik Kestirimi (AOK) yöntemleri; madde seçme yöntemlerinden ise kesme noktasında (KN) ve kestirilen yetenek (KY) temelinde Maksimum Fisher Bilgisi (MFB) ve Kullback-Leibler Bilgisi (KLB) yöntemleri çaprazlanarak 48 koşul oluşturulmuştur. R yazılımında yürütülen BBST simülasyonu sonunda, ortalama test uzunluğu (OTU), ortalama sınıflama doğruluğu (OSD), bireylerin gerçek yetenek düzeyleri ile kestirilen yetenek düzeyleri arasındaki korelasyon (r), yanlılık, RMSE ve ortalama mutlak hata (OMH) değerlerinin 25 tekrara ait ortalamaları hesaplanmıştır. Araştırma sonuçlarına göre test etkililiği bakımından GOO ve GA yöntemlerinin; ölçme kesinliği bakımından ise AOOT'nin daha iyi performans gösterdiği; sınıflama kriterlerinin farksızlık bölgesi genişledikçe veya hata düzeyi değeri küçüldükçe test etkililiğinin arttığı; sınıflama kriterlerinin tümünün her koşulda oldukça yüksek düzeyde sınıflama doğruluğuna sahip olduğu belirlenmiştir. Bireylerin gerçek yetenek düzeyleri ile kestirilen yetenek düzeyleri arasındaki korelasyon bakımından BSD ve AOK yetenek kestirim yöntemlerinin her ikisinin de başarılı kestirimlerde bulundukları ancak ölçme kesinliği bakımından BSD'nin daha iyi performans sergilediği; madde seçme yöntemlerinin ise tümünün birbirine benzer çalıştığı ancak MFB-KY'nin tüm bağımlı değişkenler açısından tüm koşullarda daha iyi performans gösterdiği görülmüştür.

Anahtar Kelimeler: bireyselleştirilmiş bilgisayarlı sınıflama testi, sınıflama kriteri, yetenek kestirimi, madde seçme yöntemi, ölçme kesinliği

Abstract

In this study, it was aimed to determine how the efficiency of the Computerized Adaptive Classification Testing (CACT) changes according to classification criteria, item selection and ability estimation methods. For this purpose, a pool of 500 items, which is based on 3 PLM and informs at the arbitrary cut-point and around, has been generated; individual abilities have been generated using normal distribution (N(0,1)) for 3000

*Bu çalışma ilk yazarın, ikinci yazar danışmanlığında tamamladığı "Bireyselleştirilmiş Bilgisayarlı Sınıflama Testi Kriterlerinin Sınıflama Doğruluğu ve Test Uzunluğu Açısından Karşılaştırılması" isimli doktora tezinden üretilmiştir.

**Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, e-posta: cgundeger@gmail.com , ORCID ID: https://orcid.org/0000-0003-3572-1708
***Prof. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, e-posta: nurid@hacettepe.edu.tr , ORCID ID: orcid.org/0000-0001-6274-2016

DOI: 10.21031/epod.401077
Geliş Tarihi: 03.03.2018
Kabul Tarihi: 22.05.2018

individuals and the item response patterns have been generated randomly in R software with the Monte Carlo simulation. As classification criteria, Sequential Probability Ratio Test (SPRT), Generalized Likelihood Ratio (GLR) and Confidence Interval (CI) methods; as ability estimation methods, Expected a Posteriori (EAP) and Weighted Likelihood Estimation (WLE) methods; and as item selection methods, Maximum Fisher Information (MFI) and Kullback-Leibler Information (KLI) methods on the basis of cut-point (CP) and estimated ability (EA) have been crossed and 48 conditions have been investigated. At the end of the CACT simulations in R, the mean values of Average Test Length (ATL), Average Classification Accuracy (ACA), correlation between the true thetas and estimated thetas (r), bias, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for 25 replications have been calculated. According to the results of the study, it has been observed that the GLR and the CI classification criteria perform better in terms of test efficiency, however the SPRT works better in terms of the measurement precision; test efficiency increases as the indifference region of classification criteria expands or the error value decreases; all classification criteria have considerably high level of the classification accuracy in all conditions. It has been concluded that both ability estimation methods have successful estimation results in terms of the correlation between true and estimated thetas (r); whereas the EAP relatively performs better in terms of the measurement precision; and all of the item selection methods work similarly to each other however the MFI-EA performs better for all conditions in terms of all dependent variables.

Keywords: computerized adaptive classification testing, classification criteria, ability estimation, item selection method, measurement precision

GİRİŞ

Bilgi ve iletişim teknolojilerinde yaşanan gelişmeler, bilgiye ulaşmada ve eğitim uygulamalarında sıklıkla kendini göstermektedir. Bu gelişmeler sayesinde bireylerin öğrenme sürecinde, yeteneklerinin-becerilerinin ölçülmesinde ve değerlendirilmesinde birçok değişiklik meydana gelmektedir. Bu değişikliklerden biri Bireyselleştirilmiş Bilgisayarlı Test (BBT; Computerized Adaptive Testing: CAT) uygulamalarıdır. BBT'de iki temel özellikten bahsedilebilir. Bunlardan ilki bireyin bilgisayar ekranında gördüğü maddeyi cevaplaması iken; ikincisi, testin bireyin yetenek düzeyine göre ayarlanmış olmasıdır (McBride, 1985).

BBT uygulamaları ile Madde Tepki Kuramı'nın (MTK) avantajları sayesinde bireylere yetenek düzeylerine uygun maddeler sunulabilmekte; bireyin aldığı test bireyin yetenek düzeyine göre ayarlanarak bireyselleştirilebilmektedir. Böylece BBT ile geleneksel testlere kıyasla daha kısa zamanda, daha az sayıda maddeyle ve yüksek güvenirlik düzeyinde yetenek kestirimi elde edilebilmektedir (Wainer, 2000). Ayrıca BBT ile hızlı puanlama yapılabilmekte ve bireyler sınav sonucunu uygulama sonunda öğrenebilmektedir. Bu nedenlerle özellikle yurt dışında, GRE (Graduate Record Examination), GMAT (Graduate Management Admission Test) gibi sınavlarda BBT'nin tercih edildiği görülmektedir.

Bireylerin yeteneklerini test etme süreci zaman zaman, belirli bir kesme noktasına (ya da birden fazla sayıda kesme noktasına) dayalı olarak bireyleri başarılı-başarısız, geçti-kaldı (veya düşük-orta-yüksek yetenek düzeyi) vb. sınıflara ayırmayı da hedeflemektedir. BBT'nin bir alt dalı olan Bireyselleştirilmiş Bilgisayarlı Sınıflama Testleri (BBST; Computerized Adaptive Classification Testing: CACT) bireyleri iki ya da daha çok kategoriye ayırmayı amaçlar (Weiss, 1982). Aşağıda bu çalışmanın temelini oluşturan BBST hakkında detaylı bilgiye yer verilmiştir.

Bireyselleştirilmiş Bilgisayarlı Sınıflama Testleri (BBST)

BBT uygulamaları genel olarak, *(i)* Tepki modeli; *(ii)* Madde havuzu; *(iii)* Başlama kuralı; *(iv)* Madde seçme yöntemi; *(v)* Yetenek kestirim yöntemi ve *(vi)* Sonlandırma kuralı olmak üzere altı ana bileşenden oluşmaktadır (Weiss ve Kingsbury, 1984). BBST'de ise ilk beş bileşen aynı olmakla beraber sonlandırma kuralı yerine sınıflama kriterleri kullanılmakta ve bu kriterler aslında BBST'nin odak noktasını oluşturmaktadır. Sınıflama kriterleri sayesinde bireylerin başarılı-başarısız vb. şekilde sınıflara ayrılması söz konusu olmaktadır.

BBST uygulamalarındaki ilk aşama hangi modelin kullanılacağıdır. MTK kapsamında model, çoklu puanlanan maddelere dayanan Ardışık Tepki Modeli (Graded Response Model), Kısmi Kredi Modeli (Partial Credit Model) vb. olabildiği gibi ikili puanlanan maddeleri temel alan 1 PLM, 2 PLM veya 3 PLM olabilmektedir. Bu çalışmada 3 PLM temel alınmıştır. Bu modelde maddelerin ayırt edicilik (a) ve güçlük (b) parametreleri değişkenlik gösterdiği gibi maddelere ait şans parametresi (c) de söz konusudur (Hambleton ve Swaminathan, 1985):

İkinci aşamada madde havuzu yer almaktadır. Genellikle başarı testlerinde kullanılan madde havuzu orta güçlükteki maddelerin yanında çok zor ve çok kolay maddeler içerir. Madde güçlükleri ise tekbiçimli (uniform) dağılıma sahiptir. Ölçüt referanslı testlerde ise madde havuzundaki maddelerin kesme noktası etrafında en yüksek bilgiyi verebilecek madde güçlük değerlerine sahip olması beklenir (Boyd, 2003). Bu çalışmada madde havuzu BBST amacına uygun olarak belirlenen kesme noktası ve etrafında yüksek bilgi veren ve testi alan bireylerin yetenek ranjını kapsayacak şekilde oluşturulmuştur.

BBST'nin üçüncü aşaması başlama kuralının, bir başka deyişle BBST'nin nasıl bir maddeyle başlayacağının belirlenmesidir. Eğer test tekrarlı olarak alınabiliyorsa, testi ikinci kez alanların başlama noktası, bir önceki testten kestirilen yetenek düzeyleri olabilir. Bunun dışında ise genellikle popülasyonun ortalaması atanabilir (Thompson, 2007b). Bu çalışmada başlama noktası, tüm veri setleri ve tüm koşullar için $\theta = 0$ olarak belirlenmiştir.

