Ahi Evran Üniversitesi
Kırşehir Eğitim Fakültesi Dergisi

ISSN: 2147 - 1037

http://kefad.ahievran.edu.tr

Investigation of Measurement Accuracy and Test Effectiveness for a Two, Three, Four Category Computerized Adaptive Classification Test

Demet Alkan
Nuri Doğan

Article Information

DOI: 10.29299/kefad.1279034

Received: 7.05.2023
Revised: 14.07.2023
Accepted: 30.07.2023

Keywords:

Computerised Classification Test,
Classification Criterion,
Test Efficiency,
Measurement Accuracy

Abstract

In this study, multi category classification was performed by using R (R Core Team, 2013) software language and simulating Computerized Adaptive Classification test. Two category, unidimensional 500 items and 1000 person data were created in R software. According to the number of classification categories, the Average Classification Accuracy (ACA) and Average Test Length (ATL) values for test effectiveness, Correlation (r) between actual and predicted abilities for measurement accuracy, Bias, Root Mean Square Error (RMSE), Mean Absolute Error (MAE) values were determined with the average of 25 repetitions. In the study, 16 conditions were created with 2 ability estimation methods as Weighted Likelihood Estimation (WLE), Expected Posterior Distribution (EPD); 4 item selection methods based on Maximum Fisher Information (MFI), Kullback Leibler Information (KLI) Ability Based (EB) and Cut Score Based (CB), 2 classification criteria as Confidence Interval (CI) and Sequential Probability Ratio Test (SPRT). In three and four category classification, MFI-EB item selection method gave better results in terms of test length, classification accuracy with WLE ability estimation. EPD ability estimation with KLI-CB item selection method made the most accurate classification with the least number of items in two category classification. EPD ability estimation showed the best performance for two category classification when crossed with the CI classification criterion.

İki, Üç, Dört Kategorili Bilgisayarda Bireyselleştirilmiş Sınıflama Testi İçin Ölçme Kesinliği ve Test Etkililiğinin İncelenmesi

Makale Bilgileri

DOI: 10.29299/kefad.1279034

Yükleme: 7.05.2023
Düzeltme: 14.07.2023
Kabul: 30.07.2023

Anahtar Kelimeler:

Bilgisayarda Bireyselleştirilmiş Sınıflama Testi,
Sınıflama Kriteri,
Test Etkililiği,
Ölçme Kesinliği

Öz

Bu çalışmada R (R Core Team, 2013) programlama dili kullanılarak bilgisayarda bireyselleştirilmiş sınıflama testi simülasyonu ile çok kategorili sınıflama yapılmıştır. R ortamında iki kategorili, tek boyutlu 500 madde ve 1000 kişilik veri oluşturulmuştur. Sınıflama kategori sayısına göre test etkililiği için ortalama sınıflamanın doğruluğu (OSD) ve ortalama test uzunluğu (OTU), ölçmenin kesinliği için gerçek yetenekler ile kestirilen yetenekler arasındaki korelasyon (r), yanlılık, ortalama hatanın karekökü (RMSE), ortalama mutlak hata (OMH) değerleri 25 tekrarın ortalaması ile belirlenmiştir. Araştırmada Ağırlıklı olabilirlik kestirimi (AOK), Beklenen sonsal dağılım (BSD) olarak 2 yetenek kestirim yöntemi; Maksimum Fisher Bilgi (MFB), Kullback Leibler bilgisi (KLB) Kestirilen Yetenek (KY) ve Kesme Noktası (KN) temelli 4 madde seçme yöntemi, Güven aralığı (GA) ve Ağırlıklı olabilirlik oran testi (AOOT), olarak 2 sınıflama kriteri ile 16 koşul oluşturulmuştur. Üç ve dört kategorili sınıflamada MFB-KY temelli madde seçme yöntemi AOK yetenek kestirimi ile test uzunluğu, sınıflamanın doğruluğu açısından daha iyi sonuçlar vermiştir. BSD yetenek kestirimi KLB-KN madde seçme yöntemi ile iki kategorili sınıflamada en az madde ile en doğru sınıflama yapmıştır. BSD yetenek kestirimi GA sınıflama kriteriyle çaprazlandığı koşullarda iki kategorili sınıflama için en iyi performansı göstermiştir.

Sorumlu Yazar : Demet Alkan, Phd.Doc., Hacettepe Üniversitesi, Türkiye, alkandemet@hotmail.com, ORCID ID: 0000-0002-1478-9183

Yazar2: Nuri Doğan Prof. Dr. Hacettepe Üniversitesi, Türkiye, nuridogan2004@gmail.com, ORCID ID: 0000-0001-6274-2016

Atıf için: Alkan, D. & Doğan, N. (2024). Iki, üç, dört kategorili bilgisayarda bireyselleştirilmiş sınıflama testi için ölçme kesinliği ve test etkililiğinin incelenmesi. *Kırşehir Eğitim Fakültesi Dergisi*, *25*(1), 29-63.

Giriş

Başarı ve yetenek testleri bir noktada yeteneği belirlemek amacıyla geliştirilirken amaç sınıflama kararı olduğunda sınıflandırma testleri kullanılır. Bilgisayar tabanlı testler özellikle değerlendirme sonuçları yüksek riskli olduğunda sınıflama kararı vermek için kağıt kalem yöntemlerine tercih edilebilir. Sınıflandırma testi prosedürlerinin amacı, bir sınav katılımcısını önceden belirlenmiş bir kesme puanına göre değerlendirmek ve kategorik bir sonuç sağlamaktır (Wainer, 1990; Weiss, 1983). Özellikle çok kategorili sınıflama yapmak için sabit formlu test yerine Bilgisayarda Bireyselleştirilmiş Sınıflama testi (BBST) kullanmak sınıflama yapabilmek için en uygun özelliklere sahip maddeleri seçmek açısından çok uygundur (Thompson, 2007). Bilgisayarda bireyselleştirilmiş sınıflama testinde kullanılan yetenek belirleme yöntemleriyle daha az sayıda madde ile daha doğru sınıflamalar elde edilebileceği yapılan çalışmalarda görülmüştür (Lewis ve Sheehan, 1990). Eğer bir testte üç ya da dört sınıflama varsa (iki veya üç kesme puanı), yüksek sayıda maddeye ihtiyaç duyan sınav katılımcılarının sayısı artar (Spray, 1993). Dolayısıyla, bu durum, tüm ölçeklerde ihtiyaç duyulan ortalama madde sayısını artırır. Bu da madde kullanım sıklığı kontrolü olmadan etkili bir test için madde havuzunda ihtiyaç duyulan madde sayısını artırır (Thompson, 2007). BBST uygulamalarında pratik kısıtlamalar sorunu, madde kullanım sıklığı kontrol yöntemleri, içerik dengeleme gibi, geçmişteki araştırmalardan kısıtlamaların getirilmesinin gerekli olmadığı özellikle simülasyon çalışmalarında pratik kısıtlamaların getirilmesinin araştırma için zararlı olduğu belirtilmektedir (Thompson,2007).

Geleneksel testlere kıyasla daha az madde kullanarak daha güvenilir sınıflama yapılmasını BBST sağladığı düşünülmektedir. (Fan, Wang, Chang, ve Douglas, 2012; Thompson, 2009). Bilgisayarda Bireyselleştirilmiş Sınıflama testlerinde madde sayısının az olması ve ortalama sınıflama doğruluğunun yüksek olmasıyla testin etkililiği artar. Hataların düşük olması, gerçek ve kestirilen yetenek düzeyleri arasındaki korelasyonun yüksek olması ölçme kesinliğini yükseltir (Thompson, 2009). BBST uygulamasında amaç bireyleri kesme noktasına göre daha az sayıda madde ile yüksek sınıflama doğrulukları ile sınıflara ayırmaktır. Özellikle sonuçları yüksek önem gösteren testlerde verilen kararlarda örneğin eğitim ve tıp gibi alanlarda mezun olma, meslek seçimi gibi önemli kararlar verildiğinden doğru sınıflama yapılması çok büyük önem göstermektedir (Thompson, 2007). BBST uygulamalarında farklı koşulların oluşturulması ve koşullara uygun desenlerin belirlenmesi önemlidir (Gündeğer, 2017). BBST araştırmasının genel amacı BBST'nin verimliliğini en üst düzeye çıkaran madde seçme ile yetenek kestirim yöntemleri, sınıflama kriterleri ile yetenek kestirim yöntemleri ve madde seçme yöntemleri ile sınıflama kriterlerinin çaprazlanmasındaki uygun koşulları belirlemektir. Yüksek sınıflama doğruluğu için az madde kullanarak test etkililiğini oluşturmaktır. Düşük standart hatalarla ölçme kesinliğini arttırmak ve en uygun sınıflama koşullarını belirlemektir (Thompson,2009).

Araştırmanın Amacı ve Önemi

Araştırmanın amacı, BBST simülasyonu ile yapılan çok kategorili sınıflamada yetenek kestirim yöntemleri ile sınıflama kriterleri, yetenek kestirim yöntemleri ile madde seçme yöntemlerinin, madde seçme yöntemleri ile sınıflama kriterlerinin çaprazlanması ile oluşturulan farklı koşullarda sınıflama kategori sayısına göre sınıflama doğruluğu ve ölçme kesinliğinin nasıl değiştiğini ve oluşturulan koşullara en uygun deseni belirlemektir. Yurt dışında alan yazında BBST koşullarının çaprazlandığı iki kategorili sınıflama örneklerine daha çok rastlanmaktadır (Lau, 1996; Reckase, 1983; Spray ve Reckase,1996). Yurt içinde daha az sayıda çalışmaya (Gündeğer, 2017; Demir, 2019; Gür ve Gülleroğlu, 2020) rastlanmaktadır. Gündeğer (2017) iki kategorili sınıflama için belirlediği yetenek kestirim yöntemlerinin, madde seçme yöntemlerinin ve sınıflama kriterlerinin oluşturduğu koşullarla performanslarını incelemiştir. Demir (2019) çok kategorili sınıflama için belirlediği koşullara ait madde kullanım sıklığı kontrol yöntemleri ve farklı içerik dengeleme yöntemlerinin performanslarını incelemiştir.

Alan yazında çok kategorili sınıflamada yetenek kestirim yöntemleri, sınıflama kriterleri ve madde seçme yöntemlerinin özellikle KLB (KN-KY) temelli madde seçme yöntemleri ile yetenek kestirim yöntemlerinin, madde seçme yöntemleri ile sınıflama kriterlerinin çaprazlanmasından oluşturulan koşulların performansının araştırıldığı araştırma örneklerine az rastlanmaktadır. Araştırmanın uygulayıcılara, iki, üç ve dört kategorili sınıflama için yetenek kestirim yöntemleri ile sınıflama kriterlerinin çaprazlanması ve yetenek kestirim yöntemleri ile madde seçme yöntemlerinin, madde seçme yöntemleri ile sınıflama kriterlerinin çaprazlanmasından oluşan koşulların ölçme kesinliği ve test etkililiği açısından performansları hakkında bilgi vermesi ve belirlenen koşullara en uygun desenin belirlenmesi beklenmektedir. Bu nedenle alan yazına katkı sağlayacağı düşünülmektedir. Bu araştırmada tek boyutlu 500 maddeden oluşan madde havuzu ile 1000 kişi üzerinde yapılan iki, üç ve dört kategorili sınıflama yapılmıştır. Sınıflamada test etkililiği açısından ortalama sınıflama doğruluğu, ortalama test uzunluğu, ölçme kesinliği açısından gerçek yetenekler ile kestirilen yetenekler arasındaki korelasyon, yanlılık, RMSE, OMH değerlerinin, yetenek kestirim yöntemlerinin sınıflama kriterleriyle, yetenek kestirim yöntemlerinin madde seçme yöntemleriyle ve madde seçme yöntemlerinin sınıflama kriterleri ile çaprazlandığı koşullardaki performanslarının sınıflama kategori sayısına göre nasıl değiştiği araştırılmıştır. Thompson (2007) pratik kısıtlamaların simülasyon çalışması için zararlı olduğunu düşünmektedir bu nedenle araştırmada pratik kısıtlamalar kullanılmamıştır.

Araştırma Problemleri

Araştırma problemleri aşağıdaki gibi belirlenmiştir.

1. BBST simülasyonu ile yapılan iki, üç, dört kategorili sınıflamada AOK ve BSD yetenek kestirim yöntemlerinin Kestirilen yetenek ve Kesme noktası temelli MFB (KN-KY), Kesme noktası ve

Kestirilen yetenek temelli KLB (KN-KY) madde seçme yöntemleriyle değerlendirildiği koşullara ait ortalama test uzunluğu, ortalama sınıflama doğruluğu ve test etkililiği nasıl değişmektedir?

