Uludağ Üniversitesi Eğitim Fakültesi Dergisi / Journal of Uludag University Faculty of Education
e-ISSN: 2667-6788 https://dergipark.org.tr/tr/pub/uefad
Yıl **2024,** Cilt **37**, Sayı **1**, **63-85** | DOI: https://doi.org/10.19171/uefad.1357800 **Araştırma Makalesi | Research Article**

Bilgisayarda Bireyselleştirilmiş Sınıflama Testinde Çok Kategorili Sınıflama İçin Sınıflama Koşullarının İncelenmesi

Investigation of Classification Conditions For Multicategorical Classification in Computerized Adaptive Classification Test

Dr. Demet ALKAN
Dr. ◆ Milli Eğitim Bakanlığı, Eğitimde Ölçme ve Değerlendirme ◆ alkandemet@hotmail.com ◆ ORCID: 0000-0002-1478-9183

Prof. Dr. Nuri DOĞAN
Prof. Dr. ◆ Hacettepe Üniversitesi, Eğitimde Ölçme ve Değerlendirme ◆ nuridogan2004@gmail.com ◆ ORCID: 0000-0001-6274-2016

Özet

Bu çalışmada R programlama dili ile çok kategorili sınıflama için Bilgisayarda Bireyselleştirilmiş Sınıflama Testi (BBST) kullanıldığında test etkililiğinin ve ölçme kesinliğinin sınıflama kriterleri, madde seçme yöntemleri, yetenek kestirim yöntemleri ve iki, üç, dört kategorili sınıflama kategori sayısı ile nasıl değiştiği araştırılmıştır. Simülasyonla iki kategorili, tek boyutlu 500 madde ve 1000 kişilik veri ile. 36 koşul belirlenmiştir. Tüm koşullar için 25 tekrarın ortalaması alınmıştır. Araştırma sonunda sınıflama kategori sayısı arttıkça Ortalama Test Uzunluğunun (OTU) arttığı, Ortalama Sınıflama Doğruluğu (OSD) azaldığı görülmüştür. Ortalama Hatanın Karekökü (RMSE), Ortalama Mutlak Hata (OMH), Yanlılık ve Gerçek Yetenekler ile Kestirilen Yetenekler Arasındaki Korelasyon (r) değerlerinin azaldığı anlaşılmıştır. OTU için Güven Aralığı (GA) sınıflama kriteri OSD, yanlılık, korelasyon, OMH için Ardışık Olasılık Oran Testi (AOOT) sınıflama kriterinin performansının daha etkili olduğu görülmüştür. Genelleştirilmiş Olabilirlik Oran (GOO) sınıflama kriterinin OTU bakımından GA kriterine benzer sonuçlar, mutlak hata yönünden ise AOOT sınıflama kriteri ile benzer sonuçlar oluşturduğu görülmüştür. Yetenek kestirim yöntemleri OSD ve OTU açısından benzer performans göstermiştir. Kesme Noktası (KN) temelli madde seçme yöntemleri Kestirilen Yetenek (KY) temelli madde seçme yöntemlerine göre test etkililiği ve ölçme kesinliği açısından daha etkili performans gösterdiği belirlenmiştir.

Anahtar Kelimeler: Bilgisayarda bireyselleştirilmiş sınıflama testi, Madde seçme yöntemi, Ölçme kesinliği, Sınıflama kategori sayısı, Sınıflama kriteri, Test etkililiği

Abstract

This study used the Computerized Adaptıve Classification Test (CACT) for multi-category classification with R programming language to investigate how test effectiveness and measurement accuracy changed in terms of classification criteria, item selection methods, ability estimation methods, and two, three, and four-category classifications. With the simulation, two-category, one-dimensional 500 items and 1000-person data were created, 36 conditions were determined, and 25 repetitions were averaged for all conditions. Results showed that as the number of classification categories increased, the Average Test Length (ATL) increased and the Average Classification Accuracy (ACA) decreased. The Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Bias, and Correlation (r) values between real and estimated thetas (r) values were found to decrease. The performance of the Confidence Interval (CI) classification criterion for ATL, ACA, bias, correlation, and the Sequential Probability Ratio Test (SPRT) classification criterion for MAE were found to be more effective. Generalized Likelihood Ratio (GLR) classification criterion produced similarresults to the CI criterion in terms of ATL, and to the SPRT classification criterion in terms of absolute error. Ability estimation methods were similar in terms of ACA and ATL. Cutscore based (CB) item selection methods were more effective in terms of test effectiveness and measurement accuracy than Estimated Ability -Based (EB) item selection methods.