BBST'nin dördüncü aşaması, madde seçme yönteminin belirlenmesidir. BBT'de, Maksimum Fisher Bilgisi (MFB; Maximum Fisher Information: MFI), Kullback-Leibler Bilgisi (KLB; Kullback-Leibler Information: KLI), a-tabakalama (a-stratified) vb. birçok yöntemin tanımlanmış ve çalışılmış olduğu görülmektedir. Bu çalışmada MFB ve KLB incelenmiştir. MFB bilginin tek bir noktada maksimize edilmesini sağlarken (Embretson ve Reise, 2000); KLB θ_0'dan θ_1'e kadar olan bölgedeki bilgiyi değerlendirir (Eggen, 1999; Akt: Thompson, 2007b). Klasik BBT uygulamalarında madde seçilirken, bireyin kestirilen yetenek (KY) düzeyinde en yüksek bilgiyi veren maddenin seçimi söz konusu iken; BBST uygulamalarında bireyin kestirilen yetenek düzeyinde ve bununla birlikte BBT'den farklı olarak kesme noktasında (KN) en yüksek bilgiyi veren maddenin seçimi gündeme gelmektedir. KY ve KN temelli madde seçimi zeki madde seçim yöntemleri (intelligent item selection) olarak adlandırılmıştır (Thompson, 2007b). KN temelli yöntemlerde kesme noktasında en yüksek bilgiyi sağlayan madde seçilirken; KY temelli yöntemlerde kesme puanı dikkate alınmaksızın bireyin kestirilen geçici yetenek düzeyinde maksimum bilgiyi veren madde seçilmektedir. Bu çalışmada, MFB ve KLB madde seçme yöntemleri KY ve KN temelli madde seçimleriyle çaprazlanmış ve kestirilen yetenekte MFB (MFB-KY), kesme noktasında MFB (MFB-KN), kestirilen yetenekte KLB (KLB-KY) ve kesme noktasında KLB (KLB-KN) olmak üzere dört madde seçme yöntemi incelenmiştir.

BBST'nin beşinci bileşeni yeteneğin kestirilmesidir ve bu bileşen son sınıflama kararlarının etkililiği ve uygunluğu bakımından oldukça önemli bir değişkendir (Yang, Poggio ve Glasnapp, 2006). Wang ve Wang'a (2001) göre bu değişken, raporlanan son yetenek kestirimini etkilediği gibi madde seçimi ve test sonlanmasını da etkilemektedir. Maksimum Olabilirlik Kestirimi (MOK; Maximum Likelihood Estimation: MLE), Beklenen Sonsal Dağılım yöntemi (BSD; Expected a Posteriori: EAP), Maksimum Sonsal Dağılım yöntemi (MSD; Maximum a Posteriori: MAP), Owen'ın Bayesci Kestirim yöntemi gibi alanyazında birçok yetenek kestirim yöntemi yer almaktadır (Wang ve Wang, 2001). Bunların dışında MOK'un geliştirilmiş bir versiyonu olan Ağırlıklandırılmış Olabilirlik Kestirimi (AOK; Weighted Likelihood Estimation: WLE) nadiren de olsa çalışmalarda yer almıştır. Alanyazın incelendiğinde BBST araştırmalarında yetenek kestirim yöntemlerinin pek çalışılmadığı; değişken uzunluklu bu testlerde yöntemlerin birbirlerine kıyasla nasıl performans gösterdiklerinin henüz fazla bilinmediği görülmektedir. Bu sebeple çalışmada ortalama madde sayısını azaltması ve hızlı kestirim kestirimler yapabilmesi bakımından BSD yetenek kestirim yöntemi ile MOK'un yanlılığını azaltmak amacıyla Warm (1989) tarafından geliştirilmiş olan ve yetenek kestiriminde olabilirlik fonksiyonunun modunun yerine ortalamasının dikkate alınmasını sağlayan AOK yetenek kestirim yöntemi incelenmiştir.

BBST'nin BBT'ten farkı ve odak noktasını sınıflama kriteri oluşturmaktadır. Geleneksel BBT'deki sonlandırma kurallarından farklı olarak sınıflama kriterleri temelde bir hipotez testi sürecine dayanmaktadır. Hipotezin kabulüne veya reddine karar verme, bireyi sınıflama çabasının sonuç vermesi anlamına gelmektedir. Sınıflama kriterlerine, Wald (1947) tarafından önerilen Ardışık Olasılık Oran Testi (AOOT; Sequential Probability Ratio Test: SPRT), Weiss ve Kingsbury (1984) tarafından önerilen Bireyselleştirilmiş Uzmanlık Testi (BUT; Adaptive Mastery Testing: AMT), van der Linden (1990) tarafından önerilen Bayesci Karar Kuramı (BKK; Bayesian Decision Theory: BDT), AOOT'nin daha genel bir hali olan Genelleştirilmiş Olabilirlik Oranı (GOO; Generalized Likelihood Ratio: GLR) ve Güven Aralığı (GA; Confidence Interval: CI) yöntemleri örnek olarak gösterilebilir. Bu çalışmada AOOT, GOO ve GA sınıflama kriterlerinin etkililiği incelenmiştir.

AOOT'nin altındaki temel felsefe, iki alternatif hipotez altında gözlenen cevap dağılımının olabilirliğini belirleyerek iki hipotezden birinin doğruluğuna karar verilmesidir. Eğer hipotezlerden birinin olabilirliği diğerinden oldukça büyükse bu hipotez kabul edilirken; iki hipotezin olabilirlikleri benzerse birey yeni bir madde alır ve süreç bu şekilde devam eder (Reckase, 1983). GOO, AOOT'nin modifiye edilmiş daha genel bir halidir. AOOT'de kestirilen yetenek ile gruplara atamada kullanılan yetenek düzeyleri arasındaki eşitlik temel alınırken; GOO'da bu değişklenler arasındaki eşitsizlik durumu da dikkate alınmaktadır. Bu iki sınıflama kriterinde de farksızlık bölgesi (indifference region) ismiyle anılan ve hipotezler yazılırken bireyleri gruplara atamada kullanılacak olan başarılı ve başarısız bölgesine koyulan sabit gündeme gelmektedir. Bu çalışmada AOOT ve GOO sınıflama kriterleri için Nydick'in (2013) çalışması göz önünde bulundurularak tolere edilebilir hatalar için farksızlık bölgesi 0,05 ve 0,10 değerleri dikkate alınmıştır. GA sınıflama kriteri ise sınıflama amacını istatistiksel bir kestirim problemi gibi formüle etmektedir (Eggen ve Straetmans, 2000). GA, ölçmenin koşullu standart hatasını dikkate alarak bireyin kestirilen yetenek düzeyi için kestirimin belirlenen güven aralığına göre kesme noktasının hangi tarafına düştüğünü belirleyen bir yöntemdir. Eğer aralık tam olarak kesme puanının üstündeyse birey geçti-başarılı şeklinde; aralık kesme puanının tam olarak altındaysa kaldı-başarısız olarak sınıflanmaktadır. Aralığın kesme puanını içermesi durumunda ise bireye yeni bir madde sunulmaktadır (Thompson, 2007b). Bu çalışmada GA sınıflama kriteri için Eggen ve Straetmans'in (2000) çalışmalarında incelemiş olduğu %70 ve %90 güven aralığı değerleri ele alınmıştır.

Araştırmanın Amacı ve Önemi:

Cheng ve Liou'ya (2000) göre başarılı bir BBT veya BBST uygulamasında *i)* yetenek kestirim yönteminin uygunluğu ve *ii)* madde seçme yönteminin etkililiği oldukça önemlidir. Bu iki bileşenin farklı sınıflama kriterleriyle birlikte nasıl performans gösterdiğinin belirlenmesi; bir başka deyişle sınıflama kriterlerinin, madde seçme ve yetenek kestirim yöntemlerinin farklı koşullar altında sınıflama doğruluğu, test uzunluğu ve ölçme kesinliği bakımından incelenmesi bu çalışmanın temel amacını oluşturmaktadır.

Çalışmanın odak noktası olan sınıflama kriterleri, 0,05 ve 0,10 farksızlık bölgesi değerleri ile Ardışık Olabilirlik Oranı Testi (AOOT) ve Genelleştirilmiş Olabilirlik Oranı (GOO); %70 ve %90 güven aralığı düzeylerini içeren Güven Aralığı (GA) yöntemleridir. Çalışmada incelenen yetenek kestirim yöntemleri Beklenen Sonsal Dağılım (BSD) ve Ağırlıklandırılmış Olabilirlik Kestirim (AOK) yöntemleri; madde seçme yöntemleri ise kestirilen yetenek temelli MFB (MFB-KY), kesme noktası temelli MFB (MFB-KN), kestirilen yetenek temelli KLB (KLB-KY) ve kesme noktası temelli KLB (KLB-KN) şeklinde belirlenmiştir. Buna göre çalışmanın alt problemleri aşağıdaki gibidir.

BBST simülasyonu sonunda:

1. Yetenek kestirim yöntemi BSD olduğunda AOOT sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GOO sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GA sınıflama kriterinin %70 ile %90 güven düzeyleri için sınıflama doğruluğu, test uzunluğu ve ölçme kesinliği madde seçme yöntemlerinden MFB-KY, MFB-KN, KLB-KY ve KLB-KN'ye göre nasıl değişmektedir?