2. BBST simülasyonu ile yapılan iki, üç, dört kategorili sınıflamada AOK ve BSD yetenek kesirim yöntemlerinin AOOT FB:0.1 ile ve %90 güven düzeyi ile GA sınıflama kriterleriyle değerlendirildiği sınıflama koşullarına ait OSD, OTU, yanlılık, RMSE, OMH değerleri nasıl değişmektedir?

3. BBST simülasyonu ile yapılan iki, üç ve dört kategorili sınıflamada sınıflama kriterlerinin AOOT FB:0,1, GA %90 güven düzeyi, kestirilen yetenek ve kesme noktası temelli madde seçme yöntemlerinin sınıflama kriterleriyle değerlendirildiği koşullara ait OSD, OTU, yanlılık, RMSE, OMH değerleri nasıl değişmektedir?

Yöntem

Araştırma simülasyon çalışmasıdır. Araştırmada yetenek parametreleri ve madde havuzu parametreleri R ortamında oluşturulmuştur (R Core Team, 2013). Araştırmada 2 yetenek kestirim yöntemi, 4 madde seçme yöntemi, 2 sınıflama kriteri ile iki, üç, dört kategorili sınıflama yapılmıştır. Farklı bir ifade ile 2 yetenek kestirim yöntemi x 4 madde seçme yöntemi x 2 sınıflama kriteri x 3 sınıflama kategori sayısı = 48 koşul oluşturulmuştur.

Veri üretimi

Araştırmada 500 maddelik madde havuzu oluşturulmuştur. Madde havuzu Thompson (2009, 2011) ve Kingsbury ve Weiss' in (1980) çalışmaları dikkate alınarak oluşturulmuştur. Maddelerin a parametresi U (0.5,1.5) dağılımından, b parametresi N (0,1) dağılımdan Kingsbury ve Weiss' in (1980) çalışması dikkate alınarak üretilmiştir. Thompson (2009) çalışması dikkate alınarak c parametresi N (0,0.3) olarak üretilmiştir. Yetenek parametresi ise 1000 birey için ortalaması 0, standart sapması 1 olacak şekilde R ortamında üretilmiştir. Yetenek kestirimi için AOK ve Bayes yetenek kestirim yöntemlerinden BSD yetenek kestirim yöntemleri kullanılmıştır. Madde seçme yöntemlerinden MFB kesme noktası ve kestirilen yetenek temelli ile KLB kesme noktası ve kestirilen yetenek temelli madde seçme yöntemleri kullanılmıştır. Sınıflama kriteri için FB: kesme puanına yakın sınıflama kararları için tolere edilebilen belirsizlik düzeyidir. Thompson (2011)'e göre farksızlık bölgesi küçük olursa sınıflama doğruluğunun yüksek olduğu düşünülmektedir. Eggen ve Straetmans (2000)'e göre güven aralığı değeri arttıkça sınıflama yapılması için gereken madde sayısı ve sınıflamanın doğruluğu artar. Bu çalışmada Nydick (2013) ile Eggen ve Straetmans' in (2000) çalışması dikkate alınarak AOOT 0.1 farksızlık bölgesi ile, %90 güven düzeyi ile GA sınıflama kriterleri kullanılmıştır.

Simülasyon Koşulları

Bilgisayarda bireyselleştirilmiş sınıflama testi için iki, üç, dört kategorili sınıflamada 2 yetenek kestirim yönteminin 2 sınıflama kriteriyle, 2 yetenek kestirim yönteminin 4 madde seçme yöntemiyle

ve 4 madde seçme yönteminin 2 sınıflama kriterleri ile çaprazlandığı koşullara ait 25 tekrarın ortalaması ile R (Core Team, 2013) ortamında analizler yapılmıştır. Sınıflama kriterleri 0,1 farksızlık bölgesi ile AOOT, %90 güven düzeyi ile GA yöntemleri Nydick (2013) ile Eggen ve Straetmans' in (2000) araştırmalarına göre seçilmiştir. Literatüre göre, farksızlık bölgesi ve güven aralığı değeri tolere edilebilir hata seviyesini göstermektedir. Güven aralığı değeri arttıkça ve farksızlık bölgesi sabiti küçüldükçe, testin sınıflandırması için gereken madde sayısı ve sınıflandırmanın doğruluğu artar (Eggen ve Straetmans, 2000). Yetenek kestirim yöntemlerinden Bayesçi yetenek kestirim yöntemlerinin araştırmalarda fazla çalışılmadığı görülmektedir. Warm'a (1989) göre AOK yanlılığı azaltan ağırlıklandırma olabilirliği üzerine çalışan, tüm maddeler doğru ya da yanlış cevaplandığında da yeteneği kestiren bir yöntemdir. Araştırmada AOK ve Bayesci yetenek kestirimlerinden BSD yetenek kestirim yöntemleri kullanılmıştır. BBST için kesme noktası temelli ve kestirilen yetenek temelli madde seçme yöntemleri bulunmaktadır. Araştırmada MFB (KN-KY), KLB (KN-KY) madde seçme yöntemlerinin yetenek kestirim yöntemleri ve sınıflama kriterleri ile çaprazlanması durumundaki performansları araştırılmıştır. Thompson'a (2007) göre, başlama noktası için yetenek düzeyi 0 alınabilir ya da önceden belirlenen yetenek düzeyleri kullanılabilir. Simülasyon koşulları araştırma problemlerine göre oluşturulmuştur. İki, üç, dört kategorili sınıflama için kesme noktaları Eggen ve Straetmans' ın (2000) çalışmalarına göre yetenek düzeyleri ikiye ayrılıp ilk bölüm 1. düzey diğer bölüm 2. düzey olarak her düzeyin %70 i alınarak belirlenebilir. Kesme noktaları rastgele de belirlenebilir. Bu çalışmada Eggen ve Straetmans' ın (2000) çalışmaları dikkate alınarak kesme noktaları belirlenmiştir. Araştırmada CatIRT (Nydick, 2014) paketi kullanılmıştır.

Verilerin analizi

Araştırmada iki, üç, dört kategorili sınıflama için oluşturulan 16 simülasyon koşulu için yapılan analizler gerçek uygulamaya en yakın sonuç elde etmek için 25 tekrarın ortalaması alınarak R da yazılan fonksiyonlarla oluşturulmuştur. Sınıflama doğruluğu için ortalama sınıflama doğruluğu, ortalama test uzunluğu değerleri hesaplanmıştır. Ölçme kesinliği için. RMSE, OMH, yanlılık, gerçek yetenekler ile kestirilen yetenek düzeyleri arasındaki korelasyon (r) hesaplanmıştır. Gerçek yetenek düzeyi ile kestirilen yetenek düzeyleri arasındaki korelasyon (r) için Pearson korelasyon katsayısı değeri hesaplanmıştır. OSD için gerçek sınıflar ile simülasyon sonucu hesaplanan sınıflar arasındaki uyum Cohen Kappa istatistiği ile hesaplanmıştır.

Yanlılık, son yetenek düzeylerinin (θi) gerçek yetenek düzeylerinden (θ_1) farkları toplamının birey sayısına (n) oranına eşittir (Miller ve Miller, 2004).

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

RMSE, kestirilen son yetenek düzeylerinin (θi) gerçek yetenek düzeylerinden (θ_1) farklarının kareleri toplamının birey sayısına (n) oranının karekökünе eşittir.

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

OMH, kestirilen son yetenek düzeylerinin (θi) gerçek yetenek düzeylerinden (θ_1) farklarının mutlak değerleri toplamının birey sayısına (n) oranına eşittir.

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

Araştırmanın Etik İzinleri

Yapılan bu çalışmada "Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesi" kapsamında uyulması belirtilen tüm kurallara uyulmuştur. Yönergenin ikinci bölümü olan "Bilimsel Araştırma ve Yayın Etiğine Aykırı Eylemler" başlığı altında belirtilen eylemlerden hiçbiri gerçekleştirilmemiştir.

Bulgular

Çalışmanın birinci araştırma probleminde BBST simülasyonu ile yapılan iki, üç, dört kategorili sınıflamada AOK ve BSD yetenek kestirim yöntemlerinin KN ve KY temelli MFB ve KLB madde seçme yöntemleriyle değerlendirildiği sınıflama koşullarına ait ortalama sınıflama doğruluğu, ortalama test uzunluğu, yanlılık, RMSE ve OHM değerlerinin sınıflama kategori sayısına göre nasıl değiştiği incelenmiştir.

Tablo 1'e göre İki kategorili sınıflamada AOK yetenek kestirim yönteminin MFB (KY-KN) madde seçme yöntemleri ile çaprazlandığı koşullara ait OTU için 19.76-17.91, OSD için 0.89-0.891 değerleri hesaplanmıştır. AOK yetenek kestiriminin MFB-KN temelli madde seçme yöntemi ile iki kategorili sınıflamada daha az madde ile sınıflama yaptığı görülmektedir. OSD için her iki madde seçme yöntemi için benzer sonuçların hesaplandığı söylenebilir. Yanlılık, RMSE, OMH ve gerçek yetenekler ile kestirilen yetenekler arasındaki korelasyon yani ölçme kesinliği açısından AOK yetenek kestirim yönteminin MFB (KY-KN) temelli madde seçme yöntemleriyle oluşturulan koşullarda 0.945-0.844 korelasyon değeri, 0.002-0.35 yanlılık, 0.337-0.706 RMSE, 0.241-0.484 OMH değerleri hesaplanmıştır. MFB-KY temelli madde seçme yöntemi ile AOK yetenek kestiriminin KN temelli madde seçme yöntemine göre ölçme kesinliği için daha düşük hata değerleri ile daha iyi performans gösterdiği görülmektedir.

AOK yetenek kestirim yöntemi ile iki kategorili sınıflamada KLB madde seçme yöntemi birlikte kullanıldığında KY-KN temelli madde seçme yöntemi ile 19.87-17.69 OTU, 0.887- 089 OSD değerlerinin oluştuğu tablodan anlaşılmaktadır. Buna göre KN temelli madde seçme yöntemi ile daha az madde ve daha yüksek sınıflama doğruluğu ile sınıflama yapıldığı görülmektedir. KLB-KN temelli madde seçme

yöntemi ile AOK yetenek kestiriminin çaprazlandığı koşulun test etkililiği için performansının daha yüksek olduğu araştırmanın bulguları arasındadır. Ölçme kesinliği için KLB (KY-KN) madde seçme yöntemi ile 0.947-0.845 korelasyon, -0.003-0.346 yanlılık, 0.332-0.699 RMSE ve 0.238-0.478 OMH değerleri hesaplanmıştır. Buna göre ölçme kesinliği açısından iki kategorili sınıflamada KY temelli KLB madde seçme yönteminin AOK yetenek kestirimi ile birlikte kullanıldığında hata değerlerinin ve yanlılık değerinin daha düşük olduğu KY temelli madde seçme yönteminin performansının daha yüksek olduğu Tablo 1'de hesaplanan değerlerden anlaşılmaktadır. Tablo 1 de Yetenek kestirim yöntemleri ile madde seçme yöntemlerinin çaprazlandığı koşullara ait değerler gösterilmektedir.