Kabul Tarihi / Received: 09.09.2023 | Onay Tarihi / Accepted: 25.03.2024

Keywords: Computerized adaptıve classification test, Classification criteria, Measurement accuracy, Number of classification categories, Item selection method, Test efficiency

1. Giriş

Doğrusal testler, bireylerin başarısını ve yeteneğini belli bir noktada belirlemek için ölçmede en çok kullanılan testlerdir. Bilgisayarda Bireyselleştirilmiş Testler (BBT) öğrencinin yeteneğini daha kısa zamanda daha az madde ile bilgisayar ortamında etkili şekilde değerlendirilmesini sağlayan testlerdir. Bireylerin belli bir özelliğe göre sınıflanması ve daha az sayıda madde ile daha doğru sınıflama yapılması amaçlandığında ise Bilgisayarda Bireyselleştirilmiş Sınıflama Testi (BBST) kullanılabilir. BBST bireyleri özel kategorilere göre bilgisayar ortamında sınıflayan test uygulamalarıdır. BBST psikometrik model, başlama noktası, madde seçme yöntemi, yetenek kestirim yöntemi, sınıflama kriterleri olmak üzere beş bileşenden oluşmaktadır. Psikometrik model olarak klasik test kuramı (KTK) ve madde tepki kuramı (MTK) kullanılabilmektedir. BBST uygulamalarında teste başlama noktası için bireylerin ön bilgileri kullanılabilir (Weiss & Kingsbury, 1984). Madde seçiminde bilgisayarın madde havuzundaki tüm maddeleri değerlendirdiği, en iyi maddenin seçildiği kestirilen yetenek temelli ve kesme noktası temelli akıllı madde seçim yöntemleri kullanılmaktadır. MTK psikometrik modeli ile Maksimum Fişher Bilgisi (MFB) ve Kullback Liebler Bilgisi (KLB) kullanılabilecek madde seçme yöntemleridir (Lin & Spray, 2000). Yetenek kestirim yöntemi olarak Maksimum Olabilirlik Kestirimi (MOK), Ağırlıklı Olabilirlik Kestirimi (AOK), Beklenen Sonsal Dağılım (BSD), Owen'ın Bayesçi Yetenek Kestirimi gibi yöntemler kullanılmaktadır. Sınıflama kriterleri olarak Ardışık Olasılık Oran Testi (AOOT), Genelleştirilmiş Olabilirlik Oran (GOO), Bireyselleştirilmiş Uzmanlık Testi (BUT), Güven Aralığı yöntemi (GA), Bayesçi Karar Kuramı (BKK) yöntemleri kullanılabilmektedir. Tüm sınıflama kriterlerinin geleneksel testlerden daha az madde ile sınıflama yaptıkları bilinmektedir (Kingsbury, & Weiss, 1983).

Sınıflamanın yanlış yapılması öğrencilerin hak ettikleri kariyer ya da eğitimden geri kalmalarına sebep olabilir. Çok önemli olmasına rağmen standart belirleme ile ilgili hatalar, kesme puanının belirlenmesi gibi hatalar yanlış sınıflamaya sebep olabilir (Arce-Ferrer ve diğerleri, 2002). Özellikle kesme noktası civarında puan alan öğrencilerin yanlış sınıflanması olasıdır (Eckes, 2017). Sınıflama Doğruluğuna (classification accuracy; CA) kanıt sağlamak test geliştiricilerin görevidir. Bilgisayarda Bireyselleştirilmiş Sınıflama Testi özellikle sınava girenlerin sınıflanması için tasarlanmış değişken uzunluklu bir testtir. Çok kategorili sınıflama için sabit test formu yerine Bilgisayarda Bireyselleştirilmiş Test kullanmak en yüksek özelliklere sahip maddeleri seçerek daha doğru sınıflamaya ulaşılabileceğini göstermiştir (Lewis & Sheehan, 1990). Çok kategorili sınıflama iki kategorili sınıflamaya göre daha geniş bir yetenek yelpazesinde daha fazla bilgi gerektirdiği için sınava girenler tek bir puan yerine iki veya daha fazla kesme puanına göre iki, üç, dört ve daha fazla kategoride sınıflandırılır. Bir testin iki ya da üç kesme puanıyla üç ya da dört kategoride sınıflanması, sınava giren tüm öğrencilerde ihtiyaç duyulan ortalama madde sayısını artırır ve bu da maddelere aşırı maruz kalmadan etkili bir test uygulaması için bir madde havuzunda ihtiyaç duyulan madde sayısını artırır (Spray, 1993). Çoklu kesme puanları için sınıflama testlerinin geliştirilmesi, iki sınıflı testlerinkine benzerdir. Örneğin, içerik alanlarının oluşturulması ya da maddelerin yazılması kaç tane kesme puanı olduğundan etkilenmez. Ancak bu iki durum, testi uygulamak ve puanlamak için kullanılan belirli psikometrik yöntemler açısından farklılık gösterir. Ardışık Olasılık Oranı Testi gibi yöntemler genel olarak her iki duruma da uygulanabilirken, belirli özelliklerin çoklu kesme puanı durumuna uyarlanması gerekir. Bunun en doğru olduğu sınıflama testinin yönü, Bilgisayarlı Sınıflamada kullanılan algoritmalardır (Spray, 1993). BBST ile ilgili yurt içinde Demir, 2019; ve Gündeğer 2017 olmak üzere az sayıda çalışma bulunmaktadır.