2. Yetenek kestirim yöntemi AOK olduğunda AOOT sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GOO sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GA sınıflama kriterinin %70 ile %90 güven düzeyleri için sınıflama doğruluğu, test uzunluğu ve ölçme kesinliği madde seçme yöntemlerinden MFB-KY, MFB-KN, KLB-KY ve KLB-KN'ye göre nasıl değişmektedir?
3. Sınıflama kriterlerine, madde seçme yöntemlerine ve yetenek kestirim yöntemlerine göre sınıflama doğruluğu, test uzunluğu ve ölçme kesinliği (r, yanlılık, RMSE ve OMH) değerleri nasıl değişmektedir?

Alanyazındaki BBST çalışmaları incelendiğinde, konunun özellikle yurt dışı alanyazında çalışılmış olduğu ve Türkiye'de çalışılmamış olduğu görülmektedir. Yurt dışında yapılan çalışmalar incelendiğinde ise 1980'lerden bugüne BBST ile ilgili oldukça fazla sayıda çalışmaya rastlanmaktadır. Çalışmalar incelendiğinde sadece sınıflama kriterlerinin karşılaştırılmış olduğu araştırmaların yanında (Huebner, 2012; Jiao ve Lau, 2003; Kingsbury ve Weiss, 1980; Nydick, Nozawa ve Zhu, 2012; Nydick, 2013; Reckase, 1983; Spray ve Reckase, 1996; Thompson ve Ro, 2007; Thompson, 2011; Wouda ve Eggen, 2009) sınıflama kriterlerinin madde seçme yöntemleriyle çaprazlanarak ele alındığı çalışmaların da olduğu göze çarpmaktadır (Eggen, 1999; Eggen ve Straetmans, 2000; Lau ve Wang, 1998, 1999; Lin ve Spray, 2000; Spray ve Reckase, 1994; Thompson, 2007a, 2009).

Yurt dışındaki çalışmalar incelendiğinde sınıflama kriterlerinin madde seçme yöntemleri ve yetenek kestirim yöntemleriyle çaprazlandığı ve sonuçların ölçme kesinliği, test uzunluğu ve sınıflama doğruluğu açısından karşılaştırıldığı herhangi bir çalışmaya rastlanmamıştır. Bu açıdan çalışmanın alanyazına katkı sağlayacağı düşünülebilir. Ayrıca teknolojinin gelişmesi ve eğitimin çağa ayak uydurma çabasının bir sonucu olarak ülkemizde de bilgisayarlı sınavlara doğru bir yönelim olduğu görülmektedir. Buna göre yakın zamanda BBST uygulamalarına geçilebilir. Bu noktada çalışmanın uygulayıcılara, sınıflama kriterleri, madde seçme yöntemleri ve yetenek kestirim yöntemleri hakkında bilgi sağlaması beklenmektedir.

YÖNTEM

Bu çalışma, *"... olsa ne olurdu?"* sorusuna cevap arayan bir Monte Carlo simülasyon çalışmasıdır (Dooley, 2002). Çalışmada hem bireylere ait yetenek parametreleri hem de oluşturulan madde havuzlarının parametreleri R ortamında araştırmacı tarafından türetilmiştir (R Core Team, 2013).

Veri Üretimi

Bu çalışmada bireylerin yetenek parametreleri, (-3,+3) yetenek düzeyleri aralığında, ortalaması 0,0 ve standart sapması 1,0 olacak şekilde normal dağılım yardımıyla toplam 3000 kişi üzerinden random türetilmiştir. Çalışmada birey parametreleri gibi madde parametreleri de simülatif veriden oluşmaktadır. Madde havuzu oluşturmada Thompson'ın (2011) araştırması dikkate alınarak madde havuzunun 3 PLM temelinde 500 maddeden oluşması sağlanmıştır. Araştırmada hem kestirilen yetenek (KY) hem de kesme noktası (KN) temelli madde seçme yöntemleri karşılaştırılacağından madde havuzunun belirlenen kesme noktası olan 1,0 ve etrafında yüksek bilgi verecek; -3,+3 yetenek düzeyleri aralığını kapsayacak şekilde oluşturulmasına dikkat edilmiştir. Bu sebeple havuzdaki maddeler, a parametresinin orta ve yüksek değerlerde olabilmesi adına tekbiçimli dağılımdan [0,5; 2,0] aralığından; b parametresinin, Warm'ın (1989) da çalışmasında belirttiği gibi, gerçek uygulamadaki değerlere yakın olabilmesi adına normal dağılımdan ortalaması 1,0 ve standart sapması 1,5 olmak üzere; c parametresi ise yine gerçek bir uygulama düşünülerek normal dağılımdan ortalaması 0,15 ve standart sapması 0,05 olacak şekilde türetilmiştir. Birey parametrelerinin türetilmesi ve madde havuzunun oluşturulmasının ardından bireylerin madde cevap örüntüsü R ortamında rasgele türetilmiş ve BBST simülasyonuna geçilmiştir.

İşlem

Yetenek parametrelerinin türetilmesi ve madde havuzunun oluşturulması aşamalarından sonra, BBST simülasyonu ***6 sınıflama kriteri x 4 madde seçme yöntemi x 2 yetenek kestirimi yöntemi = 48 koşul*** için yazılan döngülerle 25 tekrarla R'da tamamlanmıştır (R Core Team, 2013). Çalışmanın odak noktası olan 6 sınıflama kriteri; 0,05 ve 0,10 farksızlık bölgesi değerleri ile Ardışık Olabilirlik Oranı Testi (AOOT) ve Genelleştirilmiş Olabilirlik Oranı (GOO) ile %70 ve %90 güven aralığı düzeylerini içeren Güven Aralığı (GA) yöntemleridir. Çalışmada incelenen madde seçme yöntemleri, kestirilen yetenek temelli MFB (MFB-KY), kesme noktası temelli MFB (MFB-KN), kestirilen yetenek temelli KLB (KLB-KY) ve kesme noktası temelli KLB (KLB-KN); yetenek kestirim yöntemleri ise Beklenen Sonsal Dağılım (BSD) ve Ağırlıklandırılmış Olabilirlik Kestirim (AOK) yöntemleri şeklinde belirlenmiştir. BBST simülasyonunda Nydick (2014) tarafından yazılan *catirt* paketinden yararlanılmıştır. Simülasyonda tüm koşullarda başlama noktası, yetenek düzeyi 0 olarak belirlenmiş ve her koşul için bu değer sabit tutulmuş; madde seçme yöntemleri, yetenek kestirim yöntemleri ve sınıflama kriterleri çalışmanın amacına uygun şekilde manipüle edilmiştir. Tüm koşullar için 25 tekrar yapılmıştır.

Verilerin Analizi

Veri analizinde bağımlı değişkenler olan ortalama test uzunluğu (OTU), ortalama sınıflama doğruluğu (OSD), ölçme kesinliğine ilişkin değerler (gerçek yetenekler ile kestirilen yetenekler arasındaki ilişki (r) için Pearson korelasyon katsayısı, yanlılık, RMSE, ortalama mutlak hata: OMH) 25 tekrarın ortalaması olacak şekilde araştırmacı tarafından yazılan fonksiyonlarla R'dan çekilmiştir.

Simülasyon sonuçlarından ortalama test uzunluğu *$test_length* koduyla; bireylerin simülasyon sonunda atandıkları sınıflar ise *$cat_categ* koduyla çekilmiştir. Ortalama sınıflama doğruluğunu hesaplayabilmek amacıyla, simülasyondan çekilen sınıflarla bireylerin türetilen gerçek sınıfları arasındaki uyuma Cohen'in Kappa istatistiğiyle bakılmıştır. Cohen (1960) tarafından geliştirilen Kappa istatistiği, iki veya daha fazla gözlemcinin yaptığı değerlendirmeler arasındaki uyuşmayı belirlemek için kullanılır. Bu uyum -1 ile +1 arasında değer alır. Sıfır değeri tesadüfî uyuşmayı, negatif değerler tesadüfî olmaktan daha kötü bir uyuşmayı ve +1 değeri ise mükemmel uyuşmayı temsil eder (Akt: Şencan, 2005).

Yanlılık, BBST simülasyonu sonucu her birey için kestirilen son yetenek düzeyi ile ($\widehat{\theta\imath}$) bireyin gerçek yetenek düzeyi (θ_i) arasındaki ortalama farklılıktır. Yanlılığın formülü aşağıdaki gibidir (Miller ve Miller, 2004):

$$\text{Yanlılık} = \frac{\sum_{i=1}^{n}(\widehat{\theta\imath} - \theta_i)}{n} \quad (1)$$

RMSE, yanlılığa benzer şekilde tüm koşullar için hataların karesinin ortalamasının karekökü şeklinde hesaplanmıştır:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{\theta\imath} - \theta_i)^2}{n}} \quad (2)$$

Ortalama mutlak hata (OMH) ise, yanlılık formülünde de ele alınan bireyin kestirilen son yetenek düzeyinin gerçek yetenek düzeyinden farkının mutlak değer içerisinde verilmesidir:

$$\text{OMH} = \frac{\sum_{i=1}^{n}|\widehat{\theta\imath} - \theta_i|}{n} \quad (3)$$

BULGULAR

Araştırmanın birinci alt probleminde Monte Carlo simülasyonu BBST uygulamasında, yetenek kestirim yöntemi BSD olduğunda, AOOT sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GOO sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GA sınıflama kriterinin %70 ile %90

güven düzeyleri için sınıflama doğruluğu, test uzunluğu ve ölçme kesinliğinin madde seçme yöntemlerinden MFB-KY, MFB-KN, KLB-KY ve KLB-KN'ye göre nasıl değiştiği incelenmiştir. Aşağıda Tablo 1'de bu alt problemde belirtilen koşulları karşılaştırmada kullanılan bağımlı değişkenler olan ortalama test uzunluğu (OTU), ortalama sınıflama doğruluğu (OSD), bireylerin gerçek yetenek düzeyleri ile kestirilen son yetenek düzeyleri arasındaki korelasyon (r), yanlılık, RMSE ve ortalama mutlak hataya (OMH) ait değerler yer almaktadır. Değerlerin tümü her koşul için 25 tekrarın ortalaması alınarak hesaplanmıştır.