Tablo 1. *Yetenek kestirim yöntemleri ve madde seçme yöntemlerinin çaprazlandığı koşullara ait sınıflama kategori sayısına göre OTU, OSD, r, yanlılık, RMSE ve OMH değerleri*

Koşullar			Bağımlı Değişkenler					
Yetenek Kestirim Yöntemleri	Madde Seçme Yöntemleri	SKS	OTU	OSD	r	Yanlılık	RMSE	OMH
AOK	MFB-KY	İki	19.76	0.891	0.945	0.002	0.337	0.241
		Üç	26.29	0.877	0.963	-0.001	0.28	0.201
		Dört	36.07	0.868	0.976	0	0.225	0.16
	MFB-KN	İki	17.91	0.89	0.844	0.35	0.706	0.484
		Üç	28.50	0.89	0.929	0.117	0.427	0.269
		Dört	37.82	0.869	0.972	-0.004	0.245	0.17
	KLB-KY	İki	19.87	0.887	0.947	-0.003	0.332	0.238
		Üç	26.28	0.877	0.963	-0.008	0.28	0.201
		Dört	35.98	0.869	0.976	-0.002	0.226	0.16
	KLB-KN	İki	17.69	0.89	0.845	0.346	0.699	0.478
		Üç	28.45	0.891	0.93	0.116	0.425	0.267
		Dört	37.78	0.869	0.972	-0.005	0.245	0.17
BSD	MFB-KY	İki	19.83	0.89	0.95	-0.001	0.324	0.234
		Üç	26.12	0.878	0.963	-0.003	0.281	0.202
		Dört	36.32	0.871	0.976	0	0.225	0.158
	MFB-KN	İki	17.18	0.889	0.851	0.009	0.543	0.393
		Üç	27.70	0.892	0.932	0.002	0.375	0.256
		Dört	36.75	0.871	0.971	0	0.25	0.173
	KLB-KY	İki	19.94	0.888	0.951	-0.001	0.319	0.231
		Üç	26.20	0.878	0.963	-0.002	0.278	0.2
		Dört	36.25	0.871	0.976	0.001	0.225	0.159
	KLB-KN	İki	17.05	0.888	0.854	0.009	0.539	0.39
		Üç	27.69	0.892	0.933	0.003	0.372	0.253
		Dört	36.68	0.87	0.971	0.01	0.25	0.174

Ortalama test uzunluğu: OTU, Ortalama sınıflama doğruluğu: OSD, Ortalama mutlak hata: OMH, Ortalama hatanın karekökü (RMSE)

İki kategorili sınıflamada BSD yetenek kestirim yöntemi KY- KN temelli MFB madde seçme yöntemleriyle çaprazlandığı simülasyonda 19.83-17.18 OTU, 0.89-0.889 OSD değerleri hesaplanmıştır. Buna göre KN temelli madde seçme yönteminin daha az madde ile sınıflama yaptığı araştırmanın bulgularındandır. OSD için her iki madde seçme yöntemi de BSD yetenek kestirimi ile birlikte kullanıldığında benzer performans göstermişlerdir. Test etkililiği için daha az madde ile sınıflama yapmanın amaç olduğu BBST için KN temelli MFB yönteminin BSD yetenek kestirimi ile iki kategorili sınıflamada etkili olduğu bulgulanmıştır. Ölçme kesinliği için KY-KN temelli madde seçme yöntemleri ile BSD yetenek kestirimi birlikte kullanıldığında 0.95-0851 korelasyon, -0.001-0.009 yanlılık, 0.324-0.543 RMSE, 0.234-0.393 OMH değerleri hesaplanmıştır. Buna göre MFB-KY temelli madde seçme yönteminin BSD yetenek kestirimi ile ölçme kesinliği olarak iki kategorili sınıflamada performansının daha yüksek olduğu görülmektedir.

BSD yetenek kestirimi ile KLB (KY-KN) madde seçme yöntemlerinin çaprazlandığı iki kategorili simülasyonda 19.94-17.05 OTU, 0.888-0.888 OSD değerleri hesaplanmıştır. KLB-KN madde seçme yönteminin daha az madde ile ve benzer sınıflama doğruluğu ile KY temelli madde seçme yöntemine göre daha iyi performans gösterdiği görülmektedir. Korelasyon, yanlılık ve ölçmenin standart hata değerlerine bakıldığında 0.951-0.854 korelasyon, -0.001-0.009 yanlılık, 0.319-0.539 RMSE,0.231-0.39 OMH değerleri hesaplanmıştır. Test etkililiği olarak KLB-KN temelli, ölçme kesinliği olarak KLB-KY temelli madde seçme yöntemin daha iyi performans gösterdiği tablodan anlaşılmaktadır.

Üç kategorili sınıflama için AOK yetenek kestirimi MFB (KY-KN) temelli madde seçme yöntemleriyle çaprazlandığında OTU değerleri 26.29-28.50, OSD 0.877-0.89 olarak hesaplanmıştır. Buna göre test uzunluğu olarak KY temelli madde seçme yöntemi sınıflama doğruluğu olarak KN temelli madde seçme yöntemi göreceli olarak daha iyi performans gösterse de yöntemler için benzer bulgular görülmektedir. Ölçme kesinliği için AOK yetenek kestirimi ile MFB (KY-KN) temelli madde seçme yöntemleri birlikte kullanıldığında 0.963-0.929 korelasyon, -0.01-0.117 yanlılık, 0.28-0.427 RMSE ve 0.201-0.269 OMH değerleri hesaplanmıştır. Buna göre MFB-KY temelli madde seçme yönteminin daha düşük hata ve daha yüksek korelasyonla sınıflama yaptığı görülmektedir. KY temelli madde seçme yöntemi ölçme kesinliği olarak AOK yetenek kestirimi ile KN temelli madde seçme yöntemine göre daha iyi performans göstermiştir.

AOK yetenek kestiriminin KLB (KY-KN) temelli madde seçme yöntemleri ile çaprazlandığı koşullarda 26.28-28.45 OTU, 0.877-0.891 OSD değerleri hesaplanmıştır. KY temelli madde seçme yöntemi daha az madde ile KN temelli madde seçme yöntemi kısmen daha yüksek sınıflama doğruluğu ile sınıflama yapmıştır. Ölçme kesinliği için AOK yetenek kestirimi KLB (KY-KN) madde seçme yöntemleri ile birlikte kullanıldığı koşullarda 0.963-0.93 korelasyon, -0.008-0.116 yanlılık,0.28-0.425 RMSE ve 0.201-0.267 OMH değerleri hesaplanmıştır. KY temelli madde seçme yöntemi AOK yetenek

kestirimi ile daha düşük hata, yanlılık ve daha yüksek korelasyonla sınıflama yapmıştır. Test etkililiği ve ölçme kesinliği için KLB-KY temelli madde seçme yönteminin performansı daha yüksektir.

Üç kategorili sınıflamada BSD yetenek kestirimi kullanıldığında MFB (KY-KN) madde seçme yöntemleri ile oluşturulan simülasyonda 26.12-27.70 OTU, 0.878-0.892 OSD ile sınıflama yapılmıştır. BSD yetenek kestirim yöntemi KY temelli madde seçme yöntemi ile birlikte kullanıldığında daha az madde ile KN temelli madde seçme yöntemi ile birlikte kullanıldığında daha yüksek OSD ile sınıflama yapılmıştır. Ölçme kesinliği için BSD yetenek kestirimi ile MFB (KY-KN) madde seçme yöntemleri birlikte oluşturduğu koşullara ait sınıflamada 0.963-0.932 korelasyon,-0.003-0.002 yanlılık, 0.281-0.375 RMSE, 0.202-0.256 OMH değerleri hesaplanmıştır. BSD yetenek kestirimi ile KY temelli madde seçme yöntemi ölçme kesinliği olarak KN temelli madde seçme yöntemine göre daha iyi performans göstermiştir.

BSD yetenek kestiriminin KLB (KY-KN) madde seçme yöntemleri ile birlikte kullanıldığı üç kategorili sınıflamada 26.20-27.69 OTU, 0.878-0.892 OSD değerleri hesaplanmıştır. OTU için KY temelli madde seçme yöntemi, OSD için KN temelli madde seçme yönteminin daha iyi performans gösterdiği görülmektedir. Benzer şekilde ölçme kesinliği için 0.963-0.933 korelasyon, -0.002-0.003 yanlılık, 0.278-0.373 RMSE, 0.2-0.253 OMH değerleri hesaplanmıştır. KY temelli madde seçme yönteminin düşük hatalar ve yanlılık, yüksek korelasyonla KN temelli madde seçme yöntemine göre daha iyi performans gösterdiği ifade edilebilir.

Dört kategorili sınıflamada AOK yetenek kestiriminin ile MFB (KY-KN) temelli madde seçme yöntemleri ile çaprazlanmasından oluşan koşullar için 36.07-37.82 OTU, 0.868-0.869 OSD değerleri hesaplanmıştır. KY temelli madde seçme yöntemi ile daha az madde kullanarak sınıflama yapıldığı görülmektedir. OSD için her iki madde seçme yöntemiyle de yakın değerler hesaplan test etkililiği için MFB-KY temelli madde seçme yöntemi ile oluşturulan koşulun performansının daha iyi olduğu bulgular arasındadır. MFB (KY-KN) madde seçme yöntemi ile oluşturulan koşullarda 0.976-0.972 korelasyon, 0,-0.004 yanlılık,0.225-0.245 RMSE, 0.16-0.17 OMH değerleri hesaplanmıştır. Ölçme kesinliği olarak KY temelli madde seçme yönteminin daha düşük hata ve daha yüksek korelasyon ile sınıflama yaptığı tablodan anlaşılmaktadır.

KLB (KY-KN) temelli madde seçme yöntemleriyle yapılan çaprazlamada da 35.98-37.78 OTU, 0.869-0.869 OSD değerleri hesaplanmıştır. KLB-KY madde seçme yöntemi ile oluşturulan koşulda daha az madde ve yüksek sınıflama doğruluğu ile sınıflama yapıldığı görülmektedir. KLB (KY-KN) madde seçme yöntemleriyle oluşturulan koşullarda 0.976-0.972 korelasyon, -0.02,-0.005 yanlılık, 0.226-0.245 RMSE,0.16-0.17 OMH değerleri hesaplanmıştır. KLB-KY madde seçme yöntemiyle oluşturulan koşulda daha yüksek korelasyon ve düşük standart hatalarla sınıflama yapıldığı görülmektedir. KY temelli madde seçme yönteminin ölçme kesinliği olarak daha iyi performans gösterdiği tablodan anlaşılmaktadır.

Dört kategorili sınıflama için BSD yetenek kestiriminin MFB (KY-KN) temelli madde seçme yöntemleriyle çaprazlanmasından oluşturulan koşullarda 36.32-36.75 OTU, 0.871-0.871 OSD değerleri hesaplanmıştır. Her iki madde seçme yöntemi ile BSD yetenek kestiriminin benzer performans gösterdiği görülmektedir. Aynı koşula ait 0.976-0.971 korelasyon, 0-0 yanlılık,0.225- 0.25 RMSE, 0.158-0.173 OMH değerleri hesaplanmıştır. KY temelli madde seçme yönteminin KN temelli madde seçme yöntemine göre ölçme kesinliği olarak BSD yetenek kestirimi ile performansının daha yüksek olduğu görülmektedir.

KLB(KY-KN) temelli madde seçme yöntemlerinin BSD yetenek kestirimi ile çaprazlanmasından dört kategorili sınıflama için oluşturulan koşullarda 36.25-36.68 OTU, 0.871-0.87 OSD ile madde seçme yöntemleri benzer performans gösterse de ölçme kesinliği olarak 0.976-0.971 korelasyon, 0.001-0.1 yanlılık, 0.225-0.25 RMSE, 0.159-0.174 OMH değerleri hesaplanmıştır. KY temelli madde seçme yönteminin daha düşük hata ile daha kesin ölçme yaptığı söylenebilir. Sınıflama kategori sayısı arttıkça hata değerlerinin düştüğü başka bir ifade ile ölçme kesinliğinin yükseldiği araştırmanın bulgularındandır. Sınıflama kategori sayısı arttıkça OSD değeri azalmış, OTU değeri ise artmış başka bir ifade ile test etkililiği düşmüştür. MFB-KY madde seçme yönteminin AOK yetenek kestirim yöntemi ile ölçme kesinliği açısından daha iyi performans gösterdiği, KLB madde seçme yönteminin ise OSD ve OTU açısından daha iyi performans gösterdiği tablodan anlaşılmaktadır. Test etkililiği için KLB-KY madde seçme yöntemi ile BSD yetenek kestiriminin daha iyi performans gösterdiği araştırmanın bulgularındandır.

İkinci araştırma probleminde BBST simülasyonu ile yapılan iki, üç, dört kategorili sınıflamada AOK ve BSD yetenek kesirim yöntemlerinin AOOT (FB:0.1) ve GA (%90) sınıflama kriterleriyle değerlendirildiği sınıflama koşullarına ait ortalama sınıflama doğruluğu, ortalama test uzunluğu, yanlılık, RMSE ve OHM değerlerinin sınıflama kategori sayısına göre nasıl değiştiği incelenmiştir. Tablo 2 de Yetenek kestirim yöntemleri ile sınıflama kriterlerinin çaprazlandığı koşullara ait değerler gösterilmektedir.

AOK yetenek kestiriminin AOOT(FB:0.1) sınıflama kriteri ile çaprazlandığı koşullara ait iki kategorili sınıflamada OTU 33.502, OSD ise 0.897'dir. Üç kategorili sınıflamada OTU 36.346, OSD 0.895, dört kategorili sınıflama için OTU 48.304, OSD 0875 olarak hesaplanmıştır. Sınıflama kategori sayısı arttıkça OTU artmıştır. OSD değeri ise azalmıştır. Yanlılık, RMSE, OMH değerlerinin sınıflama kategori sayısı arttıkça arttığı, gerçek yetenekler ile kestirilen yetenekler arası korelasyonun azaldığı araştırmanın bulgularındandır.

AOK yetenek kestiriminin GA (%90) sınıflama kriteriyle çaprazlandığı koşullara ait iki kategorili sınıflamada OTU 11.324, üç kategorili sınıflama için OTU 24.28, dört kategorili sınıflama için 31.562 dir. İki kategorili sınıflamada en az madde sayısı ile sınıflama yapılmıştır.