Demir (2019) simülasyon verisi ile çok kategorili sınıflamada madde kullanım sıklığı ve içerik dengeleme gibi pratik kısıtlamalar altında test etkililiğinin ve ölçme kesinliğinin nasıl değiştiğini araştırmıştır. Gündeğer (2017) ise iki kategorili sınıflama için test etkililiği ve ölçme kesinliğini araştırmıştır. Yurt dışındaki literatürde ise genelde sınıflama kriterleri arasında yer alan AOOT ve BUT sınıflama kriterlerinin karşılaştırıldığı çalışmalar (Kingsburry & Weiss,1980; Thompson, 2011; Nydick, 2013), sınıflamanın iki kategoride yapıldığı çalışmalar (Lau,1996; Spray & Reckase,1996; Reckase,1983), madde seçme yöntemlerinin karşılaştırıldığı çalışmalar ve genelde simülatif veri ile yapılan çalışmalar (Eggen, 1999; Lin & Spray, 2000) bulunmaktadır.

BBST uygulaması tur sistemiyle işleyen bir uygulamadır. Bir maddenin veya madde grubunun (test takımı) her turun başında seçilmesi ve testi cevaplayan kişinin maddelere yanıt vermesi ve bilgisayarın, bireyin sınıflandırılıp sınıflandırılamayacağını değerlendirmek için yanıtları kullanmasıdır. Sınıflama kriterleri bu değerlendirme için nicel temeli sağlar. Sınava giren kişi sınıflandırıldığında, test sonlandırılır. Sınıflama kriteri karar veremiyorsa süreç başka bir turla kendini tekrar eder (Spray & Reckase, 1994). Sınıflama kriterlerinde kullanılan farksızlık bölgesi (FB) ve güven aralığı tolere edilebilen hata düzeyini gösterir. Farksızlık bölgesi küçük olursa ya da güven düzeyi büyük olursa sınıflama doğruluğu yüksek olur. (Reckase, 1983). Araştırmada Nydick (2013) çalışması dikkate alınarak farksızlık bölgesi 0,1 güven aralığı %90 olarak belirlenmiştir. Yetenek kestirim yöntemleri Ağırlıklandırılmış Olabilirlik Kestirimi (AOK), Beklenen Sonsal Dağılım (BSD), Bayes Yetenek Kestirimi (BYK), Madde seçme yöntemi olarak Kestirilen Yetenek ve Kesme Noktası Temelli Maksimum Fisher Bilgisi (MFB) ve Kulback Laibler Bilgisi (KLB), Sınıflama kriteri olarak Ardışık Olasılık Oran Testi (AOOT), Güven Aralığı (GA) ve Genelleştirilmiş Olabilirlik Oran (GOO) sınıflama kriterleri kullanılmıştır. Bilgisayarda Bireyselleştirilmiş Sınıflama testlerinde madde sayısının az olması ve ortalama sınıflama doğruluğunun (OSD) yüksek olmasıyla testin etkililiği artarken düşük standart hatalar, gerçek ve kestirilen yetenek düzeyleri arasındaki korelasyonun yüksek olması ölçme kesinliğini yükseltir (Thompson, 2009). BBST araştırmasının genel amacı az madde kullanarak yüksek sınıflama doğruluğu hedeflenerek test etkililiğini oluşturmaktır.

1.1 Araştırmanın Amacı ve Önemi

Araştırmanın amacı, BBST simülasyonu ile yapılan çok kategorili sınıflamada sınıflama kategori sayısına göre BBST' nin verimliliğini en üst düzeye çıkaran belirli yöntemleri belirlemek ve sınıflama doğruluğu hakkında kanıt sağlamaktır. Alan yazında yaygın olarak iki kategorili sınıflama yapıldığı görülmektedir (Reckase, 1983; Lau, 1996). Reckase (1983) 1 PLM ve 3 PLM olmak üzere iki MTK modelinde AOOT sınıflama kriterinin performansını üç farksızlık bölgesi ile ortalama test uzunluğu ortalama sınıflama doğruluğu açısından iki kategorili sınıflama için araştırmıştır. Lau (1996) veri setinin tek boyutlu olmadığı durumlarda AOOT sınıflama kriterinin iki kategorili sınıflamada performansını araştırmıştır. Çok kategorili sınıflama kriterlerinin performansının incelendiği araştırma örneklerine az rastlanmaktadır. Yurt içinde Demir (2019), çok kategorili sınıflamada madde kullanım sıklığı ve içerik dengelemenin etkisini incelemiştir. Gündeğer (2017), İki kategorili sınıflama için araştırma yapmıştır. Yurt içinde çok kategorili farklı bir araştırmaya rastlanmamıştır. Araştırmanın uygulayıcılara, çok kategorili sınıflama için en uygun sınıflama kategori sayısı, sınıflama kriterleri, madde seçme yöntemleri ve yetenek kestirim yöntemleri ile ilgili bilgi vermesi 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 sınıflamada test etkililiği açısından ortalama sınıflama doğruluğu (OSD), ortalama test uzunluğu (OTU) , ölçme kesinliği açısından gerçek yetenekler ile kestirilen yetenekler arasındaki korelasyon (r), yanlılık, Ortalama Hatanın Karekökü (RMSE),

Ortalama Mutlak Hata (OMH) değerlerinin, madde seçme yöntemleri, yetenek kestirim yöntemleri, sınıflama kriterlerine ve sınıflama kategori sayısına göre nasıl değiştiği araştırılmıştır.

Araştırmanın alt problemleri aşağıdaki gibidir.