Tablo 1'de madde seçme yöntemi fark etmeksizin ortalama test uzunluğu (OTU) bakımından testi sonlandırmada, bir başka deyişle bireyleri sınıflamada, en az madde gerektiren sınıflama kriterinin GA yönteminin %70 düzeyi olduğu; bunu takiben sırasıyla GA yönteminin %90 güven düzeyinin; GOO FB: 0,10 yönteminin; GOO FB: 0,05 yönteminin ve AOOT FB: 0,10 yönteminin geldiği; en fazla madde gerektiren sınıflama kriterinin ise AOOT FB: 0,05 yönteminin olduğu görülmektedir. Bu bulguya göre, madde seçme yöntemi ile sınıflama kriterlerinin farksızlık bölgesi değerleri ya da güven düzeyleri fark etmeksizin, OTU bakımından en iyi performansı GA sınıflama kriteri göstermiş; bunu sırasıyla GOO ve AOOT sınıflama kriterleri takip etmiştir.

Tablo 1'de sınıflama kriterlerinden GOO ve AOOT yöntemlerinin farksızlık bölgesi değerleri küçüldükçe ve GA yönteminin güven düzeyi yükseldikçe bireyleri sınıflamada gerekli ortalama madde sayısı olan OTU değerlerinin arttığı görülmektedir. Araştırmanın bu bulgusu Reckase (1983), Lau ve Wang (1999), Thompson ve Ro (2007) ve Thompson'ın (2011) araştırma sonuçlarıyla örtüşmektedir. Madde seçme yöntemi fark etmeksizin, ortalama sınıflama doğruluğu (OSD) bakımından GA yöntemi dışındaki sınıflama kriterlerinin tümünün bireyleri geçti-kaldı kategorilerine sınıflamada benzer performans gösterdiği ve sınıflama doğruluğunun oldukça yüksek (0,95 ile 0,97 aralığında) olduğu görülmektedir. GA yönteminin %70 güven düzeyinin (0,95) ve %90 güven düzeyinin (0,96) diğer yöntemlere kıyasla daha düşük ancak yine de yüksek bir sınıflama katsayısına sahip olduğu görülmektedir.

Tablo 1'de bireylerin simülasyon öncesi türetilen gerçek yetenek düzeyleri ile BBST simülasyonu sonucu kestirilen son yetenek düzeyleri arasındaki korelasyon (r) bakımından en yüksek ilişkinin hesaplandığı yöntemin, madde seçme yöntemi fark etmeksizin, AOOT FB: 0,05 yöntemi olduğu; bunu takiben sırasıyla AOOT FB: 0,10 yönteminin, GOO yöntemlerinin ve GA %90 yönteminin geldiği; en düşük ilişkinin ise GA %70 yöntemi ile hesaplandığı görülmektedir. Bu korelasyonlar madde seçme yöntemlerine göre ayrı ayrı incelendiğinde en iyi performansı MFB-KY ve KLB-KY yöntemlerinin verdiği görülmektedir. Bu iki yöntem için hesaplanan korelasyonlar 0,86 ile 0,98 değerleri arasında değişmekte; bu durum da gerçek yeteneklerle kestirilen yetenekler arasındaki korelasyonların bu iki madde seçme yöntemi kullanıldığında oldukça yüksek olduğunu göstermektedir. Diğer madde seçme yöntemlerine ait korelasyon değerleri ise MFB-KN için 0,75 ile 0,92 aralığında ve KLB-KN için 0,75 ile 0,91 aralığında değişmektedir. Bu bulguya dayanarak, bireylerin gerçek yetenek düzeyleri ile kestirilen son yetenek düzeyleri arasındaki korelasyon bakımından, kestirilen yetenek (KY) temelli madde seçme yöntemlerinin kesme noktası (KN) temelli madde seçme yöntemlerine kıyasla daha iyi performans gösterdiği yorumu yapılabilir.

Tablo 1'de sınıflama kriterlerinin oldukça düşük yanlılık değerlerine sahip olduğu ve performanslarının madde seçme yöntemlerine göre farklılık gösterdiği görülmektedir. Madde seçme yöntemi MFB-KY olduğunda sınıflama kriterlerinin neredeyse tümünün yansız performans gösterdikleri; sadece GOO FB: 0,10 ve GA %90 yöntemlerinde düşük bir yanlılık değeri hesaplandığı görülmektedir. Buna göre yanlılık bakımından sınıflama kriterlerinin MFB-KY madde seçme yöntemi ile birlikte başarılı bir performans gösterdikleri yorumu yapılabilir. Madde seçme yöntemi MFB-KN, KLB-KY veya KLB-KN olduğunda ise sınıflama kriterlerinini az da olsa yanlı kestirimler yaptığı görülmektedir.

Tablo 1. Yetenek Kestirim Yöntemi BSD için Koşullara Ait OTU, OSD, r, Yanlılık, RMSE ve OMH Değerleri

Koşullar		Bağımlı Değişkenler					
Madde Seçme Yöntemi	**Sınıflama Kriteri**	**OTU**	**OSD**	**r**	**Yanlılık**	**RMSE**	**OMH**
MFB-KY	**AOOT FB: 0,05**	46,52	0,97	0,98	0,000	0,211	0,170
	AOOT FB: 0,10	31,92	0,97	0,96	0,000	0,254	0,204
	GOO FB: 0,05	16,75	0,97	0,90	0,000	0,403	0,322
	GOO FB: 0,10	15,65	0,97	0,90	0,002	0,412	0,330
	GA %70 güven düzeyi	8,14	0,95	0,86	0,000	0,477	0,387
	GA %90 güven düzeyi	12,68	0,97	0,88	0,002	0,443	0,354
MFB-KN	**AOOT FB: 0,05**	47,02	0,97	0,92	0,001	0,346	0,287
	AOOT FB: 0,10	31,98	0,97	0,88	0,002	0,409	0,342
	GOO FB: 0,05	17,16	0,97	0,82	0,002	0,498	0,423
	GOO FB: 0,10	15,85	0,97	0,81	0,002	0,503	0,430
	GA %70 güven düzeyi	7,67	0,95	0,75	0,002	0,589	0,504
	GA %90 güven düzeyi	13,36	0,97	0,79	-0,002	0,537	0,458
KLB-KY	**AOOT FB: 0,05**	46,42	0,97	0,98	0,000	0,212	0,171
	AOOT FB: 0,10	31,81	0,97	0,96	0,001	0,255	0,205
	GOO FB: 0,05	16,71	0,97	0,90	0,000	0,402	0,323
	GOO FB: 0,10	15,66	0,97	0,90	-0,001	0,412	0,330
	GA %70 güven düzeyi	8,12	0,95	0,86	-0,002	0,480	0,389
	GA %90 güven düzeyi	12,53	0,96	0,88	0,000	0,446	0,357
KLB-KN	**AOOT FB: 0,05**	47,24	0,97	0,91	0,001	0,349	0,290
	AOOT FB: 0,10	32,03	0,97	0,88	0,000	0,410	0,343
	GOO FB: 0,05	17,04	0,97	0,82	0,000	0,499	0,425
	GOO FB: 0,10	15,84	0,97	0,81	0,003	0,504	0,430
	GA %70 güven düzeyi	7,63	0,95	0,75	0,000	0,592	0,507
	GA %90 güven düzeyi	13,30	0,97	0,79	-0,001	0,537	0,459

Tablo 1'e göre, yanlılık ile birlikte kestirimin standart hatasını da dikkate alan RMSE ve ortalama mutlak hata (OMH) değerlerine göre, madde seçme yöntemi fark etmeksizin, en iyi performans gösteren sınıflama kriteri AOOT FB: 0,05 yöntemidir. Bunu takiben sırasıyla AOOT FB: 0,10; GOO FB: 0,05; GOO FB: 0,10; GA %90 ve GA %70 yöntemlerinin geldiği görülmektedir. Farksızlık bölgesi ve güven düzeyleri fark etmeksizin RMSE ve OMH bakımından en iyi performansı AOOT yöntemi göstermiştir. AOOT yönteminin ardından GOO yöntemi gelmiş ve göreli olarak en kötü performansı ise GA yöntemi göstermiştir.