Tablo 2. *Yetenek kestirim yöntemleri ile sınıflama kriterlerinin çaprazlandığı koşullara ait OTU, OSD, yanlılık, korelasyon ve RMSE, OMH değerleri*

Koşullar			Bağımlı değişkenler					
Yetenek kestirim yöntemleri	Sınıflama kriterleri	SKS	OTU	OSD	r	Yanlılık	RMSE	OMH
AOK	AOOT	İki	33.502	0.897	985	0.002	0.183	0.141
		Üç	36.346	0.895	0.941	0.114	0.392	0.239
		Dört	48.304	0.875	0.981	-0.004	0.198	0.141
	GA	İki	11.324	0.877	0.921	-0.004	0.404	0.303
		Üç	24.28	0.884	0.93	0.106	0.422	0.272
		Dört	31.562	0.862	0.969	-0.015	0.255	0.182
BSD	AOOT	İki	33.078	0.897	0.985	-0.001	0.181	0.139
		Üç	36.092	0.896	0.948	0.004	0.33	0.22
		Dört	47.906	0.875	0.981	-0.001	0.199	0.141
	GA	İki	12.009	0.879	0.928	-0.002	0.386	0.291
		Üç	22.357	0.888	0.924	0.001	0.396	0.275
		Dört	29.001	0.866	0.964	0	0.277	0.196

Ortalama test uzunluğu: OTU, Ortalama sınıflama doğruluğu: OSD, Ortalama mutlak hata: OMH, Ortalama hatanın karekökü (RMSE)

OSD için üç kategorili sınıflamada GA(%90) sınıflama kriterinin 0.884 ile en yüksek performansı gösterdiği görülmektedir. AOK yetenek kestiriminin GA(%90) sınıflama kriteri ile çaprazlanması ile oluşturulan koşullarda yanlılık, RMSE, OMH gibi hata değerlerinin dört kategorili sınıflamada en düşük değerlerin hesaplandığı görülmektedir. Gerçek yetenekler ile kestirilen yetenekler arasındaki korelasyonun en yüksek değeri dört kategorili sınıflamada hesaplanmıştır. Ölçmenin kesinliği açısından oluşturulan koşulların dört kategorili sınıflamada iyi performans gösterdiği araştırmanın bulgularındandır. Testin etkililiği açısından iki kategorili sınıflamada koşulların daha etkili performans gösterdiği görülmektedir.

BSD yetenek kestirimi AOOT (FB:0.1) sınıflama kriteri ile çaprazlandığı koşullarda iki kategorili sınıflamada 33.078 OTU, üç kategorili sınıflamada 36.092 OTU, dört kategorili sınıflamada 47.906 OTU değerleri hesaplanmıştır. Sınıflama kategori sayısı arttıkça OTU artmıştır. OSD olarak iki ve üç kategorili sınıflamalar için benzer değerler hesaplanmıştır. Ölçme kesinliği açısından hata değerlerinin iki kategorili sınıflamada daha düşük olduğu araştırmanın bulgularındandır.

BSD yetenek kestirimi GA (%90) sınıflama kriteri ile çaprazlandığı koşullarda 12.009 OTU ile en az madde ile sınıflama yapmıştır. Sınıflama kategori sayısı arttıkça OTU değeri artmış OSD değeri düşmüştür. Ölçmenin kesinliğinin yüksek olduğu hatanın düşük olduğu performansı BSD yetenek kestiriminin GA (%90) sınıflama kriteri ile dört kategorili sınıflamada gösterdiği araştırmanın bulguları arasındadır.

Üçüncü araştırma probleminde iki, üç ve dört kategorili sınıflamada sınıflama kriterlerinin, madde seçme yöntemleri ile değerlendirildiği koşulların ölçme kesinliği ve test etkililiği bakımından nasıl değiştikleri incelenmiştir. Tablo 3 de Madde seçme yöntemleri ile sınıflama kriterlerinin çaprazlandığı koşullara ait değerler gösterilmektedir.

Tablo 3. *Madde seçme yöntemleri ile sınıflama kriterlerinin çaprazlandığı koşullara ait değerler*.

Koşullar			Bağımlı Değişkenler					
Madde Seçme Yöntemi	Sınıflama Kriteri	SKS	OTU	OSD	r	Yanlılık	RMSE	OMH
MFB-KY	AOOT	İki	33.09	0.897	0.984	-0.001	0.182	0.14
		Üç	37.91	0.085	0.985	-0.001	0.18	0.137
		Dört	46.69	0.874	0.987	0	0.169	0.136
	GA	İki	11.61	0.876	0.923	-0.002	0.398	0.299
		Üç	17.93	0867	0.944	-0.003	0.342	0.253
		Dört	28.29	0.863	0.969	-0.001	0.256	0.182
MFB-KN	AOOT	İki	33.09	0.897	0.984	-0.001	0.182	0.14
		Üç	36.26	0.895	0.942	0.061	0.363	0.228
		Dört	48.11	0.874	0.981	-0.003	0.201	0.141
	GA	İki	11.61	0.876	0.923	-0.002	0.398	0.299
		Üç	22.95	0.885	0.922	0.057	0.417	0.278
		Dört	30.08	0.864	0.966	-0.008	0.271	0.192
KLB-KY	AOOT	İki	33.38	0.895	0.985	-0.001	0.181	0.139
		Üç	37.94	0.884	0.985	-0.001	0.179	0.137
		Dört	46.70	0.874	0.987	0	0.169	0.127
	GA	İki	11.57	0.876	0.925	-0.007	0.393	0.296
		Üç	18.03	0.868	0.944	-0.01	0.341	0.252
		Dört	28.12	0.864	0.969	-0.002	0.256	0.183
KLB-KN	AOOT	İki	33.38	0.895	0.985	-0.001	0.181	0.139
		Üç	36.22	0.895	0.922	0.057	0.417	0.278
		Dört	48.15	0.874	0.981	-0.004	0.202	0.141
	GA	İki	11.57	0.876	0.925	0.007	0.393	0.296
		Üç	22.89	0.885	0.923	0.058	0.414	0.275
		Dört	29.97	0.863	0.966	-0.008	0.271	0.191

Ortalama test uzunluğu: OTU, Ortalama sınıflama doğruluğu: OSD, Ortalama mutlak hata: OMH, Ortalama hatanın karekökü (RMSE)

Araştırmanın üçüncü problemine ait tüm koşullar için 25 tekrarın ortalaması alınarak elde edilen değerlere göre oluşturulan Tablo 3 de MFB (KY-KN) temelli madde seçme yöntemlerinin ikisi için de testi sonlandırmak bireyleri sınıflamak için AOOT (FB:01) sınıflama kriteri ile oluşturulan koşullara ait iki kategorili sınıflamada 33.09-33.09 OTU, 0.897-0.897 OSD, üç kategorili sınıflamada

37.91-36.26 OTU, 0.085-0.895 OSD, dört kategorili sınıflamada 46.69-48.11 OTU, 0.874-0.874 OSD değerleri hesaplanmıştır. GA (%90) sınıflama kriteri ile oluşturulan koşullarda iki kategorili sınıflama için 11.61-11.61 OTU, 0.876-0.876 OSD, üç kategorili sınıflamada 17.93-22.95 OTU,0.867-0.885 OSD, dört kategorili sınıflamada 28.29-30.08 OTU, 0.863-0.864 OSD değerleri hesaplanmıştır. GA sınıflama kriterinin AOOT sınıflama kriterine göre daha az madde ile sınıflama yaptığı görülmektedir. AOOT sınıflama kriterinin ise daha yüksek ortalama sınıflama doğruluğu değerleri ile sınıflama yaptığı araştırmanın bulguları arasındadır. Sınıflama kategori sayısı arttıkça OTU arttığı, OSD ise azaldığı araştırmanın bulgularındandır. KY ve KN temelli madde seçme yöntemleriyle sınıflama kriterleri benzer performans gösterse de MFB-KY temelli madde seçme yönteminin performansının AOOT sınıflama kriteri ile KN temelli madde seçme yöntemine göre daha yüksek olduğu ifade edilebilir.

MFB (KY-KN) madde seçme yöntemlerinin sınıflama kriterleri ile çaprazlandığı koşullara ait ölçmenin standart hatasını gösteren değerler ise AOOT (FB:01) sınıflama kriteri ile iki kategorili sınıflamada 0.984-0.984 korelasyon, -0.001,-0.001 yanlılık, 0.182-0.182 RMSE, 0.14-0.14 OMH değerleri, üç kategorili sınıflamada 0.985-0.944 korelasyon, -0.001-0.061 yanlılık, 0.18-0.363 RMSE ve 0.137-0.228 OMH, dört kategorili sınıflamada 0.987-0.981 korelasyon, 0,-0.003 yanlılık, 0.169-0.201 RMSE, 0.136-0.141 OMH değerleri hesaplanmıştır. AOOT sınıflama kriterinin ölçme kesinliği olarak KY temelli madde seçme yöntemi ile daha iyi performans gösterdiği bulgular arasındadır. Sınıflama kategori sayısı arttıkça hata değerleri azalmıştır. GA (%90) sınıflama kriteri ile iki kategorili sınıflamada 0.923-0.923 korelasyon, -0.002,-0.002 yanlılık, 0.398-0.398 RMSE ve 0.299-0.299 OMH değerleri, üç kategorili sınıflamada 0.944-0.922 korelasyon, -0.003,-0.57 yanlılık, 0.342-0.363 RMSE ve 0.253-0.278 OMH değerleri hesaplanmıştır. GA sınıflama kriteri KY ve KN temelli madde seçme yöntemleriyle benzer performans gösterse de KY temelli madde seçme yöntemi ile yapılan sınıflamanın performansının daha yüksek olduğu, sınıflama kategori sayısı arttıkça hata değerlerinin düştüğü görülmektedir. GA sınıflama kriterinin test etkililiği açısından performansı AOOT sınıflama kriterinden, AOOT sınıflama kriterinin ise ölçme kesinliği olarak performansı GA sınıflama kriterinden daha yüksek olduğu söylenebilir.

KLB (KY-KN) madde seçme yöntemleri ile AOOT (FB:01) ve GA(%90) sınıflama kriterlerinin çaprazlandığı koşullarda AOOT(FB:0.1) sınıflama kriteri ile iki kategorili sınıflamada 33.38-33.38 OTU, 0.895-0.895 OSD, üç kategorili sınıflamada 37.94-36.22 OTU, 0.884-0.895 OSD, dört kategorili sınıflamada 46.70-48.15 OTU ve 0.874-0.874 OSD değerleri hesaplanmıştır. GA (%90) sınıflama kriteri ile iki kategorili sınıflamada 11.57-11.57 OTU, 0.876-0.876 OSD, , üç kategorili sınıflamada 18.03-22.89 OTU, 0.868-0.885 OSD değerleri, dört kategorili sınıflamada 28.12-29.97 OTU ve 0.864-0.863 OSD değerleri hesaplanmıştır. GA sınıflama kriterinin iki kategorili sınıflamada en az madde ile sınıflama yaptığı görülmektedir. Test etkililiği olarak GA sınıflama kriterinin AOOT sınıflama kriterine göre daha yüksek performans gösterdiği görülmektedir. Sınıflama kategori sayısı arttıkça OTU artmış, OSD

azalmıştır. KY ve KN temelli madde seçme yöntemleri benzer performans gösterse de KY temelli madde seçme yönteminin test etkililiği açısından daha etkili olduğu araştırmanın bulgularındandır.

Ölçme kesinliğini gösteren kestirimin hata değerlerinde ise KLB (KY-KN) madde seçme yöntemleri ile AOOT(FB:01) sınıflama kriterinin çaprazlandığı koşullara göre iki kategorili sınıflamada 0.985-0.985 korelasyon, -0.001,-0.001 yanlılık, 0.181-0.181 RMSE ve 0.139-0.139 OMH değerleri, üç kategorili sınıflamada 0.985-0.922 korelasyon, -0.001-0.057 yanlılık, 0.179-0.414 RMSE ve 0.137-0.278 OMH değerleri, dört kategorili sınıflamada 0.987-0.981 korelasyon, 0,-0.004 yanlılık, 0.169-0.202 RMSE ve 0.202-0.127 OMH değerleri hesaplanmıştır. Sınıflama kategori sayısı arttıkça hata değerleri azalmıştır. Ölçme kesinliği artmıştır. KY ve KN temelli madde seçme yöntemleri ile AOOT sınıflama kriteri birlikte kullanıldığında madde seçme yöntemlerinin performansları benzer olsa da KY temelli madde seçme yöntemi ile oluşturulan koşulların daha az hata ile sınıflama yaptığı araştırmanın bulgularındandır. GA sınıflama kriteri ile iki kategorili sınıflamada 0.925-0.925 korelasyon, -0.007-0.007 yanlılık, 0.393-0.393 RMSE ve 0.296-0.296 OMH değerleri, üç kategorili sınıflamada 0.944-0.923 korelasyon, -0.001-0.0058 yanlılık, 0.341-0.414 RMSE, 0.252-0.275 OMH değerleri, dört kategorili sınıflamada 0.969-0.966 korelasyon, -0.002-0.008 yanlılık, 0.256-0.271 RMSE ve 0.183-0.191 OMH değerleri hesaplanmıştır. GA sınıflama kriterinin sınıflama kategori sayısı arttıkça ölçme kesinliği değerleri yükselmiş hata değerleri düşmüştür. KY temelli madde seçme yöntemi ile oluşturulan koşulların KN temelli madde seçme yöntemine göre daha iyi performans gösterdiği görülmektedir. AOOT sınıflama kriterinin GA sınıflama kriterinden ölçme kesinliği olarak daha iyi performans gösterdiği araştırmanın bulgularındandır.