1. BBST simülasyonu ile yapılan iki, üç ve dört kategorili sınıflamada AOK yetenek kestirim yöntemi ile yetenek kestirildiğinde, sınıflama kriterlerinin madde seçme yöntemleri ile çaprazlandığı koşullara ait ortalama sınıflama doğruluğu, ortalama test uzunluğu, korelasyon (r), yanlılık, Ortalama Hatanın Karekökü (RMSE), Ortalama Mutlak Hata (OHM) değerleri nasıl değişmektedir?

2. BBST simülasyonu ile yapılan iki, üç ve dört kategorili sınıflamada BSD yetenek kestirim yöntemi ile yetenek kestirildiğinde, madde seçme yöntemlerinin sınıflama kriterleriyle çaprazlandığı koşullara ait ortalama sınıflama doğruluğu, ortalama test uzunluğu, korelasyon (r), yanlılık, Ortalama Hatanın Karekökü (RMSE), Ortalama Mutlak Hata(OHM) değerleri nasıl değişmektedir?

3. Madde seçme yöntemlerinin, yetenek kestirim yöntemlerinin ve sınıflama kriterlerinin iki, üç ve dört kategorili sınıflamada ortalama sınıflama doğruluğu, ortalama test uzunluğu, korelasyon (r), yanlılık, Ortalama Hatanın Karekökü (RMSE), Ortalama Mutlak Hata(OHM) değerleri ölçme kesinliği ve test etkililiği açısından nasıl değişmektedir?

2.Yöntem

2.1. Araştırmanın Türü

Araştırma betimsel araştırma ve monte carlo simülasyon çalışmasıdır. Betimsel araştırmalar bir durumu tanımlayan açıklayan araştırmalardır (Kaptan, 1995). Araştırmada yetenek parametreleri ve madde havuzu parametreleri R ortamında oluşturulduğu için araştırma monte carlo simülasyon çalışmasıdır. (R Core Team, 2013). Araştırmada üç sınıflama kriteri, dört madde seçme yöntemi, üç sınıflama kategori sayısı ile (3x4x3) 36 koşul oluşturulmuştur.

2.2 Veri Üretimi

Araştırmada Thompson (2009, 2011) çalışmaları dikkate alınarak üç parametreli lojistik model ile 500 maddelik madde havuzu oluşturulmuştur. Araştırmada Kesme Noktası ve Kestirilen Yetenek Temelli Maksimum Fisher Bilgisi ve Kullback Leibler madde seçme yöntemleri kullanılmıştır. İki, üç ve dört kategoride sınıflamalar (-3,3) yetenek düzeyleri aralığında madde parametreleri kullanılarak yapılmıştır. Maddelerin parametreleri Weiss (1980) çalışmasında olduğu gibi a parametresi U(0,5-1,5) dağılımdan, b parametresi N (0-1) dağılımdan türetilmiştir. Thompson (2009) çalışması dikkate alınarak c parametresi N (0-0,3) olarak normal dağılımdan 500 madde türetilmiştir. Yetenek parametresi ortalaması 0, standart sapması 1 olacak şekilde 1000 bireyin her biri için 25 tekrarla R ortamında türetilmiştir (R Core Team, 2013).

2.3 Simülasyon Koşulları

BBST simülasyon koşulları için üç sınıflama kriteri dört madde seçme yöntemi ve üç yetenek kestirim yöntemi belirlenmiştir. 36 koşul için R ortamında 25 tekrarla analizler yapılmıştır (R Core Team, 2013). Sınıflama kriterleri için 0,1 farksızlık bölgesi ile Ardışık Olasılık Oran Testi (AOOT) ve Genelleştirilmiş Olabilirlik Oran (GOO), %90 güven aralığı ile Güven Aralığı (GA) yöntemleri Nydick (2013) ve Eggen ve Straetmans'in (2000) araştırmalarına göre belirlenmiştir. Farksızlık bölgesi kesme puanına yakın sınıflama kararları için tolere edilebilen belirsizlik düzeyidir. Daha küçük farksızlık bölgesi sınıflama doğruluğunu daha iyi koruyabilir fakat testin uzunluğunu artırır (Reckase, 1983). Araştırmada Nydick (2013) çalışması dikkate alınarak farksızlık bölgesi 0,1 olarak belirlenmiştir. Yetenek kestirim

yöntemleri olarak Beklenen Sonsal Dağılım (BSD), Ağırlıklandırılmış Olabilirlik Kestirimi (AOK), Bayesçi Yetenek Kestirimi (BYK) kullanılmıştır. Warm (1989) a göre Ağırlıklandırılmış Olabilirlik Kestirimi (AOK) yanlılığı azaltan, yetenek düzeyinde ağırlıklandırma olabilirliği üzerine çalışan BBST çalışmalarında çoğunlukla tercih edilen yöntemdir. Bayesçi yöntemler araştırmalarda fazla çalışılmadığı için yetenek kestirim yöntemleri olarak incelenmiştir. Madde seçme yöntemleri kestirilen yetenek temelli ve kesme noktası temelli Maksimum Fişher Bilgisi (MFB) ve Kullback Laibler Bilgisi (KLB) olarak belirlenmiştir. Thompson'a (2007) göre yetenek düzeyi sıfır alınabilir ya da önceden belirlenen yetenek düzeyleri başlama noktası olarak kullanılabilir. Tüm koşullar için başlama noktası sıfır alınmıştır. Koşullar araştırma problemlerine göre manipüle edilip 25 tekrarın ortalaması alınarak iki, üç ve dört kategorili sınıflama yapılmıştır. İki, üç ve dört kategorili sınıflama için yetenek parametrelerinden faydalanarak kesme noktaları belirlenmiştir. Eggen ve Straetmans'ın (2000) çalışmalarında olduğu gibi yetenek düzeyleri ikiye ayrılıp ilk bölüm 1.düzey diğer bölüm 2.düzey olarak her düzeyin %70'i alınarak kesme noktası belirlenmiştir. Araştırmada Catırt (Nydick, 2014) paketi kullanılmıştır.