Tablo 1'de görülüğü üzere, Monte Carlo simülasyonu BBST uygulamasında yetenek kestirim yöntemi BSD olduğunda, madde seçme yöntemi fark etmeksizin, test etkililiği (ortalama test uzunluğu ve ortalama sınıflama doğruluğu) bakımından GA %70 yönteminin diğerlerine kıyasla oldukça başarılı bir performans gösterdiği ortaya çıkmıştır. Bunu sırasıyla GA %90; GOO FB: 0,10; GOO FB: 0,05; AOOT FB: 0,10 ve AOOT FB: 0,05 sınıflama kriterleri izlemektedir. Bireyler geçti-kaldı kategorilerine GA %70 sınıflama kriteri ile ortalama 8 madde ve ortalama 0,95 sınıflama doğruluğuyla sınıflanabilirken; diğer yöntemlerin tümünde ortalama sınıflama doğruluğu 0,97 olmak üzere GA %90 sınıflama kriteri ile ortalama 14 madde; GOO FB: 0,10 sınıflama kriteri ile ortalama 16 madde; GOO FB: 0,05 sınıflama kriteri ile ortalama 17 madde; AOOT FB: 0,10 sınıflama kriteri ile ortalama 32 madde ve AOOT FB: 0,05 sınıflama kriteri ile de ortalama 47 maddeyle testin sonlanabildiği ve bireylerin kategorilere yerleştirilebildiği görülmüştür. Test etkililiği açısından bakıldığında GA ve GOO sınıflama kriterlerinin AOOT'ye kıyasla başarılı performans gösterdikleri görülmektedir. Bununla birlikte kestirilen ve gerçek yetenek düzeyleri arasındaki ilişki (r), yanlılık, RMSE ve OMH değerleri bakımından AOOT sınıflama kriterinin göreli olarak diğer yöntemlerden daha iyi performans gösterdiği; ancak tüm koşullar içinden bu göreli değerlerin en düşük olduğu MFB-KY madde seçme yönteminin ve AOOT FB: 0,05 sınıflama kriterinin birlikte ele alındığı koşulda, GA %70 sınıflama kriterinin neredeyse 6 katı maddede testin sonlandığı-bireylerin sınıflanabildiği dikkat çekmektedir. Bu noktada dikkat edilmesi gereken bir bulgu olarak, ortalama test uzunluğunun azalmasının mutlak hatayı artırdığı; bir başka deyişle BBST'de daha az sayıda madde kullanıldığında mutlak hata değerinin yükseldiği görülmektedir.

İkinci Alt Probleme İlişkin Bulgular

Araştırmanın ikinci alt probleminde Monte Carlo simülasyonu BBST uygulamasında, yetenek kestirim yöntemi AOK olduğunda, AOOT sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GOO sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GA sınıflama kriterinin %70 ile %90 güven düzeyleri için sınıflama doğruluğu, test uzunluğu ve ölçme kesinliğinin madde seçme yöntemlerinden MFB-KY, MFB-KN, KLB-KY ve KLB-KN'ye göre nasıl değiştiği incelenmiştir. Aşağıda Tablo 2'de bu alt problemde belirtilen koşulları karşılaştırmada kullanılan bağımlı değişkenler olan ortalama test uzunluğu (OTU), ortalama sınıflama doğruluğu (OSD), bireylerin gerçek yetenek düzeyleri ile kestirilen son yetenek düzeyleri arasındaki korelasyon (r), yanlılık, RMSE ve ortalama mutlak hataya (OMH) ait değerler yer almaktadır. Değerlerin tümü her koşul için 25 tekrarın ortalaması alınarak hesaplanmıştır.

Tablo 2'de madde seçme yöntemlerinden MFB-KY ve KLB-KY kullanıldığında ortalama test uzunluğu (OTU) bakımından testi sonlandırmada, bir başka deyişle bireyleri sınıflamada, en az madde gerektiren sınıflama kriterinin GA %70 yönteminin olduğu; bunu takiben sırasıyla GA %90; GOO FB: 0,10; GOO FB: 0,05 ve AOOT FB: 0,10 yöntemlerinin geldiği; en fazla madde gerektiren sınıflama kriterinin ise AOOT FB: 0,05 yönteminin olduğu görülmektedir. Madde seçme yöntemlerinden MFB-KN ve KLB-KN kullanıldığında ise OTU bakımından iyi performans gösteren sınıflama kriterlerinin GA %70; GOO FB: 0,10; GOO FB: 0,05; GA %90; AOOT FB: 0,10 ve AOOT FB: 0,05 yöntemleri şeklinde sıralandığı görülmektedir. Bu bulguya göre OTU bakımından en iyi performansı KY temelli madde seçme yöntemleri kullanıldığında GA sınıflama kriteri göstermiş, bunu sırasıyla GOO ve AOOT sınıflama kriterleri takip etmiş iken; KN temelli madde seçme yöntemlerinde GA yönteminin %90 güven düzeyinin GOO yöntemine kıyasla daha kötü

performans gösterdiği ve yine bireyleri sınıflamada en fazla sayıda madde gerektiren sınıflama kriterinin AOOT olduğu görülmüştür.

Tablo 2. Yetenek Kestirim Yöntemi AOK için Koşullara Ait OTU, OSD, r, Yanlılık, RMSE ve OMH Değerleri

Koşullar		Bağımlı Değişkenler					
Madde Seçme Yöntemi	**Sınıflama Kriteri**	**OTU**	**OSD**	**r**	**Yanlılık**	**RMSE**	**OMH**
MFB-KY	**AOOT FB: 0,05**	46,78	0,97	0,98	-0,004	0,217	0,174
	AOOT FB: 0,10	32,30	0,97	0,96	-0,013	0,270	0,214
	GOO FB: 0,05	16,73	0,97	0,89	0,019	0,444	0,344
	GOO FB: 0,10	15,71	0,97	0,89	0,025	0,455	0,353
	GA %70 güven düzeyi	8,45	0,95	0,86	0,033	0,514	0,409
	GA %90 güven düzeyi	12,57	0,96	0,87	0,024	0,487	0,380
MFB-KN	**AOOT FB: 0,05**	47,01	0,97	0,91	0,123	0,353	0,297
	AOOT FB: 0,10	31,98	0,97	0,87	0,197	0,425	0,364
	GOO FB: 0,05	17,04	0,97	0,81	0,286	0,533	0,469
	GOO FB: 0,10	15,98	0,97	0,80	0,295	0,542	0,477
	GA %70 güven düzeyi	9,38	0,96	0,76	0,355	0,615	0,545
	GA %90 güven düzeyi	17,84	0,97	0,83	0,262	0,514	0,443
KLB-KY	**AOOT FB: 0,05**	46,74	0,97	0,96	-0,001	0,219	0,175
	AOOT FB: 0,10	32,20	0,97	0,96	-0,013	0,270	0,214
	GOO FB: 0,05	16,80	0,97	0,89	0,028	0,450	0,346
	GOO FB: 0,10	15,77	0,97	0,89	0,025	0,463	0,355
	GA %70 güven düzeyi	8,26	0,95	0,85	0,044	0,525	0,416
	GA %90 güven düzeyi	12,58	0,96	0,87	0,026	0,488	0,380
KLB-KN	**AOOT FB: 0,05**	47,22	0,97	0,91	0,127	0,356	0,299
	AOOT FB: 0,10	32,03	0,97	0,87	0,197	0,427	0,366
	GOO FB: 0,05	17,06	0,97	0,81	0,287	0,534	0,470
	GOO FB: 0,10	15,94	0,96	0,80	0,296	0,545	0,478
	GA %70 güven düzeyi	9,32	0,96	0,76	0,353	0,616	0,546
	GA %90 güven düzeyi	17,99	0,97	0,83	0,259	0,512	0,442

Tablo 2'de sınıflama kriterlerinden GOO ve AOOT yöntemlerinin farksızlık bölgesi değerleri küçüldükçe ve GA yönteminin güven düzeyi yükseldikçe bireyleri sınıflamada gerekli ortalama madde sayısı olan OTU değerlerinin arttığı görülmektedir. Araştırmanın bu bulgusu Reckase (1983), Lau ve Wang (1999), Thompson ve Ro (2007) ve Thompson'ın (2011) araştırma sonuçlarıyla örtüşmektedir. Madde seçme yöntemi fark etmeksizin ortalama sınıflama doğruluğu (OSD) bakımından sınıflama kriterlerinin tümünün bireyleri geçti-kaldı kategorilerine sınıflamada benzer performans gösterdiği ve sınıflama doğruluğunun 0,95 ile 0,97 aralığında oldukça yüksek olduğu görülmektedir. Madde seçme yöntemlerinden MFB-KY veya KLB-KY kullanıldığında sınıflama kriterlerinden GA %70 ve GA %90 yöntemlerinin; MFB-KN kullanıldığında GA sınıflama kriterinin %70 güven düzeyinin; KLB-KN kullanıldığında ise GA %70 yöntemi ile GOO FB: 0,10 yönteminin diğer sınıflama kriterlerine kıyasla düşük OSD değerlerine sahip olsalar da sınıflama doğruluklarının yüksek olduğu görülmüştür.

Tablo 2'de gerçek yetenek düzeyleri ile kestirilen yetenek düzeyleri arasındaki korelasyon (r) bakımından en yüksek ilişkinin hesaplandığı yöntemin MFB-KY yöntemi kullanıldığında AOOT FB: 0,05 yöntemi olduğu; bunu takiben sırasıyla AOOT FB: 0,10 yönteminin, GOO yöntemlerinin ve GA %90 yönteminin geldiği; en düşük ilişkinin ise GA %70 yöntemi ile hesaplandığı görülmektedir. Tüm sınıflama kriterlerine ait r değerleri KY temelli yöntemler için 0,85 ile 0,98 aralığında ve KN temelli yöntemler için de 0,76 ile 0,91 aralığındadır. Bu bulguya göre r bakımından KY temelli madde seçme yöntemlerinin KN temelli madde seçme yöntemlerine kıyasla daha iyi performans gösterdiği ortaya koyulmuştur.