Tartışma ve Sonuç

Bu araştırmada BBST uygulamalarında simülasyonla yetenek kestirim yöntemleri ile sınıflama kriterlerinin, madde seçme yöntemleri ile yetenek kestirim yöntemlerinin ve madde seçme yöntemleri ile sınıflama kriterlerinin çaprazlandığı iki, üç, dört kategorili sınıflama koşullarındaki performansları incelenmiştir. Araştırma sonunda çok kategorili sınıflamada test etkililiği ve ölçme kesinliği için oluşturulan koşullara en uygun desenler belirlenmiştir.

Araştırma koşullarının tamamında kategori sayısı arttıkça OTU artmış, OSD ise azalmıştır. Kategori sayısı arttıkça yapılan sınıflamalarda madde havuzunda daha az madde kaldığı için OTU' nun arttığı yorumu yapılabilir. Ayrıca kategori sayısı arttıkça ölçmenin standart hata değerleri azalmıştır başka bir ifadeyle daha hassas ölçme yapılmıştır. Bu sonuçlara göre çok kategorili sınıflama yapıldığında testin sonlanması için gereken madde sayısı arttığı için bireylerin son yetenek düzeyleri daha hassas ölçülerek belirlenmiştir ve sınıflamanın kategori sayısı hata değerlerine göre belirlenebilir yorumu yapılabilir. Araştırmanın bu bulgusu Demir (2019), Eggen(1999), Nydick ve diğerleri (2012) araştırma sonuçlarıyla benzerlik göstermektedir.

Birinci araştırma probleminde yetenek kestirimi ile madde seçme yöntemlerinin çaprazlandığı koşullar için iki kategorili sınıflamada AOK ve BSD yetenek kestirimlerinin her ikisi için de MFB ve KLB her iki madde seçme yöntemi de KN temelli madde seçme yöntemi ile daha az sayıda madde ile daha yüksek sınıflama doğruluğu ile sınıflama yaptığı sonucuna ulaşılmıştır. Ölçme kesinliği için AOK ve BSD her iki yetenek kestiriminin MFB ve KLB her iki madde seçme yöntemiyle de KY temelli madde seçme yöntemi ile daha az hata değerleri ile daha iyi performans gösterdiği araştırmanın sonuçlarındandır. İki kategorili sınıflamada KLB madde seçme yönteminin MFB madde seçme yöntemine göre test etkililiği ve ölçme kesinliği olarak performansının daha yüksek olduğu sonucuna ulaşılmıştır. Araştırmanın bu sonuçları Gündeğer (2017), Thompson (2009) araştırma sonuçlarıyla benzerlik göstermektedir. Gündeğer (2017), KY temelli madde seçme yönteminin KN temelli madde seçme yöntemine göre ölçme kesinliği olarak daha iyi performans gösterdiğini belirtmektedir.

Üç kategorili sınıflamada AOK ve BSD her iki yetenek kestirim yöntemi de KLB-MFB her iki madde seçme yönteminde KY temelli madde seçme yönteminin test etkililiği ve ölçme kesinliği olarak daha iyi performans gösterdiği sonucuna ulaşılmıştır. BSD yetenek kestiriminin MFB-KY temelli madde seçme yöntemi ile oluşturulan koşulun OTU ve OSD olarak bireyleri sınıflamada en az madde en yüksek sınıflama doğruluğu ile sınıflama yaptığı belirlenmiştir. Korelasyon, hata değerleri olarak da etkili performans gösterdiği görülmüştür.

Dört kategorili sınıflamada da KY temelli MFB ve KLB her iki madde seçme yönteminin de AOK ve BSD yetenek kestirimleri ile ölçme kesinliği ve test etkililiği olarak benzer ve iyi performans gösterdiği sonucuna ulaşılsa da BSD yetenek kestirimi ile nispeten daha az madde ile sınıflama yapıldığı için BSD yetenek kestirimi ile KLB-KY temelli madde seçme yönteminin oluşturduğu desenin performansının daha yüksek olduğu sonucu belirlenmiştir. Üç ve dört kategorili sınıflamada BSD yetenek kestiriminin AOK yetenek kestiriminden daha az madde ile sınıflama yaptığı sonucu Yi, Wang ve Ban (2000) araştırmasında AOK 'un BSD' dan daha fazla sayıda madde ile sınıflama yaptığı sonucuyla uyuşmaktadır.

İkinci araştırma probleminde AOK ve BSD her iki yetenek kestirimi ile GA sınıflama kriterinin iki kategorili sınıflamada en az madde ile sınıflama yaptığı test etkililiğinin yüksek olduğu belirlenmiştir. AOOT sınıflama kriterinin kullanıldığı koşullarda ise daha fazla madde ile hata değerlerinin düşük olduğu başka ifadeyle ölçme kesinliğinin yüksek olduğu sınıflamanın yapıldığı görülmüştür. AOOT sınıflama kriterinin düşük farksızlık bölgesi ile daha yüksek sınıflama doğruluğunda daha uzun testlerle sınıflama yapacağı özelliğinden dolayı AOOT sınıflama kriterinin daha fazla madde ile yüksek doğrulukta sınıflama yaptığı söylenebilir. Araştırmanın bu bulgusu Nydick (2012), Thompson (2009) ve Demir (2019) araştırma sonuçlarıyla da benzerlik göstermektedir.

Üçüncü araştırma probleminde GA sınıflama kriterinin KY temelli madde seçme yöntemleriyle iki kategorili sınıflamada en az madde ve yüksek sınıflama doğruluğu ile sınıflama yaptığı sonucuna

ulaşılmıştır. GA sınıflama kriteri her maddeden sonra belirlenen yetenek düzeylerini kullanarak belirlenen güven aralığını kesme puanı ile karşılaştırdığı için KY temelli madde seçme yöntemleri ile daha başarılı performans gösterdiği yorumu yapılabilir. AOOT sınıflama kriterinin ise KY temelli madde seçme yöntemleriyle birlikte kullanıldığı koşullarda düşük hata değerleri ile ölçme kesinliği yüksek sınıflama yapıldığı belirlenmiştir. GA sınıflama kriteri test etkililiği için iki kategorili sınıflamada KY temelli madde seçme yöntemleri ile en etkili deseni oluşturduğu görülmektedir. AOOT sınıflama kriteri ise sınıflama kategori sayısı arttıkça daha düşük hata değerleri ile KY temelli madde seçme yöntemleri ile etkili desen oluşturduğu sonucuna ulaşılmıştır. Araştırmanın bu sonuçları Nydic ve diğerleri (2012), Thompson ve Ro (2007) ile benzerlik göstermektedir.

Araştırmanın tüm koşullarında BSD yetenek kestiriminin AOK yetenek kestirimine göre daha düşük yanlılıkla sınıflama yaptığı görülmüştür. BBST uygulamalarında BSD yetenek kestiriminin AOK yetenek kestirimine göre ölçme kesinliği yanlılık açısından daha iyi performans gösterdiği sonucuna ulaşılmıştır. Gündeğer (2017), araştırmasında iki kategorili sınıflama için OTU ve OSD ve gerçek yeteneklerle kestirilen yetenekler arasındaki korelasyon açısından BSD ve AOK yetenek kestiriminin benzer performans gösterdiğini yanlılık ve hata değerleri olarak BSD yetenek kestirim yönteminin AOK' a göre daha etkili olduğu sonucuna ulaşmıştır. Çok kategorili sınıflamada da benzer sonuç oluştuğu bu araştırmada görülmektedir. Günder'in (2017) araştırması ile araştırmanın bu sonucu farklı sınıflama kategorilerinde de benzerlik göstermektedir.

Öneriler

Araştırma sonuçları genel olarak değerlendirildiğinde Sınıflama kategori sayısı arttıkça test etkililiğinin düştüğü ölçme kesinliğinin arttığı daha hassas ölçme yapıldığı sonucuna göre BBST amaç yüksek sınıflama doğruluğu ile az madde ile sınıflama yapmak olduğu için uygulayıcılara hata değerlerine bakılarak sınıflama kategori sayısının belirlenmesi önerilebilir.

GA sınıflama kriteri test etkililiği için AOOT sınıflama kriteri ise ölçme kesinliği için uygulayıcılara önerilebilir.

BSD yetenek kestirimi KY temelli madde seçme yöntemi ile daha az sayıda madde ile sınıflama yaptığı için AOK yetenek kestirimine göre uygulayıcılara önerilebilir.

BSD yetenek kestirimi AOK yetenek kestirimine göre daha düşük yanlılık ile yeteneği belirlediği için uygulayıcılara önerilebilir.

MFB madde seçme yöntemi KLB madde seçme yöntemine göre koşullara göre değişmekle birlikte performansı daha iyi olduğu için uygulayıcılara önerilebilir.

Araştırmacılar için ise araştırmada tek boyutlu MTK modellerinden ikili puanlamaya dayalı 3 PLM kullanılmıştır. Tüm koşullardaki başlama kuralı $\theta = 0$ olarak belirlenmiştir. Bireylere ait ön bilgiler varsa başlama kuralı olarak belirlenebilir.

İçerik dengeleme ve madde kullanım sıklığı dikkate alınmamıştır. Yetenek kestirim yöntemlerinden AOK ve BSD yöntemlerinin performansı araştırılmıştır. Maksimum olabilirlik kestirimi (MOK), Maksimum sonsal dağılım (MSD) gibi yetenek kestirim yöntemlerinin performansları da araştırmacılar tarafından araştırılabilir.

Sınıflama kriterlerinden GA ve AOOT performansları araştırılmıştır. Weiss ve Kingsbury (1984) tarafından önerilen Bireyselleştirilmiş Uzmanlık Testi (BUT), AOOT' nin daha genel bir hali olan Genelleştirilmiş olabilirlik oran (GOO), van der Linden (1990) tarafından önerilen Bayesci Karar Kuramı (BKK) sınıflama kriterlerinin performansları da araştırmacılar tarafından araştırılabilir.

Ahi Evran University
Journal of Kırşehir Education Faculty
ISSN: 2147 - 1037
http://kefad.ahievran.edu.tr

ENGLISH VERSION

Introduction

While achievement and ability tests are developed to identify ability at a point in time, classification tests are used when the aim is to make a classification decision. Computer-based tests may be preferable to paper and pencil methods for making classification decisions, especially when assessment results are high stakes. Classification test procedures evaluate a test taker against a predetermined cut score and provide a categorical result (Weiss, 1983; Wainer, 1990). Using a Computerized Adaptİve Classification Test (CACT) instead of a fixed form test, especially for multi-category classification, is very convenient for selecting items with the most appropriate characteristics for classification (Thomson, 2007). Studies have shown that more accurate classifications can be obtained with fewer items with the ability determination methods used in the computerized adaptive classification test (Lewis and Sheehan, 1990). If a test has three or four classifications (two or three cut scores), the number of test takers requiring many items increases (Spray, 1993). This, in turn, increases the average number of items needed across all scales. It increases the number of items needed in the item pool for an effective test without item exposure control (Thomson, 2007). The problem of practical constraints, such as item exposure control methods, and content balancing in CACT applications is stated from past research that the introduction of constraints is not necessary, especially in simulation studies, the introduction of practical constraints is harmful to research (Thompson, 2007).

It is thought that CACT provides more reliable classification by using fewer items than traditional tests. (Fan, Wang, Chang, and Douglas, 2012; Thompson, 2009). In Computerized Adaptive Classification Tests, the effectiveness of the test increases with a low number of items and high average classification accuracy. Low errors and high correlation between actual and predicted ability levels increase measurement accuracy (Thompson, 2009). The CACT aims to categorize individuals into classes with high classification accuracies with fewer items according to the cut score. Especially in tests whose results show high importance, accurate classification is crucial since important decisions are made, such as graduation and career choice in fields such as education and medicine (Thompson, 2007). It is important to create different conditions and determine the appropriate designs for them in CACT applications (Gündeğer, 2017). The general aim of the CACT research is to determine the appropriate

conditions for crossing item selection and ability estimation methods, classification criteria and ability estimation methods, and item selection methods and classification criteria that maximize the efficiency of CACT. To establish test efficiency by using fewer items for high classification accuracy. To increase measurement accuracy with low standard errors and to determine the most appropriate classification conditions (Thompson, 2009).