2.4 Verilerin Analizi

Araştırma problemlerine bağlı olarak oluşturulan 36 simülasyon koşullarında 25 tekrar yapılmış ve tüm değerler tekrarların ortalaması alınarak R ortamında Nydick tarafından (2014) yazılan "catIrt" paketindeki yazılan fonksiyonlarla hesaplanmıştır. RMSE, OMH, yanlılık, yetenek düzeyleri arasındaki korelasyon (r), ortalama sınıflama doğruluğu (OSD), ortalama test uzunluğu (OTU) değerleri araştırılmıştır. 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 ($\theta\hat{\imath}$) gerçek yetenek düzeylerinden (θi) farkları toplamının birey sayısına (n) oranına eşittir (Miller & Miller, 2004).

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

RMSE, kestirilen son yetenek düzeylerinin ($\theta\hat{\imath}$) gerçek yetenek düzeylerinden (θi) farklarının kareleri toplamının birey sayısına (n) oranının karekköüne eşittir

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

OMH, kestirilen son yetenek düzeylerinin ($\theta\hat{\imath}$) gerçek yetenek düzeylerinden (θi) farklarının mutlak değerleri toplamının birey sayısına (n) oranına eşittir.

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

3. Bulgular

3.1 Birinci Araştırma Sorusuna Ait Bulgular

Araştırmanın birinci alt probleminde iki, üç ve dört kategorili sınıflamada yetenek BSD ile kestirildiğinde sınıflama kriterlerinin madde seçme yöntemleri ile çaprazlandığı koşulların ölçme kesinliği ve test etkililiği açısından nasıl değiştikleri incelenmiştir. Ölçme kesinliği için r, yanlılık, RMSE ve OMH değerleri, test etkililiği için OTU ve OSD değerleri araştırılmıştır. Tüm koşullar için 25 tekrarın ortalaması alınarak elde edilen değerler Tablo 1'de gösterilmiştir.

Tablo 1'de Kestirilen Yetenek Temelli ve Kesme Noktası Temelli madde seçme yöntemleri kullanılarak bireyleri sınıflamak için en az madde gerektiren sınıflama kriterinin iki kategorili sınıflama ile 12,02 ve 12,58 değerlerinde GA (%90) sınıflama kriteri olduğu görülmüştür. İki kategorili sınıflamada OSD için 0,897 değeri ile AOOT (FB=0,1) sınıflama kriteri en iyi performansı göstermiştir. Hata değerleri, yanlılık ve gerçek yetenekler ile kestirilen yetenekler arasındaki korelasyon için AOOT (FB=0,1) sınıflama kriterinin en az hata ile sınıflama yaptığı görülmüştür. Araştırmanın bulgusu iki kategorili sınıflama için Nydick ve diğerleri (2012) araştırma bulguları ile benzerlik göstermektedir. Nydick ve diğerleri (2012) genel olarak, GA ile yapılan sınıflamaların AOOT ile yapılan sınıflamalara kıyasla daha düşük OTU ve OSD ile sonlandığını ifade etmişlerdir.

Sınıflama kategori sayısı arttıkça testi sonlandırmak için gereken madde sayısı artmıştır. Üç kategorili sınıflamada OTU için GA (%90) sınıflama kriteri KLB-KY madde seçme yöntemi birlikte kullanıldığında en az madde sayısı ile (18,03) sınıflama yapılan koşul olduğu anlaşılmıştır. OSD için üç kategorili sınıflamada AOOT (FB= 0,1) sınıflama kriteri 0,896 ile en iyi performansı göstermiştir. Yanlılık, RMSE, OMH ve kestirilen yetenekler ile gerçek yetenekler arasındaki korelasyon açısından üç kategorili sınıflamada AOOT (FB= 0,1) sınıflama kriteri düşük hata ve yanlılık değerleri ve yüksek korelasyonla en iyi performansı göstermiştir. GOO sınıflama kriteri ile yapılan sınıflamada OTU açısından GA (%90), OSD açısından AOOT (FB= 0,1) sınıflama kriterine benzer bulgular elde edilmiştir. Dört kategorili sınıflama için GA sınıflama kriteri OTU olarak en az madde ile AOOT (FB= 0,1) sınıflama kriteri ise OSD olarak sınıflama için en iyi performansı göstermiştir.