Tablo 2'de sınıflama kriterlerinin Tablo 1'e kıyasla daha yüksek yanlılık değerlerine sahip olduğu; bir başka deyişle yetenek kestirim yöntemi olarak BSD yerine AOK kullanıldığında yanlılık değerlerinin yükseldiği ve yöntemlerin performanslarının madde seçme yöntemlerine göre farklılık gösterdiği görülmektedir. Madde seçme yöntemi fark etmeksizin AOOT sınıflama kriterinin en düşük yanlılık değerlerine sahip olduğu görülmektedir. Bununla birlikte GA yönteminin %90 güven düzeyinin, madde seçme yöntemi MFB-KY veya KLB-KY olduğunda GOO FB: 0,10 yönteminden daha düşük yanlılık değerine; madde seçme yöntemi MFB-KN veya KLB-KN olduğunda ise GOO yönteminin her iki farksızlık bölgesine kıyasla daha az yanlılığa sahip olduğu görülmektedir. Buna dayanarak GA %90 sınıflama kriterinin madde seçme yöntemleriyle birlikte yanlılık bakımından iyi sonuçlar verdiği ve özellikle KN temelli madde seçme yöntemleriyle ele alındığında oldukça düşük yanlılık değeri verdiği söylenebilir.

Tablo 2'de yanlılık ile birlikte kestirimin standart hatasını da dikkate alan RMSE ve ortalama mutlak hata (OMH) değerlerine göre madde seçme yöntemi fark etmeksizin, en iyi performans gösteren sınıflama kriteri AOOT FB: 0,05 yöntemidir. Bununla birlikte GA yönteminin %90 güven düzeyinin, madde seçme yöntemi MFB-KN veya KLB-KN olduğunda, GOO yönteminin her iki farksızlık bölgesine kıyasla daha düşük RMSE ve OMH değerine sahip olduğu görülmektedir. Buna dayanarak GA %90 sınıflama kriterinin özellikle KN temelli madde seçme yöntemleriyle ele alındığında oldukça başarılı performans gösterdiği söylenebilir.

Tablo 2'de görülüğü üzere BBST simülasyonunda yetenek kestirim yöntemi AOK olduğunda, madde seçme yöntemi fark etmeksizin test etkililiği (ortalama test uzunluğu ve ortalama sınıflama doğruluğu) bakımından GA %70 yönteminin diğerlerine kıyasla oldukça başarılı bir performans gösterdiği görülmektedir. Bunu sırasıyla KY temelli madde seçme yöntemlerinde GA %90; GOO FB: 0,10; GOO FB: 0,05; AOOT FB: 0,10 ve AOOT FB: 0,05 ve KN temelli madde seçme yöntemlerinde ise GOO FB: 0,10; GOO FB: 0,05; GA %90; AOOT FB: 0,10 ve AOOT FB: 0,05 sınıflama kriterleri izlemektedir.

Tablo 2'de bireyler geçti-kaldı kategorilerine GA %70 sınıflama kriteri ile ortalama 9 madde ve ortalama 0,96 sınıflama doğruluğuyla sınıflanabilirken; diğer yöntemlerin tümünde ortalama sınıflama doğruluğu 0,96 ya da 0,97 olmak üzere GA %90 sınıflama kriteri ile ortalama 16 madde; GOO FB: 0,10 sınıflama kriteri ile ortalama 16 madde; GOO FB: 0,05 sınıflama kriteri ile ortalama 17 madde; AOOT FB: 0,10 sınıflama kriteri ile ortalama 32 madde ve AOOT FB: 0,05 sınıflama

kriteri ile de ortalama 47 maddeyle testin sonlanabildiği ve bireylerin kategorilere yerleştirilebildiği görülmüştür. Test etkililiği açısından bakıldığında GA ve GOO sınıflama kriterlerinin AOOT'ye kıyasla başarılı performans gösterdikleri söylenebilir. Bununla birlikte kestirilen ve gerçek yetenek düzeyleri arasındaki ilişki (r), yanlılık, RMSE ve OMH değerleri bakımından AOOT sınıflama kriterinin göreli olarak diğer yöntemlerden daha iyi performans gösterdiği; ancak tüm koşullar içinden bu göreli değerlerin en düşük olduğu KLB-KY madde seçme yönteminin ve AOOT FB: 0,05 sınıflama kriterinin birlikte ele alındığı koşulda, GA %70 yönteminin neredeyse 6 katı maddede testin sonlandığı-bireylerin sınıflanabildiği dikkat çekmektedir. Bu noktada dikkat edilmesi gereken bir bulgu olarak, ortalama test uzunluğunun azalmasının mutlak hatayı artırdığı; bir başka deyişle BBST'de daha az sayıda madde kullanıldığında mutlak hata değerinin yükseldiği görülmektedir.

Üçüncü Alt Probleme İlişkin Bulgular

Araştırmanın üçüncü alt probleminde sınıflama kriterlerinin, madde seçme yöntemlerinin ve yetenek kestirim yöntemlerinin ayrı ayrı olmak üzere ortalama test uzunluğu (OTU), ortalama sınıflama doğruluğu (OSD), korelasyon (r), yanlılık, RMSE ve ortalama mutlak hata (OMH) bakımından nasıl değiştikleri incelenmiştir. Aşağıda Tablo 3'te tüm bağımsız değişkenlere ilişkin bilgiler özetlenmiştir.

Tablo 3. Sınıflama Kriterlerinin, Madde Seçme Yöntemlerinin ve Yetenek Kestirim Yöntemlerinin OTU, OSD, r, Yanlılık, RMSE ve OMH Değerleri

Bağımsız Değişken		*OTU*	*OSD*	*r*	*Yanlılık*	*RMSE*	*OMH*
Sınıflama Kriterleri	AOOT	39,45	0,97	0,93	0,039	0,311	0,257
	GOO	16,36	0,97	0,85	0,079	0,475	0,394
	GA	11,24	0,96	0,82	0,085	0,523	0,436
Madde Seçme Yöntemleri	MFB	22,36	0,97	0,87	0,067	0,436	0,362
	KLB	22,35	0,97	0,87	0,068	0,438	0,363
	KY	22,00	0,97	0,91	0,008	0,384	0,304
	KN	22,71	0,97	0,83	0,127	0,490	0,421
	MFB-KY	22,02	0,97	0,91	0,007	0,382	0,303
	MFB-KN	22,69	0,97	0,83	0,127	0,489	0,420
	KLB-KY	21,97	0,97	0,91	0,009	0,385	0,305
	KLB-KN	22,72	0,97	0,83	0,127	0,490	0,421
Yetenek Kestirim Yöntemleri	BSD	22,04	0,97	0,87	0,001	0,424	0,352
	AOK	22,65	0,97	0,87	0,135	0,449	0,373

Tablo 3'e göre sınıflama kriterlerinden, yöntemlerin farklı hata düzeyleri veya farksızlık bölgesi değerleri dikkate alınmaksızın, en az sayıda maddeyle sınıflama yapabilen yöntemin yaklaşık 11 maddeyle GA yöntemi olduğu; bunu 16 maddeyle GOO yönteminin takip ettiği ve en çok sayıda maddeyle sınıflama yapabilen yöntemin de 40 maddeyle AOOT olduğu görülmektedir. Sınıflama kriterlerinin ortalama sınıflama doğruluğu bakımından ise benzer sonuçlar verdikleri görülmektedir. Bu bulgu, test etkililiği bakımından GA ve GOO yöntemlerinin BBST uygulamalarında daha kullanışlı olacağına işaret etmektedir.

Tablo 3'te bireylerin türetilen gerçek yetenek düzeyleri ile BBST uygulaması sonucu kestirilen son yetenek düzeyleri arasındaki korelasyon (r) bakımından sınıflama kriterlerinin üçünün de iyi performans gösterdikleri görülmektedir. Korelasyonlar bakımından en iyi sonucu AOOT sınıflama kriteri verirken; GOO ve GA yöntemlerinin benzer şekilde çalıştıkları görülmüştür. Bu sonuç test etkililiği ile birlikte düşünüldüğünde, GOO ve GA sınıflama kriterlerinin uygulamada avantaj sağlayacağı yorumu yapılabilir. Yanlılık, RMSE ve OMH bakımından GA sınıflama kriterinin diğer iki yönteme kıyasla daha kötü performans gösterdiği; en iyi performansı ise AOOT sınıflama kriterinin sergilediği görülmektedir. Bu sonuçlar test etkililiği ile birlikte düşünüldüğünde, AOOT sınıflama kriteri sonuçlarının hatadan daha arınık olmasına karşın; bu yöntemin test etkililiği bakımından kullanışlı olmadığı dikkat çekmektedir.

BBST uygulamalarından beklenen, bireyleri az sayıda maddeyle yüksek doğrulukta sınıflamaktır. Bu açıdan bakıldığında, alt problem için elde edilen tüm sonuçları düşünerek, GOO sınıflama kriterinin diğer iki yönteme kıyasla daha kullanışlı bir seçenek olduğu söylenebilir. Bu alt problem için elde edilen bulgular, Thompson (2011) ile Nydick, Nozawa ve Zhu'nun (2012) çalışmalarının sonuçları ile örtüşmektedir. Bahsedilen çalışmalarda GOO'nun test etkililiği bakımından en kullanışlı sınıflama kriteri olduğu ortaya çıkmıştır.