Purpose and Importance of the Research

The study aims to determine how classification accuracy and measurement precision change according to the number of classification categories in different conditions created by crossing ability estimation methods and classification criteria, ability estimation methods and item selection methods, item selection methods and classification criteria in multi-categorical classification made with CACT simulation and to determine the most appropriate design for the created conditions. In the literature abroad, there are more examples of two-category classification in which CACT conditions are crossed (Lau, 1996; Reckase, 1983; Spray and Reckase, 1996). Fewer studies (Gündeğer, 2017; Demir, 2019) are encountered in Turkey. Gündeğer (2017) examined the performance of the ability estimation methods, item selection methods, and classification criteria for two-category classification with the conditions created. Demir (2019) examined the performances of item exposure control methods and different content balancing methods for the conditions he determined for multi-categorical classification.

In the literature, there are few examples of studies investigating the performance of ability estimation methods, classification criteria and item selection methods in multi-categorical classification, especially the performance of conditions formed by crossing ability estimation methods and ability estimation methods, item selection methods and classification criteria with Kullback Leibler Information (KLI) methods based on cut score (CB) and estimated ability (EB) item selection methods. This study is expected to provide information to practitioners about the performance of the conditions formed by crossing ability estimation methods and classification criteria for two, three and four-category classification and crossing ability estimation methods and item selection methods, item selection methods and classification criteria in terms of measurement precision and test efficiency and to determine the most appropriate design for the conditions determined. Therefore, it is thought to contribute to the literature. In this study, two, three and four-category classification was made on 1000 individuals with an item pool of 500 one-dimensional items. The study investigated average classification accuracy and average test length in terms of test efficiency, correlation between real and estimated abilities in terms of measurement precision, bias, RMSE, MAE values, ability estimation methods with classification criteria, ability estimation methods with item selection methods and item selection methods with classification criteria. Additionally, the study explored how their performances change according to the number of classification categories. Thompson (2007) believes that practical constraints are detrimental to simulation studies; therefore, the study didn't use practical constraints.

Research Problems

The research problems were as follows.

1. How do the average test length, average classification accuracy and test efficiency change in the two, three and four-category classification with CACT simulation when evaluating weighted likelihood estimation (WLE) and expected a posterior (EAP) ability estimation methods with the maximum fisher information (MFI) based on estimated ability and cut score, Kullback Leibler Information (KLI) based on cut score and estimated ability (CB-EB) item selection methods?

2. How do the ACA, ATL, bias, r, RMSE, and MAE values change for the classification which evaluates WLE and EAP ability estimatıon methods with SPRT δ: 0.1 indifference region and CI classification criteria with 90% confidence level in two, three, four-category classification with CACT simulation?

3. In two, three and four category classification with CACT simulation, how do the ACA, ATL, bias, r, RMSE, and MAE values change when the classification criteria are SPRT δ:0.1 region of indifference, CI 90% confidence level and when evaluating the estimated ability and cut score based MFB-KLB item selection methods with the classification criteria?

Method

The research which was a simulation study and created ability parameters and item pool parameters in R software (R Core Team, 2013). Moreover, the research created two, three and four-category classification with 2 ability estimation methods, 4 item selection methods and 2 classification criteria. In other words, it produced 2 ability estimation methods x 4 item selection methods x 2 classification criteria x 3 classification categories = 48 conditions.

Data generation

The study produced an item pool of 500 items. For the item pool, the study considered the studies by Thompson (2009, 2011) and Weiss (1980). Also, it generated the a parameter of the items from the U (0.5,1.5) distribution and the b parameter from the N (0,1) distribution considering the study by Weiss (1980). Additionally, the study created the c parameter as N (0,0.3) considering the study byThompson (2009). The study also generated the ability parameter in the R software with a mean of 0 and a standard deviation of 1 for 1000 individuals. Moreover, the study used weighted likelihood estimation and expected posteriori ability estimation methods from Bayesian ability estimation methods for ability estimation. It also used item selection methods based on maximum fisher information cut score and estimated ability and item selection methods based on Kulback laiber cut score and estimated ability. The indifference region (İR) for the classification criterion is the level of uncertainty that can be tolerated for classification decisions close to the cut score. According to Thompson (2011), classification accuracy is considered to be high if the region of indifference is small. According to Eggen and

Straetmans (2000), as the confidence interval value increases, the number of items required for classification and the accuracy of classification increases. This study, taking into account the study by Nydick (2013) and Eggen and Straetmans (2000), used CI classification criteria with SPRT δ: 0.1 indifference region and 90% confidence level.

Simulation Conditions

For the computerized classification test, the study performed analyses in the R software with an average of 25 replications of the conditions in which 2 ability estimation methods were crossed with 2 classification criteria and 2 ability estimation methods were crossed with 4 item selection methods in two, three, four-category classification (R Core Team, 2013). Classification criteria were chosen according to the research by Nydick (2013) and Eggen and Straetmans (2000) with the methods of Sequential Probability Ratio Test (SPRT) with 0.1 indifference region, Confidence Interval (CI) with a 90% confidence level. According to the literature, the indifference zone and confidence interval value indicate the tolerable level of error. As the confidence interval value increases and the smaller the indifference zone constant, the number of items required for the test to classify and the accuracy of the classification increase (Eggen and Straetmans, 2000). Research in literature hasn't studied Bayesian ability estimation methods among the ability estimation methods. According to Warm (1989), Weighted Likelihood Estimation (WLE) is a method that works on weighting likelihood that reduces bias and estimates ability when all items are answered correctly or incorrectly. The study used Weighted likelihood estimation (WLE) and Expected a posterior (EAP) ability estimation methods from Bayesian ability estimation. CACT include cut score-based and ability-based item selection methods. This study investigated the performances of maximum fisher information MFI (CB-EB), cut score and based on estimated ability Kullback-leibler KLI (CB-EB) item selection methods when crossed with ability estimation methods. According to Thompson (2007), ability level 0 is the starting point, or predetermined ability levels can be used. Furthermore, the study created simulation conditions according to the research problems. According to Eggen and Straetmans (2000), the cut-off points for two, three, and four-category classification can be determined by dividing the ability levels into two and taking 70% of each level as the first part as level 1 and the second part as level 2. The cut score can also be determined randomly. This study determined the cut score considering the studies by Eggen and Straetmans (2000) and usedCatIRT (Nydick, 2014) package.

Analyzing the data

The research performed analyses for 48 simulation conditions in two, three, and four-category classifications with functions written in R by averaging 25 repetitions to obtain the closest results to the real application. For classification accuracy, the study calculated average classification accuracy and average test length values. In addition, the study calculated precision RMSE, MAE, bias, and correlation (r) between real and estimation thetas levels for measurement. It also calculated the Pearson correlation

coefficient value for the correlation (r) between real and estimation thetas levels. Besides, it calculated the agreement between the real classes and the simulated classes by Cohen's Kappa statistic for ACA.

Bias was equal to the ratio of the sum of the differences of the final ability levels ($\theta i\hat{}$) from the real ability levels (θi) to the number of individuals (n) (Miller and Miller, 2004).

$$\text{Bias}=\frac{\sum_{i=1}^{n}(\widehat{\theta_1}-\theta_i)}{n} \quad \textbf{(1)}$$

RMSE was equal to the square root of the ratio of the sum of squares of the differences of the final estimated ability levels ($\theta i\hat{}$) from the real ability levels (θi) to the number of individuals (n)

$$\text{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(\widehat{\theta_1}-\theta_i)^2}{n}} \quad \textbf{(2)}$$

MAE was equal to the ratio of the sum of the absolute values of the differences of the estimated final ability levels ($\theta i\hat{}$) from the real ability levels (θi) to the number of individuals (n).

$$\text{MAE}=\frac{\sum_{i=1}^{n}|\widehat{\theta_1}-\theta_i|}{n} \quad \textbf{(3)}$$

Research Ethics Permissions

This study followed all the rules specified in the "Directive on Scientific Research and Publication Ethics of Higher Education Institutions". However, the study carried out none of the actions specified under the second section of the Directive, "Actions Contrary to Scientific Research and Publication Ethics".

Findings

The first question of the research examined how the average classification accuracy, average test length, bias, RMSE and MAE values of the classification conditions in which WLE and EAP ability estimation methods were evaluated with CB and EB, MFI and KLI item selection methods in two, three and four category classification with CACT simulation changed according to the number of classification categories. Table 1 shows the values for the conditions in which ability estimation and item selection methods were crossed.

According to Table 1, the values were 19.76-17.91 for ATL and 0.89-0.891 for ACA according to the conditions in which the WLE ability estimation method was crossed with MFI (EB-CB) item selection methods in two-category classification. The WLE ability estimation made classification with fewer items in the two-category classification with the MFI-CB item selection method. Hence, the study obtained similar results with both item selection methods for ACA.

Table 1. *ATL, ACA, r, bias, RMSE and MAE values according to the number of classification categories for the conditions in which ability estimation methods and item selection methods were crossed*

Conditions			Dependent Variables					
Ability Estimation Methods	Item Selection Methods	CC	ATL	ACA	r	Bias	RMSE	MAE
WLE	MFI-EB	Two	19.76	0.891	0.945	0.002	0.337	0.241
		Three	26.29	0.877	0.963	-0.001	0.28	0.201
		Four	36.07	0.868	0.976	0	0.225	0.16
	MFI-CB	Two	17.91	0.89	0.844	0.35	0.706	0.484
		Three	28.50	0.89	0.929	0.117	0.427	0.269
		Four	37.82	0.869	0.972	-0.004	0.245	0.17
	KLI-EB	Two	19.87	0.887	0.947	-0.003	0.332	0.238
		Three	26.28	0.877	0.963	-0.008	0.28	0.201
		Four	35.98	0.869	0.976	-0.002	0.226	0.16
	KLI-CB	Two	17.68	0.89	0.845	0.346	0.699	0.478
		Three	28.45	0.891	0.93	0.116	0.425	0.267
		Four	37.78	0.869	0.972	-0.005	0.245	0.17
EAP	MFI-EB	Two	19.83	0.89	0.95	-0.001	0.324	0.234
		Three	26.12	0.878	0.963	-0.003	0.281	0.202
		Four	36.32	0.871	0.976	0	0.225	0.158
	MFI-CB	Two	17.18	0.889	0.851	0.009	0.543	0.393
		Three	27.70	0.892	0.932	0.002	0.375	0.256
		For	36.75	0.871	0.971	0	0.25	0.173
	KLI-EB	Two	19.94	0.888	0.951	-0.001	0.319	0.231
		Three	26.20	0.878	0.963	-0.002	0.278	0.2
		Four	36.25	0.871	0.976	0.001	0.225	0.159
	KLI-CB	Two	17.05	0.888	0.854	0.009	0.539	0.39
		Three	27.69	0.892	0.933	0.003	0.372	0.253
		Four	36.68	0.87	0.971	0.01	0.25	0.174

Average test length: ATL, Average classification accuracy: ACA, Mean absolute error :MAE, Root mean square error (RMSE), Classification Categoria: CC

In terms of bias, RMSE, MAE and the correlation between real abilities and estimated abilities, i.e. measurement precision, 0.945-0.844 correlation value, 0.002-0.35 bias, 0.337-0.706 RMSE, 0.241-0.484 MAE values were calculated when WLE ability estimation method was crossed with MFI (EB-CB) item selection methods. The WLE ability estimation with the MFI-EB item selection method performed better than the CB item selection method with lower error values for measurement precision.

Considering the table, when the WLE ability estimation and KLI item selection method were used together in two-category classification, 19.87-17.69 ATL and 0.887- 089 ACA values were obtained

with the EB-CB item selection methods. Accordingly, with the CB item selection method, classification was with fewer items and higher classification accuracy. According to the study findings, the KLI-CB item selection method and the condition in which the WLE ability estimation was crossed with the KLI-CB item selection method had a higher performance for test efficiency. For measurement precision, the KLI (EB-CB) item selection method calculated the values as 0.947-0.845 for correlation, -0.003-0.346 for bias, 0.332-0.699 for RMSE and 0.238-0.478 for MAE. Accordingly, in the two-category classification in terms of measurement precision, considering the values calculated in Table 1, the error values and bias values were lower when the KLI item selection method was used together with WLE ability estimation, and the performance of the EB item selection method was higher. Table 1 shows the values for the conditions in which the ability estimation methods and item selection methods were crossed.

In the simulation in which the EAP ability estimation method was crossed with the MFI (EB-CB) item selection methods in two-category classification, the values were 19.83-17.18 for ATL and 0.89-0.889 for ACA. Accordingly, the study found that the CB item selection method made classification with fewer items. Both item selection methods performed similarly when used together with EAP ability estimation. For CACT, where classification with fewer items was the goal for test efficiency, MFI (CB) method was - effective in two-category classification with EAP ability estimation. For measurement precision, the values were 0.95-0851for correlation, -0.001-0.009 for bias, 0.324-0.543 for RMSE, and 0.234-0.393 for MAEwhen the EB-CB item selection methods and EAP ability estimation were used together. Accordingly, the performance of the MFI-EB item selection method with EAP ability estimation was higher in the two-category classification as measurement precision.