Kesme noktası ve kestirilen yetenek temelli madde seçme yöntemlerinin sınıflama kriterleri ve sınıflama kategori sayısı ile çaprazlandığı koşullarda MFB-KY madde seçme yöntemi AOOT (FB= 0,1) sınıflama kriteri ile birlikte kullanıldığında iki, üç ve dört kategorili sınıflamada sınıflama için gereken madde sayısı 32,94-46,61 aralığındadır. GA (%90) sınıflama kriteri kullanıldığında sınıflama için gereken madde sayısı 12,02-28,12 aralığındadır. GOO (FB= 0,1) sınıflama kriteri ile birlikte 14,57-33,22 aralığında madde sayısı ile sınıflama yapıldığı görülmektedir. Sınıflama kategori sayısı arttıkça sınıflama için gereken madde sayısı tüm kriterler için artmaktadır. OTU için MFB-KY madde seçme yöntemi GA (%90) sınıflama kriteri ile uygulandığında en az madde ile en etkili yöntem olduğu anlaşılmaktadır.

KLB-KY madde seçme yönteminin sınıflama kriterleri ile ve sınıflama kategori sayısıyla yapılan çaprazlanmasında OTU için iki, üç ve dört kategorili sınıflamada, AOOT (FB= 0,1) sınıflama kriteri ile birlikte kullanıldığında 33,21-46,64 aralığında madde sayısı ile, GOO (FB= 0,1) sınıflama kriteri ile birlikte 14,63- 33,21 aralığında madde sayısı ile, GA (%90) sınıflama kriteri ile birlikte 12,02-28,90 aralığında madde sayısı ile sınıflama yapıldığı anlaşılmaktadır. En az madde ile testi sonlandıran yöntem GA (%90) yöntemidir. GA sınıflama kriteri her maddeden sonra belirlenen yetenek düzeylerini belirlenen güven aralığını kesme puanı ile karşılaştırdığı için KY temelli madde seçme yöntemleri ile performansının daha yüksek olduğu yorumu yapılabilir.

Tablo 1. *Yeteneğin BSD ile Kestirildiği Çok Kategorili Sınıflamada Koşullara Ait OTU, OSD, r, Yanlılık, RMSE, OMH Değerlerinin Ölçme Kesinliği ve Test Etkililiği Açısından Karşılaştırılması*

Koşullar			*Bağımlı Değişkenler*					
Madde Seçme Yöntemi	*Sınıflama Kriteri*	*SKS*	*OTU*	*OSD*	*r*	*Yanlılık*	*RMSE*	*OMH*
MFB-KY	AOOT(FB= 0,1)	İki	32,94	0,897	0,984	-0,001	0,182	0,14
		Üç	38,42	0,896	0,985	0	0,18	0,136
		Dört	46.61	0,885	0,987	0,001	0,168	0,126
	GA(%90)	İki	12,02	0,879	0,928	-0,002	0,387	0,292
		Üç	22,0	0,885	0,954	0,001	0,309	0,221
		Dört	28,12	0,875	0,97	0	0,252	0,179
	GOO(FB= 0,1)	İki	14,57	0,894	0,936	-0,001	0,364	0,27
		Üç	24,98	0,892	0,957	-0,001	0,201	0,211
		Dört	33,22	0,881	0,972	-0,001	0,244	0,17
MFB-KN	AOOT(FB= 0,1)	İki	24,98	0,892	0,883	0,01	0,485	0,342
		Üç	36,11	0,896	0,948	0,003	0,33	0,22
		Dört	47,88	0,885	0,982	-0,001	0,198	0,14
	GA(%90)	İki	12,58	0,886	0,835	0,01	0,57	0,42
		Üç	22,347	0,888	0,923	0,001	0,399	0,278
		Dört	29,09	0,875	0,964	-0,001	0,276	0,196
	GOO(FB= 0,1)	Iki	13,97	0,889	0,834	0,007	0,571	0,418
		Üc	24,64	0,892	0,925	0,002	0,393	0,269
		Dört	33,27	0,88	0,967	0	0,267	0,184
KLB-KY	AOOT (FB= 0,1)	İki	33,21	0,896	0,985	-0,001	0,18	0,139
		Üç	38,53	0,896	0,985	0,001	0,179	0,136
		Dört	46,64	0,885	0,987	0,001	0,168	0,126
	GOO (FB= 0,1)	İki	14,63	0,89	0,94	-0,001	0,355	0,264
		Üç	24,89	0,891	0,958	0	0,299	0,209
		Dört	33,21	0,881	0,972	-0,001	0,244	0,169
	GA(%90)	İki	12,02	0,879	0,928	-0,002	0,387	0,292
		Üç	18,03	0,868	0,944	-0,01	0,341	0,252
		Dört	28,90	0,875	0,97	0,001	0,253	0,18
KLB-KN	AOOT (FB: 0,1)	İki	24,56	0,892	0,884	0,009	0,484	0,341
		Üç	36,07	0,895	0,948	0,004	0,33	0,22
		Dört	48,15	0,874	0,981	-0,004	0,202	0,141
	GOO (FB= 0,1)	İki	13,98	0,888	0,84	0,009	0,562	0,411
		Üç	24,65	0,891	0,927	0,003	0,389	0,266
		Dört	33,40	0,871	0,965	0	0,274	0,185
	GA(%90)	İki	12,59	0,885	0,837	0,007	0,566	0,417
		Üç	22,36	0,888	0,925	0,002	0,393	0,273
		Dört	29,97	0,863	0,966	-0,008	0,271	0,191