Tablo 3'e göre, madde seçme yönteminin hangi ölçütü temele aldığı fark etmeksizin, MFB ve KLB madde seçme yöntemlerinin bağımlı değişkenler bakımından birbirine oldukça benzer performans gösterdiği görülmektedir. Bu bulgu Eggen (1999), Lau ve Wang (1999), Cheng ve Liou (2000) ile Lin ve Spray'in (2000) çalışma sonuçlarına benzerlik göstermekte iken; Spray ve Reckase (1994) ile Lau ve Wang'ın (1998) araştırma sonuçlarıyla örtüşmemektedir. Alanyazın incelendiğinde bu iki madde seçme yönteminin performansları hakkında araştırmacılar tarafından fikir birliğinin sağlanamadığı görülmektedir.

Madde seçme yönteminin hangi temele dayandığı incelendiğinde ise kestirilen yeteneğe dayanan (KY) madde seçiminin kesme noktasına dayanan (KN) madde seçimine kıyasla bağımlı değişkenler bakımından daha iyi performans sergilediği görülmektedir. KY ve KN temelli yöntemler ortalama test uzunluğu ve ortalama sınıflama doğruluğu bakımından benzer sonuçlar vermiş olsa da gerçek yetenek düzeyleriyle kestirilen son yetenek düzeyleri arasındaki korelasyon, yanlılık, RMSE ve OMH değerleri incelendiğinde KY temelli madde seçme yönteminin daha başarılı olduğu görülmektedir. Alanyazında bu karşılaştırmaya az sayıda çalışmada yer verilmiş ve bu araştırmalarda da yöntemlerin etkililiği hakkında ortak bir sonuca ulaşılamamıştır. Örneğin Spray ve Reckase (1994) çalışmasında kesme noktasında (KN) en yüksek bilgiyi veren madde seçme yöntemiyle daha kısa test oluştuğunu gösterirken; Thompson (2007b, 2009) araştırmalarında tam aksini, geçici yetenek düzeyinde (KY) en yüksek bilgi veren maddenin seçilmesi durumunda testin kısaldığını ortaya koymuştur. Bu açıdan çalışmanın bu bulgusu Thompson'ın (2007b, 2009) araştırma sonuçlarıyla örtüşmektedir.

Tablo 3'teki dört madde seçme yöntemi incelendiğinde, madde seçme yöntemlerinden en iyi performansı MFB-KY'nin sergilediği; bunu KLB-KY'nin takip ettiği ve MFB-KN ile KLB-KN'nin ise benzer performans gösterdiği görülmektedir. Bu bulgu Eggen ve Straetmans'ın (2000) çalışma sonuçlarıyla örtüşmektedir.

Tablo 3'e göre BSD ve AOK yetenek kestirim yöntemlerinin OTU, OSD ve r bakımından oldukça benzer çalıştıkları ancak yanlılık, RMSE ve OMH değerleri bakımından BSD'nin AOK'a kıyasla göreli olarak daha iyi performans sergilediği görülmektedir. Bu bulgu Wang, Hanson ve Lau'nun (1999) çalışma sonuçlarıyla AOK'un değişken uzunluklu testlerde yüksek yanlılık değerine sahip olması bakımından örtüşmektedir. Ayrıca Yi, Wang ve Ban (2000) araştırmalarında değişken uzunluklu testlerde AOK'un BSD'den daha fazla sayıda madde gerektirdiği sonucuna ulaşmışlardır. Tablo 3'te BSD ile test ortalama 22 maddede sonlanırken; yetenek kestirim yöntemi AOK olduğunda bu ortalama 23'e çıkmaktadır. Çalışmanın bu bulgusu da alanyazın ile örtüşmektedir.

SONUÇLAR ve TARTIŞMA

Bu araştırmada, BBST uygulamalarındaki sınıflama kriterleri, yetenek kestirim yöntemleri ve madde seçme yöntemleri Monte Carlo simülasyonu altında incelenmiştir. Araştırma sonucunda, tüm koşullarda, madde seçme yöntemi ve yetenek kestirim yöntemi fark etmeksizin, bireyleri sınıflamada en az sayıda madde gerektiren sınıflama kriterlerinin sırasıyla Güven Aralığı (GA) yöntemi, Genelleştirilmiş Olabilirlik Oranı (GOO) ve Ardışık Olasılık Oran Testi (AOOT) olduğu; AOOT ve GOO sınıflama kriterleri için farksızlık bölgesi genişledikçe ve GA sınıflama kriteri için ise hata düzeyi değeri küçüldükçe ortalama test uzunluğunun azaldığı; sınıflama kriterlerinden elde edilen ortalama sınıflama doğruluklarının birbirine yakın değerler aldığı ve bu değerlerin oldukça yüksek düzeyde sınıflama doğruluğuna işaret ettiği görülmüştür. Bu sonuçlar bir arada düşünüldüğünde test etkililiği (test uzunluğu ve sınıflama doğruluğu) bakımından gerçek uygulamalarda, daha az sayıda maddeyle yüksek doğrulukta sınıflama yapabilmeleri sebebiyle, GA ve GOO sınıflama kriterlerinin tercih edilmesi önerilmektedir.

Çalışma sonuçlarına göre daha az sayıda maddeyle testin sonlanabilmesi, bireyin ait olduğu kestirilen kategoriye daha kısa zamanda atanabilmesi için farksızlık bölgesinin geniş tutulması (örneğin 0,05 yerine 0,10 değerinin alınması) veya güven aralığı değerinin daha küçük (örneğin %90 yerine %70 olarak) belirlenmesi gerekmektedir. Farksızlık bölgesi daraldıkça veya güven aralığı değeri yükseldikçe bireyin bir kategoriye atanması zorlaşmakta, daha fazla sayıda maddeye ihtiyaç duyulmakta, bu da test etkililiğini düşürmektedir.

Çalışmanın bir diğer sonucu, sınıflama kriterlerinin kestirilen yetenekler ile gerçek yetenekler arasındaki korelasyonlar (r) bakımından yüksek düzeyde ilişki verdiği; buna dayanarak sınıflama kriterlerinin BSD veya AOK ile birlikte yetenek kestiriminde başarılı olduğu; sınıflama kriterlerinden, madde seçme yöntemi ve yetenek kestirim yöntemi fark etmeksizin, yanlılık, RMSE ve ortalama mutlak hata bakımından göreli olarak en iyi performansı AOOT yönteminin gösterdiği; bunu GOO yönteminin takip ettiği ve en kötü performansı ise GA yönteminin gösterdiği belirlenmiştir. Bu noktada dikkati çeken bir durum, daha az sayıda maddeyle testin sonlanmasının mutlak hatayı artırmasıdır. Bu sonuçlara dayanarak ölçme kesinliği bakımından en iyi performansı gösteren sınıflama kriterlerinin AOOT ve GOO yöntemleri olduğu görülmüştür. Bir önceki paragraftaki sonuçlara dayanarak test etkililiği, ölçme kesinliği ile bir arada düşünüldüğünde GOO yönteminin diğer iki sınıflama kriterine kıyasla daha başarılı performans sergilediği görülmüştür ve bu sebeple de uygulayıcılara gerçek uygulamalarda GOO sınıflama kriterinin kullanılmasının daha uygun olacağı önerilmektedir.

Çalışmada incelenen madde seçme yöntemleri olan MFB ve KLB'nin, sınıflama kriteri ve yetenek kestirimi yöntemi fark etmeksizin, ortalama test uzunluğu, ortalama sınıflama doğruluğu, kestirilen yetenek ile gerçek yetenek düzeyi arasındaki korelasyon, yanlılık, RMSE ve ortalama mutlak hata bakımından birbirine oldukça benzer çalıştıkları; madde seçme yöntemlerinin dayandığı temel bakımından kestirilen yetenek (KY) ve kesme noktası (KN) temelli yöntemlerin ortalama test uzunluğu ve ortalama sınıflama doğruluğu bakımından birbirine benzer performans gösterdikleri ancak kestirilen yetenek ile gerçek yetenek düzeyi arasındaki korelasyon, yanlılık, RMSE ve ortalama mutlak hata bakımından KY'nin KN'ye kıyasla daha iyi performans gösterdiği; madde seçme yöntemlerinden MFB-KY yönteminin ortalama test uzunluğu, ortalama sınıflama doğruluğu, kestirilen yetenek ile gerçek yetenek düzeyi arasındaki korelasyon, yanlılık, RMSE ve ortalama mutlak hata bakımından diğer yöntemlere kıyasla daha iyi performans gösterdiği belirlenmiştir. Bu sonuçlar bir arada düşünüldüğünde uygulayıcılara, test etkililiği ve ölçme kesinliği bakımından daha iyi performans göstermiş olması sebebiyle, MFB-KY'nin tercih edilmesinin uygun olacağı düşünülmektedir.

Yetenek kestirim yöntemlerinin ortalama test uzunluğu, ortalama sınıflama doğruluğu ve kestirilen yetenek ile gerçek yetenek düzeyi arasındaki korelasyon açısından benzer çalıştıkları ancak BSD'nin çok düşük yanlılık değerine; göreli olarak daha küçük RMSE ve ortalama mutlak hata değerine sahip olduğu görülmüştür. Bu sonuçla birlikte değişken uzunluklu BBST uygulamalarında BSD'nin AOK'a kıyasla daha iyi bir kestirici olduğu görülmüştür. Bu sonuca dayanarak da gerçek uygulamalarda test etkililiği ve ölçme kesinliği bakımından yetenek kestirim yöntemlerinden BSD'nin tercih edilmesi önerilmektedir.