In the two-category simulation where EAP ability estimation and KLI (EB-CB) item selection methods were crossed, the values were 19.94-17.05 for ATL and 0.888-0.888 for ACA. Hence, the KLI-CB item selection method performed better than the PB-based item selection method with fewer items and similar classification accuracy. Measurement values were 0.951-0.854 for correlation, -0.001-0.009 for bias, 0.319-0.539 for RMSE, and 0.231-0.39 for MAE. Considering the table, KLI-CB item selection methods performed better in terms of test efficiency and KLI-EB item selection methods performed better in terms of measurement precision.

When the WLE ability estimation for three-category classification was crossed with MFI (EB-CB) based item selection methods, ATL values were calculated as 26.29-28.50 and ACA as 0.877-0.89. Accordingly, similar findings were observed for the item selection methods, although the item selection method based on EB performed relatively better in terms of test length than the item selection method based on CB in terms of classification accuracy. For measurement precision, the values were 0.963-0.929 for correlation, -0.01-0.117 for bias, 0.28-0.427 for RMSE, and 0.201-0.269 for MAE values when WLE ability estimation and MFI (EB-CB) item selection methods were used together. Accordingly, the MFI-EB item selection method provided classification with lower error and higher correlation. The EB item

selection method performed better than the CB item selection method with WLE ability estimation regarding measurement accuracy.

When WLE ability estimation was crossed with KLI (EB-CB) item selection methods, the values were 26.28-28.45 for ATL and 0.877-0.891 for ACA values. The EB item selection method classified fewer items with slightly higher classification accuracy than the CB item selection method. For measurement precision, WLE ability estimation was used with KLI (EB-CB) item selection methods with the values 0.963-0.93 for correlation, -0.008-0.116 for bias, 0.28-0.425 for RMSE, and 0.201-0.267 for MAE. The EB item selection method produced classification with lower error, bias and higher correlation than the WLE ability estimation. For test efficiency and measurement precision, the performance of the KLI-EB item selection method was higher.

When EAP ability estimation was used in the three-category classification, the values were 26.12-27.70 for ATL and 0.878-0.892 for ACA in the simulation created with MFI (EB-CB) item selection methods. When the EAP ability estimation method was used together with the EB item selection method, classification was fewer items and with higher ACA when used together with the CB item selection method. For measurement precision, the values were 0.963-0.932 for correlation, -0.003-0.002 for bias, 0.281-0.375 for RMSE, and 0.202-0.256 for MAE in the classification of the conditions in which EAP ability estimation and MFI (EB-CB) item selection methods were used together. The EAP ability estimation and the EB item selection method performed better than the CB item selection method regarding measurement precision.

In the three-category classification where EAP ability estimation was used together with KLI (EB-CB) item selection methods, the values were 26.20-27.69 for ATL and 0.878-0.892 for ACA. Hence, the EB item selection method performed better for ATL and the CB item selection method performed better for ACA. Similarly, the values were 0.963-0.933 for correlation, -0.002-0.003 for bias, 0.278-0.373 for RMSE, and 0.2-0.253 for MAE for measurement precision. Moreover, the EB item selection method performed better than the CB item selection method with low errors and bias and high correlation.

In the four-category classification, the values were 36.07-37.82 for ATL and 0.868-0.869 for ACA for the conditions crossing the WLE ability estimation with the MFI (EB-CB) item selection methods. Additionally, classification was made by using fewer items with the EB item selection method. For the test effectiveness, close values were calculated for ACA with both item selection methods, and the performance of the condition created with the MFI-EB item selection method was better. In the conditions created with MFI (EB-CB) item selection method, the values were 0.976-0.972 for correlation, 0,-0.004 for bias, 0.225-0.245 for RMSE, 0.16-0.17 for MAE. Considering the table, the item selection method based on EB had a lower error and higher correlation regarding measurement accuracy.

In the crossover with KLI (EB-CB) item selection methods, the values were 35.98-37.78 for ATL and 0.869-0.869 for ACA -. According to the condition created with the KLI-CB item selection method,

classification was made with fewer items and high classification accuracy. In the conditions created with KLI (EB-CB) item selection methods, the values were 0.976-0.972 for correlation, -0.02,-0.005 for bias, 0.226-0.245 for RMSE, and 0.16-0.17 for MAE values. Hence, the condition created with the KLI-EB item selection method had higher correlations and lower standard errors. It is clear from the table that the item selection method based on EB performed better in terms of measurement precision.

For the four-category classification, the values were 36.32-36.75 for ATL and 0.871-0.871 for ACA values in the conditions created by crossing the EAP ability estimation with the MFI (EB-CB) item selection methods. Additionally, EAP ability estimation with both item selection methods showed similar performance. For the same condition, the values were 0.976-0.971 for correlation, 0-0 for bias, 0.225-0.25 for RMSE, and 0.158-0.173 for MAE values. Also, the performance of the EB item selection method with EAP ability estimation was higher than the CB item selection method regarding measurement precision.

Although the item selection methods with 36.25-36.68 for ATL and 0.871-0.87 for ACA showed similar performance in the conditions created for the four-category classification from the crossover of the KLI (EB-CB) item selection methods with EAP ability estimation, the values were 0.976-0.971 for correlation, 0.001-0.1 for bias, 0.225-0.25 for RMSE, and 0.159-0.174 for MAE values regarding measurement precision. Hence, the item selection method based on EB provided more precise measurement with lower error. It was one of the study findings that the error values decreased as the number of classification categories increased, in other words, the measurement precision increased. As the number of classification categories increased, the ACA and the ATL value increased, in other words, the test efficiency decreased. Considering the table, the MFI-EB item selection method performed better than the WLE ability estimation method in terms of measurement precision, while the KLI item selection method performed better in terms of ACA and ATL. Additionally, the study found that the KLI-EB item selection method and EAP ability estimation performed better for test efficiency.

The second research problem analyzed how the average classification accuracy, average test length, bias, RMSE and OHM values changed according to the number of classification categories in the two, three and four-category classification with CACT simulation for the classification conditions where WLE and EAP ability estimation methods were evaluated with SPRT (δ: 0.1) and CI (90%) classification criteria. Table 2 shows the values for the conditions in which the ability estimation methods and classification criteria were crossed.

In the two-category classification of the conditions in which the WLE ability estimation was crossed with the SPRT (δ:0.1) classification criterion, the RMSE was 33.502, and the MAE was 0.897. For the three-category classification, ATL was 36.346 and ACA was 0.895, and for the four-category classification, ATL was 48.304 and ACA was 0.875. As the number of classification categories increased, ATL increased and ACA value decreased. Besides, Bias, RMSE, and MAE values increased as the

number of classification categories increased, and the correlation between actual and predicted abilities decreased.

Table 2. *ATL, ACA, bias, correlation, and RMSE, MAE values for the conditions in which the classification criteria are crossed with the ability estimation methods*

Conditions			Dependent Variables					
Item Selection Methods	Classification Criteria	CC	ATL	ACA	r	Bias	RMSE	MAE
WLE	SPRT	Two	33.502	0.897	985	0.002	0.183	0.141
		Three	36.346	0.895	0.941	0.114	0.392	0.239
		Four	48.304	0.875	0.981	-0.004	0.198	0.141
	CI	Two	11.324	0.877	0.921	-0.004	0.404	0.303
		Three	24.28	0.884	0.93	0.106	0.422	0.272
		Four	31.562	0.862	0.969	-0.015	0.255	0.182
EPD	SPRT	Two	33.078	0.897	0.985	-0.001	0.181	0.139
		Three	36.092	0.896	0.948	0.004	0.33	0.22
		Four	47.906	0.875	0.981	-0.001	0.199	0.141
	CI	Two	12.009	0.879	0.928	-0.002	0.386	0.291
		Three	22.357	0.888	0.924	0.001	0.396	0.275
		Four	29.001	0.866	0.964	0	0.277	0.196

Average test length: ATL, Average classification accuracy: ACA, Mean absolute error :MAE, Root mean square error (RMSE), Classification Categoria: CC

For the conditions in which the WLE ability estimation was crossed with the CI (90%) classification criterion, the RMSE was 11.324 for the two-category classification, 24.28 for the three-category classification, and 31.562 for the four-category classification. In the two-category classification, classification was made with the minimum number of items.

In the study, the CI (90%) classification criterion showed the highest performance with 0.884 in the three-category classification for ACA. Additionally, the error values such as bias, RMSE, and MAE in the conditions created by crossing the WLE ability estimation with the CI (90%) classification criterion were calculated at the lowest values in the four-category classification. The highest correlation value between actual abilities and predicted abilities was calculated in the four-category classification. The study found that the conditions created in terms of the accuracy of the measurement performed well in the four-category classification. In terms of the effectiveness of the test, the conditions performed more effectively in the two-category classification.

When the EAP ability estimation was crossed with SPRT (δ:0.1) classification criterion, the values were 33.078 for ATL in two-category classification, 36.092 for ATL in three-category classification, and 47.906 for ATL in four-category classification. Hence, as the number of classification categories increased, the ATL increased. The study calculated similar values for two and three-category classifications as ACA. The study found that the error values in terms of measurement accuracy were lower in the two-category classification.

The EAP ability estimation made the classification with the least number of items with 12.009 ATL when crossed with the CI (90%) classification criterion. As the number of classification categories increased, the ATL value increased and the ACA value decreased. that the study found that the performance of the EAP ability estimation with high accuracy and low error was in four-category classification with CI (90%) classification criterion.

The third research problem examined how the conditions in which the classification criteria were evaluated with item selection methods in two, three and four-category classification changed regarding measurement precision and test effectiveness. Table 3 shows the values of the conditions in which item selection methods and classification criteria were crossed.

Table 3, which was created according to the values obtained by averaging 25 repetitions for all conditions of the third problem of the research shows that for both of the MFI (EB-CB) item selection methods, the values were 33.09-33.09 for ATL and 0.897-0.897 for ACA in two-category classification, 37.91-33.26 for ATL and 0.085-0.895 for ACA in three-category classification, and 37.91-33.26 for ATL and 0.085-0.895 for ACA in four-category classification. Besides, the values were 09-33.09 for ATL and 0.897-0.897 for ACA in two-category classification, 37.91-36.26 for ATL and 0.085-0.895 for ACA in three-category classification, and 46.69-48.11 for ATL and0.874-0.874 for ACA in four-category classification. In the conditions created with CI (90%) classification criterion, 11.61-11.61 SPRT, 0.876-0.876 ACA values were calculated for two-category classification, 17.93-22.95 SPRT, 0.867-0.885 ACA values for three category classification, 28.29-30.08 SPRT, 0.863-0.864 ACA values for four category classification. Also, the CI classification criterion made classification with fewer items than the SPRT classification criterion. The study found that the SPRT classification criterion performed classification with higher average classification accuracy values. It also determined that as the number of classification categories increased, the ATL increased and the ACA decreased. Although the classification criteria showed similar performance with EB and CB item selection methods, the performance of the MFI-EB item selection method was higher than the SPRT classification criterion and CB item selection method. The values showing the standard error of measurement for the conditions in which the MFI (EB-CB) item selection methods were crossed with the classification criteria were 0.984-0.984 for correlation, -0.001- -0.001 for bias, 0.182-0.182 for RMSE, and 0.14-0.14 for MAE in two-category classification, 0.985-0.944 for correlation, -0.001-0.061 for bias, 0.18-0.363 for RMSE, and 0.137-0.228 for MAE in three-category

classification, and 0.987-0.981 for correlation, 0,-0.003 for bias, 0.169-0.201 for RMSE, and 0.136-0.141 for MAE in four-category classification.