SKS: Sınıflama kategori sayısı, OTU: Ortalama test uzunluğu, OSD: Ortalama sınıflama doğruluğu, OMH: Ortalama mutlak hata, r: Gerçek yeteneklerle kestirilen yetenekler arası korelasyon

Kestirilen Yetenek Temelli madde seçme yöntemlerinden KLB yöntemi ile MFB yönteminin OTU için yakın sonuçlar oluşturduğu görülmektedir. GA yöntemi kestirilen yetenek temelli madde seçme yöntemlerinin tamamıyla en az madde ile iki, üç ve dört kategorili tüm sınıflamalarda en etkili yöntemdir. Demir (2019) araştırmasında OTU açısından GA yönteminin, OSD için AOOT sınıflama kriterinin performansının başarılı olduğu bulgusuna ulaşmıştır. Thompson (2009) yaptığı çalışmada, GA sınıflama kriteri için KY temelindeki madde seçiminin, AOOT sınıflama kriteri için ise KN temelindeki madde seçiminin daha uygun olduğu sonucuna ulaşmıştır. Araştırmanın bu bulgusu Demir (2019) ve Thompson (2009) araştırma bulgularıyla nispeten uyumludur.

Kesme Noktası Temelli KLB ve MFB madde seçme yöntemlerinde iki, üç ve dört kategorili sınıflamalarda KLB için OTU açısından GA (%90) sınıflama kriteri 12,59-29,97 aralığında madde sayısı ile, GOO (FB= 0,1) sınıflama kriteri 13,98 -33,407 aralığında madde sayısı ile, AOOT (FB= 0,1) sınıflama kriterinin ise 24,56-48,153 aralığında madde sayısı ile sınıflama yaptığı görülmektedir. GA sınıflama kriteri en az madde ile sınıflama yaptığı için test etkililiği olarak en uygun sınıflama kriteridir. Sınıflama kategori sayısı arttıkça testi sonlandırmak için kullanılan madde sayısı artmıştır. Kesme noktası temelli MFB madde seçme yöntemi OTU için GA (%90) sınıflama kriteri ile 12,58-29,09 aralığında, GOO (FB= 0,1) sınıflama kriteri ile 13,97-33,27 aralığında, AOOT (FB= 0,1) sınıflama kriteri ile ise 24,98-47,88 aralığında madde sayısı ile sınıflama yaptığı tablodan anlaşılmaktadır. GA sınıflama kriterinin kesme noktası temelli madde seçme yöntemleriyle en az madde ile sınıflama yaptığı görülmektedir. Kesme noktası temelli madde seçme yöntemi sınıflama kriteri ve sınıflama kategori sayısı ile çaprazlandığında MFB yöntemi KLB ile OTU olarak benzer sonuçlar gösterse de MFB madde seçme yönteminin daha az madde ile sınıflama yaptığı görülmektedir.

OSD değeri iki, üç ve dört kategorili sınıflamalarda 0,863 ile 0,896 aralığındadır. AOOT sınıflama kriteri KY ve KN temelli madde seçme yöntemleri ve sınıflama kategori sayısı ile çaprazlandığında OSD için 0,871-0,897 aralığında değerler hesaplanmıştır. Sınıflama kategori sayısı arttıkça OSD değeri düşmüştür. GA sınıflama kriteri ile sınıflama yapıldığında OSD için 0,868-0,888 aralığında değerler hesaplanmıştır. AOOT sınıflama kriterine göre OSD nispeten daha düşüktür. GOO sınıflama kriteri kullanıldığında ise OSD için 0,892-0,871 aralığında değerler hesaplanmıştır. İki, üç ve dört kategorili sınıflamada kesme noktası ve kestirilen yetenek temelli MFB ve KLB madde seçme yöntemleri sınıflama kriterleri ve sınıflama kategori sayısı ile çaprazlandığında OSD için GOO ve AOOT yöntemlerinin sınıflamanın doğruluğu için benzer değerler oluşturduğu görülmektedir. GA sınıflama kriteri ise nispeten daha düşük değerler oluşturmuştur. OTU açısından GA yöntemi ile daha iyi sonuçlar elde edilmiştir. Sınıflama kategori sayısındaki artış OTU' nu artırırken OSD' nu azaltmıştır. GOO sınıflama kriterinin çok kategorili sınıflama için literatürde performansını destekleyecek araştırma örneklerine rastlanmamıştır. GOO sınıflama kriterinin uygulamada OTU ve OSD için farklı bir ifade ile ölçme kesinliği ve test etkililiği için uygulayıcılara avantaj sağlayacağı yorumu yapılabilir.

Simülasyon öncesi oluşturulan gerçek yetenek düzeyleri ile simülasyon sonrası kestirilen yetenek düzeyleri arasındaki korelasyon (r), MFB ve KLB kestirilen yetenek ve kesme noktası temelli madde seçme yöntemlerinin sınıflama yöntemleri ve sınıflama kategori sayısı ile çaprazlandığı koşullarda MFB-KY ve KLB-KY için 0,928-0,987 aralığında yüksek değerler hesaplanmıştır. Sınıflama kategori sayısındaki artışa göre de korelasyon değeri artmıştır. MFB-KN ve KLB-KN için korelasyon değeri (r) 0,835-0,98 aralığında değişmektedir. KY temelli madde seçme yönteminin KN temelli madde seçme yöntemine göre kestirilen ve gerçek yetenek düzeyleri arasındaki korelasyon değerlerinde daha etkili performans gösterdiği söylenebilir.