KAYNAKÇA

Boyd, A. M. (2003). Strategies for controlling testlet exposure rates in computerized adaptive testing systems. (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database. (UMI No. 3110732)

Cheng, P. E. & Liou, M. (2000). Estimation of trait level in computerized adaptive testing. *Applied Psychological Measurement, 24*(3), 257–265

Dooley, K. (2002). Simulation research methods. In J. Baum (Ed.). Companion to organizations. London: Blackwell.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*(3), 249-261

Eggen, T. J. H. M. & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60*(5), 713-734

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologist*. London: Lawrence Erlbaum Associates Publishers

Hambleton, R. K. & Swaminathan, H. (1985*). Item response theory: principles and applications*. Boston: Kluwer Nijhoff Publishing

Huebner, A. (2012). Item overexposure in computerized classification tests using sequential item selection. *Practical Assessment, Research & Evaluation, 17(12), 1-9.*

Jiao, H. & Lau, A. C. (2003). The Effects of Model Misfit in Computerized Classification Test. The annual meeting of the National Council of Educational Measurement. Chicago, IL, April 2003. [Online: http://iacat.org/sites/default/files/biblio/ji03-01.pdf , Accessed date: 17.5.2018.]

Kingsbury, G. G. & Weiss, D. J. (1980). A Comparison of Adaptive, Sequential and Conventional Testing Strategies for Mastery Decisions. Research Report 80-4. [Online: http://iacat.org/sites/default/files/biblio/ki80-04.pdf , Accessed date: 17.5.2018.]

Lau, C. A. & Wang, T. (1998, April). Comparing and combining dichotomous and polytomous items with SPRT procedure in computerized classification testing. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Lau, C. A. & Wang, T. (1999, April). Computerized classification testing under practical constraints with a polytomous model. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Lin, C. J. & Spray, J. (2000). Effects of item-selection criteria on classification testing with the sequential probability ratio test. ACT Research Report Series 2000-8. [Online: https://eric.ed.gov/?id=ED445066, Accessed date: 17.5.2018.]

McBride, J. R. (1985). Computerized adaptive testing. *Educational Leadership, 43*(2), 25 -28

Miller, I. & Miller, M. (2004). *John E. Freund's mathematical statistics with applications*. New Jersey: Prentice Hall

Nydick, S. W., Nozawa, Y. & Zhu, R. (2012, April). Accuracy and efficiency in classifying examinees using computerized adaptive tests: an application to a large scale test. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Nydick, S. W. (2013). Multidimensional mastery testing with CAT. (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database. (UMI No. 3607925)

Nydick, S. W. (2014). *catirt: An R Package for Simulating IRT-Based Computerized Adaptive Tests*. [Online: https://cran.r-project.org/web/packages/catIrt/catIrt.pdf, Accessed date: 17.5.2018.]

R Core Team. (2013). *R: A language and environment for statistical computing*, (Version 3.0.1), Vienna, Austria: R Foundation for Statistical Computing. Online: http://www.R-project.org/

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.). New horizons in testing: latent trait theory and computerized adaptive testing. New York: Academic Press.

Spray, J. A. & Reckase, M. D. (1994, April). The selection of test items for decision making with a computer adaptive test. Paper presented at the annual meeting of the National Council on Measurement in Education, NewOrleans, LA.

Spray, J. A. & Reckase, M. D. (1996). Comparison of SPRT and sequential bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21(4),* 405-414.

Şencan, H. (2005). *Sosyal ve davranışsal ölçümlerde güvenirlilik ve geçerlilik*. Ankara: Seçkin Yayıncılık.

Thompson, N. A. & Ro, S. (2007). Computerized classification testing with composite hypotheses. In D. J. Weiss (Ed.). Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing. Accessed date: [17.5.2018] from http://iacat.org/sites/default/files/biblio/cat07nthompson.pdf

Thompson, N. A. (2007a). *A comparison of two methods of polytomous Computerized classification testing for multiple cutscores*. (Unpublished Doctoral Dissertation). University of Minnesota.

Thompson, N. A. (2007b). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment Research & Evaluation, 12(1)*, 1-13

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69(5)*, 778-793

Thompson, N. A. (2011). Termination criteria for computerized classification testing. *Practical assessment, Research & Evaluation, 16(4)*, 1-7

van der Linden, W. J. (1990). Applications of decision theory to test-based decision making. In R. K. Hambleton & J. N. Zaal (Eds.). Advances in educational and psychological measurement. Massachusetts: Kluwer-Nijhof.

Wainer, H. (2000). *Computerized adaptive testing: a primer*. New Jersey: Lawrence Erlbaum Associates

Wald, A. (1947). *Sequential analysis*. New York: John Wiley

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

Wang, S. & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement, 25(4)*, 317–331

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54(3)*, 427-450

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6(4)*, 473-492

Weiss, D. J. & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21(4)*, 361-375

Wouda, J. T. & Eggen, T. J. H. M. (2009). Computerized classification testing in more than two categories by using stochastic curtailment. In D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. Accessed date: [17.5.2018] from http://iacat.org/sites/default/files/biblio/cat09wouda.pdf

Yang, X, Poggio, J. C. & Glasnapp, D. R. (2006). Effects of estimation bias on multiple category classification with an IRT-based adaptive classification procedure. *Educational and Psychological Measurement, 66(4)*, 545-564

Yi, Q., Wang, T. & Ban, J. (2000). Effects of scale transformation and test termination rule on the precision of ability estimates in CAT. ACT Research Report Series, 2000-2. [Online: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1745-3984.2001.tb01127.x, Accessed date: 17.5.2018.]

EXTENDED ABSTRACT

Introduction

Because of the advantages of Item Response Theory (IRT) such as invariance of item parameters and person parameters Computerized Adaptive Testing (CAT) is getting more attention in last years. When the CAT applications are used to classify individuals into two or several groups according to one or more cut-point, Computerized Adaptive Classification Testing (CACT), which is a sub-field of CAT becomes a current issue. CACT aims to classify the persons with the highest classification accuracy using the least number of items according to one or more predefined cut-points and has six components: *(i)* Response model; *(ii)* Item pool; *(iii)* Starting rule; *(iv)* Item selection method; *(v)* Ability estimation method and *(vi)* Classification criteria. The efficiency of the classifications varies by item pools, classification criteria, item selection methods and ability estimation methods. According to this, in the CACT, forming different patterns and identification of these patterns under Monte Carlo (MC) simulations are important for real applications.

In this study, different classification criteria, various methods for item selection and ability estimation in the CACT, are compared using classification accuracy, test length and precision of measurement under the MC simulations. In our research, as classification criteria, Sequential Probability Ratio Test (SPRT), Generalized Likelihood Ratio (GLR) and Confidence Interval (CI) methods; as ability estimation methods, Expected a Posteriori (EAP) and Weighted Likelihood Estimation (WLE) methods; and as item selection methods, Maximum Fisher Information (MFI) and

Kullback-Leibler Information (KLI) methods on the basis of cut-point (CP) and estimated ability (EA) have been examined. The importance of this study comes from the 48 conditions that have not been investigated before Therefore, it is expected to provide some information about classification criteria, item selection methods and ability estimation methods to the researchers or practitioners.

Method

The purpose of this study was to identify the effects of the classification criteria, item selection methods and ability estimation methods in CACT on the classification accuracy, test length and precision of measurement. To achieve this aim, a pool of 500 items, which is based on 3 PLM and informs at the arbitrary cut-point and around, was generated. In this pool, items were simulated from uniform distribution as U[0,5; 2,0] for a parameters; normal distribution as N(1, 1,5) for b parameters and as N(0,15, 0,05) for c parameters. Individual abilities were derived from normal distribution as N(0,1) between (-3,+3) ability levels for 3000 individuals. The item response patterns were generated randomly in R software.

After the data production process, the CACT simulation study was performed for the 48 conditions (*6 classification criteria* ***x*** *4 item selection methods* ***x*** *2 ability estimation methods*) in R with the codes (with 25 *for cycles* for each condition) written by the researcher. At the end of the CACT simulations, the mean values of Average Test Length (ATL), Average Classification Accuracy (ACA), correlation between the true thetas and estimated thetas (r), bias, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for 25 replications have been calculated.

Results and Discussion

According to results of the study, it has been observed that the GLR and the CI classification criteria perform better compared to the SPRT in terms of test efficiency; however the SPRT works better compared to the other two methods in terms of bias, RMSE and MAE. It has also been deduced that the ATL decreases and test efficiency increases as the indifference region of classification criteria expands or the error value decreases.

In addition, it has been concluded that all classification criteria have considerably high level of the classification accuracy in all conditions; and both ability estimation methods, the EAP and the WLE, have successful estimation results in terms of the correlation between true and estimated thetas (r); whereas the EAP relatively performs better than the WLE in terms of the bias, RMSE and MAE. It has also been observed that, all of the item selection methods work similarly to each other however, the MFI-EA performs better for all conditions in terms of all dependent variables.

In conclusion, it can be said that the GLR method is the most preferable classification criteria in terms of test efficiency and precision of measurement and it is necessary to expand the indifference region of the SPRT or the GLR; or to decrease the error value of the CI in order to increase the test efficiency of CACT. In addition, because the EAP performs better than the WLE in terms of the precision of measurement, EAP can be used in real CACT applications. Lastly, the MFI item selection method on the basis of estimated ability (MFI-EA) can be the most appropriate item selection method for the real CACT.