Table 3. *The values for the conditions in which item selection methods and classification criteria are crossed.*

Conditions			Dependent Variables					
Item selection method	Classification Criteria	CC	ATL	ACA	r	Bias	RMSE	MAE
MFB-KY	SPRT(	Two	33.09	0.897	0.984	-0.001	0.182	0.14
		Three	37.91	0.085	0.985	-0.001	0.18	0.137
		Four	46.69	0.874	0.987	0	0.169	0.136
	CI	Two	11.61	0.876	0.923	-0.002	0.398	0.299
		Three	17.93	0867	0.944	-0.003	0.342	0.253
		Four	28.29	0.863	0.969	-0.001	0.256	0.182
MFB-KN	SPRT	Two	33.09	0.897	0.984	-0.001	0.182	0.14
		Three	36.26	0.895	0.942	0.061	0.363	0.228
		Four	48.11	0.874	0.981	-0.003	0.201	0.141
	CI	Two	11.61	0.876	0.923	-0.002	0.398	0.299
		Three	22.95	0.885	0.922	0.057	0.417	0.278
		Four	30.08	0.864	0.966	-0.008	0.271	0.192
KLB-KY	SPRT	Two	33.38	0.895	0.985	-0.001	0.181	0.139
		Three	37.94	0.884	0.985	-0.001	0.179	0.137
		Four	46.70	0.874	0.987	0	0.169	0.127
	CI	Two	11.57	0.876	0.925	-0.007	0.393	0.296
		Three	18.03	0.868	0.944	-0.01	0.341	0.252
		Four	28.12	0.864	0.969	-0.002	0.256	0.183
KLB-KN	SPRT	Two	33.38	0.895	0.985	-0.001	0.181	0.139
		Three	36.22	0.895	0.922	0.057	0.417	0.278
		Four	48.15	0.874	0.981	-0.004	0.202	0.141
	CI	Two	11.57	0.876	0.925	0.007	0.393	0.296
		Three	22.89	0.885	0.923	0.058	0.414	0.275
		Four	29.97	0.863	0.966	-0.008	0.271	0.191

Average test length: ATL, Average classification accuracy: ACA, Mean absolute error :MAE, Root mean square error (RMSE), Classification Categoria: CC

The study explored that the SPRT classification criterion performed better than the EB item selection method regarding measurement precision. Hence, error values decreased as the number of classification categories increased. With the CI (90%) classification criterion, the values were 0.923-0.923 for correlation, -0.002- -0.002 for bias, 0.398-0.398 for RMSE, and 0.299-0.299 for MAE in two-category classification, and 0.944-0.922 for correlation, -0.003- -0.57 for bias, 0.342-0.363 for RMSE, and 0.253-0.278

for MAE in three-category classification. Although the CI classification criterion showed similar performance with the EB and CB item selection methods, the performance of the classification made with the EB item selection method was higher, and the error values decreased as the number of classification categories increased. Hence, the performance of the CI classification criterion in terms of test efficiency was higher than the SPRT classification criterion, and the performance of the SPRT classification criterion in terms of measurement accuracy was higher than the CI classification criterion.

In conditions where KLB (KY-KN) item selection methods and SPRT (δ:01) and CI (90%) classification criteria were crossed, the values were 33. 38-33.38 for ATL and 0.895-0.895 for ACA in two-category classification, 37.94-36.22 for ATL and 0.884-0.895 for ACA in three-category classification, and 46.70-48.15 for ATL and 0.874-0.874 for ACA in-four category classification. With the CI (90%) classification criterion, the values were 11.57-11.57 for ATL and 0.876-0.876 for ACA in two-category classification, 18.03-22.89 for ATL and 0.868-0.885 for ACA in three-category classification, and 28.12-29.97 for ATL and 0.864-0.863 for ACA in four-category classification. Hence, the CI classification criterion made classification with the least number of items in the two-category classification. In terms of test efficiency, the CI classification criterion showed higher performance than the SPRT classification criterion. As the number of classification categories increased, ATL increased and ACA decreased. Although the item selection methods based on EB and CB showed similar performance, the study found that the item selection method based on EB was more effective in terms of test efficiency.

In the error values of the estimation showing the measurement accuracy, according to the conditions in which KLI (EB-CB) item selection methods and SPRT (δ: 01) classification criterion were crossed, the values were 0.985-0.985 for correlation, -0.001- -0.001 for bias, 0.181-0.181 for RMSE, and 0.139-0. 139 for MAE values, 0.985-0.922 for correlation, -0.001- -0.057 for bias, 0.179-0.414 for RMSE, and 0.137-0.278 for MAE in three-category classification, and 0.987-0.981 for correlation, 0,-0.004 for bias, 0.169-0.202 for RMSE, and 0.202-0.127 for MAE in four-category classification. Error values decreased as the number of classification categories increased and measurement precision increased. Although the performances of the item selection methods were similar when the EB and CB item selection methods and the SPRT classification criterion were used together, the study determined that that the conditions created with the EB item selection method made classification with less error. With the CI classification criterion, the values were 0.925-0.925 for correlation, -0.007-0.007 for bias, 0.393-0.393 for RMSE, and 0.296-0.296 for MAE in two-category classification, 0.944-0.923 for correlation, -0. 001-0.0058 for bias, 0.341-0.414 for RMSE, and 0.252-0.275 for MAE values in three-category classification, and 0.969-0.966 for correlation, -0.002-0.008 for bias, 0.256-0.271 for RMSE, and 0.183-0.191 for MAE in four-category classification. As the number of classification categories of the CI classification criterion increased, measurement accuracy values increased and error values decreased. It was clear that the conditions created with the EB item selection method performed better than the CB item selection method. The

study found that the SPRT classification criterion performed better than the CI classification criterion regarding measurement precision.

Discussion and Conclusion

This study examined the performances of ability estimation methods and classification criteria, ability estimation methods and item selection methods, ability estimation methods and item selection methods, and item selection methods and classification criteria in two, three, and four-category classification conditions by simulation in CACT applications. At the end of the research, the most appropriate designs for the conditions created for test efficiency and measurement precision in multi-categorical classification were determined.

In all of the research conditions, as the number of classification categories increased, ATL increased and ACA decreased. Hence, as the number of categories increased, the ATL increased because fewer items remained in the item pool. In addition, as the number of categories increased, the standard error values of the measurement decreased, in other words, more precise measurement was made. According to these results, since the number of items required for the end of the test increased when a multi-category classification was made, the final ability levels of individuals were determined by measuring more precisely, and the number of categories of the classification could be determined according to the error values. This study finding is similar to the results by Demir (2019), Eggen (1999), and Nydick et al.

The first research problem, for the conditions in which ability estimation and item selection methods were crossed, concluded that both WLE and EAP ability estimations performed better than both MFI and KLI item selection methods with higher classification accuracy with fewer items in two-category classification. For measurement accuracy, WLE and EAP, both ability estimation methods, performed better with less error values than MFI and KLI, both item selection methods and CB item selection method. The two-category classification concluded that the performance of the MFI item selection method was higher than the KLI item selection method in terms of test efficiency and measurement precision. These study results are similar to those of Gündeğer (2017) and Thompson (2009). Gündeğer (2017) stated that the item selection method based on the EB performed better in measurement precision than the item selection method based on the CB.

The three-category classification, WLE and EAP both ability estimation methods and KLI-MFI both item selection methods, concluded that the EB item selection method performed better in test efficiency and measurement precision. Besides, the condition created with the MFI-EB item selection method of EAP ability estimation classified individuals as ATL and ACA with the least number of items and the highest classification accuracy. The study also found that it performed effectively regarding correlation and error values.

In the four-category classification, although both item selection methods of EB-MFI and KLI performed similarly and well in terms of measurement precision and test efficiency with WLE and EAP ability estimations, the performance of the pattern formed by the KLI-EB item selection method with EAP ability estimation was higher since the classification was made with relatively fewer items with EAP ability estimation. In three and four-category classification, the result that EAP ability estimation made classification with fewer items than WLE ability estimation was consistent with the result by Yi, Wang and Ban (2000) that WLE made classification with more number of items than EAP.

The second research problem determined that the CI classification criterion with both ability estimations of WLE and EAP and the CI classification criterion made classification with the least number of items in the two-category classification and the test efficiency was high. In the conditions where the SPRT classification criterion was used, classification was made with more items with low error values, in other words, with high measurement accuracyThus, the SPRT classification criterion performed high accuracy classification with more items due to the feature that the SPRT classification criterion would classify with longer tests at higher classification accuracy with low indifference region. This study finding is similar to the results by Nydick (2012), Thompson (2009) and Demir (2019).

The third research problem concluded that the CI classification criterion made classification with the least number of items and high classification accuracy in two-category classification with EB item selection methods. Since the CI classification criterion compared the confidence interval determined by using the ability levels determined after each item with the cut score, it performed more successfully with EB item selection methods. Additionally, the study determined that the SPRT classification criterion was used in conjunction with the EB item selection methods to classify items with low error values and high measurement accuracy. Also, the CI classification criterion formed the most effective pattern with the EB item selection methods in the two-category classification for test effectiveness. The SPRT classification criterion, on the other hand, formed an effective pattern with the item selection methods based on EB with lower error values as the number of classification categories increased. These study results are similar to those of Nydic et al. (2012) and Thompson and Ro (2007).

In all conditions of the study, EAP ability estimation was found to classify with lower bias than WLE ability estimation. The study concluded that EAP ability estimation performed better than WLE ability estimation regarding measurement accuracy bias in CACT applications. Gündeğer (2017) concluded that EAP and WLE ability estimation performed similarly in terms of ATL and ACA for two-category classification and correlation between real and estimated abilities, but EAP ability estimation method was more effective than WLE in terms of bias and error values. The study showed a similar result in multi-categorization. Günder's (2017) study and this study result showed similarities in different classification categories.

Recommendations

According to evaluations of the research results, as the number of classification categories increased, test effectiveness decreased, measurement accuracy increased and more precise measurement was made. Since the aim of CACT was to classify with fewer items with high classification accuracy, it can be recommended that practitioners determine the number of classification categories by looking at the error values.

CI classification criterion can be recommended to practitioners for test efficiency and the SPRT classification criterion for measurement precision.

EAP ability estimation can be recommended to practitioners compared to WLE ability estimation since it performs classification with fewer items by using the EB item selection method.

EAP ability estimation can be recommended to practitioners since it determines ability with a lower bias than WLE ability estimation.

The MFI item selection method can be recommended to the practitioners since its performance is better than the KLI item selection method, although it varies according to the conditions.

The study used 3 PLM based on binary scoring from unidimensional IRT models. The starting rule in all conditions was set as $\theta = 0$. If there is preliminary information about the individuals, it can be determined as the starting rule.

This study didn't consider content balancing and item exposure control methods. However, it investigated the performance of WLE and EAP methods among the ability estimation methods. Researchers can investigate the performances of ability estimation methods such as maximum likelihood estimation (MLE) and maximum posterior distribution (MPD).

Among the classification criteria, the study performed the performances of CI and SPRT. Researchers can investigate the performances of classification criteria such as Individualised Expertise Test (IET) proposed by Weiss and Kingsbury (1984), Generalized Likelihood Ratio (GLR), which is a more general version of SPRT, Bayesian Decision Theory (BDT) proposed by van der Linden (1990).

Kaynakça

Demir, S. (2019). *Bireyselleştirilmiş bilgisayarlı sınıflama testlerinde sınıflama doğruluğunun incelenmesi* (Yayınlanmış Doktora Tezi). Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara. https://tez.yok.gov.tr/UlusalTezMerkezi/

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*, 713-734. https://doi.org/10.1177/00131640021970862

Fan, Z., Wang, C., Chang, H., & Douglas, J. (2012). Utilizing Response Time Distributions for Item Selection in CAT. *Journal of Educational and Behavioral Statistics, 37*(5), 655-670. https://doi.org/10.3102/1076998611422912

Gündeğer, C. (2017*).Bireyselleştirilmiş bilgisayarlı sınıflama testi kriterlerinin sınıflama doğruluğu ve test uzunluğu açısından karşılaştırılması*.(Yayımlanmamış Doktora Tezi). Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara. https://tez.yok.gov.tr/UlusalTezMerkezi/

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

Lewis, C. & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement*, *14*, 367-386.

Miller, I., & Miller, M. (2004). *John E. Freund's Mathematical Statistics with Applications*. (7th Edition). New Jersey: Prentice Hall.

Nydick, S. W. (2013). *Multidimensional mastery testing with CAT. Unpublished Doctoral Dissertation*. University of Minnesota, USA.

Nydick, S. W. (2014). *catirt: An R Package for Simulating IRT-Based Computerized Adaptive Tests*.

https://cran.rproject.org/web/packages/catIrt/catIrt.pdf

R Core Team (2013). R: A language and environment for statistical computing, (Version 3.0.1) [Computer software], Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.Rproject.org/

Spray, J. A. (1993). *Multiple-category classification using sequential probability ratio test.* ACT Research Report Series, 93-7.

Thompson, N. A. (2007). *A comparison of two methods of polytomous computerized classification testing for multiple cutscores.* Unpublished doctoral dissertation, University of Minnesota, Twin Cities.

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69*(5), 778-793. https://doi.org/ 10.1177/0013164408324460

Thompson, N. A. (2011). Termination criteria for computerized classification testing. *Practical Assessment, Research & Evaluation, 16*(4), 1-7. https://pareonline.net/getvn.asp?v=16&n=4 adresinden erişilmiştir.

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, 129-156. Massachusetts: Kluwer-Nijhof.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54(3), 427-450.* doi: 10.1007/BF02294627

Wainer, H. (Ed.) (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Weiss, D. J. (Ed.) (1983). *New horizons in testing*. New York: Academic Press.

Weiss, D. J. & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*(4), 361-375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x