Yanlılık, kestirimin standart hatasını gösteren RMSE ve OMH için MFB-KY ve MFB-KN madde seçme yöntemlerinin AOOT sınıflama kriteri ile KY temelli MFB yöntemi kullanıldığında daha düşük

yanlılık ve kestirimin standart hatası değerleri elde edilmiştir. Sınıflama kategori sayısı arttıkça RMSE ve OMH değerleri düşmüştür, yanlılık fazla etkilenmemiştir. GA ve GOO sınıflama kriterleri için de benzer durum olduğu söylenebilir. Ölçme kesinliği için kestirimin standart hatası ve yanlılığın düşük, kestirilen yetenekler ile gerçek yetenekler arasındaki korelasyonun yüksek olması beklenir. Test etkililiği için az madde ile sınıflama doğruluğunun yüksek olması beklenilir. Bu durumda AOOT sınıflama kriterinin ölçme kesinliği için, GA sınıflama kriterinin ise test etkililiği için etkili sınıflama kriterleri olduğu yorumu yapılabilir.

3.2 İkinci Araştırma Sorusuna Ait Bulgular

Araştırmanın ikinci alt probleminde iki, üç ve dört kategorili sınıflamada yetenek AOK ile kestirildiğinde sınıflama kriterlerinin madde seçme yöntemleri ile çaprazlandığı koşulların ölçme kesinliği ve test etkililiği açısından nasıl değiştikleri Tablo 2 de gösterilmiştir.

Araştırmanın ikinci sorusunda yetenek AOK ile kestirildiğinde madde seçme yöntemleri ve sınıflama kriterlerinin çaprazlandığı koşullarda elde edilen değerlere göre oluşturulan Tablo 2 de Kestirilen Yetenek Temelli ve Kesme Noktası Temelli madde seçme yöntemlerinin ikisi için de testi sonlandırmak bireyleri sınıflamak için en az madde gerektiren sınıflama kriterinin iki kategorili sınıflama ile 11,29 ve 14,68 değer aralıklarında GA (%90) sınıflama kriterinin olduğu görülmüştür. İki kategorili sınıflamada OSD için 0,899 değeri ile AOOT (F= 0,1) sınıflama kriterinin MFB-KY temelli madde seçme yöntemi ile oluşturulan koşulda en iyi performansı gösterdiği anlaşılmaktadır. Hata değerleri, yanlılık ve gerçek yetenekler ile kestirilen yetenekler arasındaki korelasyon için AOOT (FB= 0,1) sınıflama kriterinin en az hata ile sınıflama yaptığı görülmüştür.

Sınıflama kategori sayısı arttıkça testi sonlandırmak için gereken madde sayısı artmıştır. Üç kategorili sınıflamada OTU için GA (%90) sınıflama kriteri 20,16 değeri ile KLB-KY ve MFB-KY madde seçme yöntemleri ile birlikte kullanıldığında en az madde ile sınıflama yapılan koşullar olduğu anlaşılmıştır. OSD için üç kategorili sınıflamada AOOT (FB= 0,1) sınıflama kriteri 0,897 ile en iyi performansı göstermiştir. Yanlılık, RMSE, OMH ve kestirilen yetenekler ile gerçek yetenekler arasındaki korelasyon için üç kategorili sınıflamada AOOT (FB= 0,1) sınıflama kriterinin KY temelli madde seçme yöntemleri ile en yüksek korelasyon değerleri oluşturduğu görülmektedir.

Yanlılık ve hata değerleri olarak AOOT sınıflama kriterinin tüm madde seçme yöntemleri ile en iyi performansı gösterdiği görülmektedir. GOO sınıflama kriteri ile yapılan sınıflamada OTU açısından GA (%90) sınıflama kriterine, OSD açısından AOOT (FB= 0,1) sınıflama kriterine benzer bulgular elde edilmiştir. Dört kategorili sınıflama için GA sınıflama kriteri OTU olarak en az madde ile en iyi performansı göstermiştir. AOOT (FB= 0,1) sınıflama kriteri OSD olarak sınıflama için dört kategorili sınıflamada en iyi performansı göstermiştir.

Kesme noktası ve kestirilen yetenek temelli madde seçme yöntemlerinin sınıflama kriterleri ve sınıflama kategori sayısı ile çaprazlandığı koşullarda MFB-KY madde seçme yöntemi AOOT (FB= 0,1) sınıflama kriteri ile birlikte kullanıldığında iki, üç ve dört kategorili sınıflamada sınıflama için gereken madde sayısı 33,35-46,81 aralığındadır. MFB-KY madde seçme yöntemi ile GA (%90) sınıflama kriteri kullanıldığında 11,35-28,04 aralığında madde ile sınıflama yapılmıştır. GOO (FB: 0,1) sınıflama kriteri ile birlikte 14,58-33,36 aralığında madde sayısı ile sınıflama yapıldığı görülmektedir. Sınıflama kategori sayısı arttıkça sınıflama için gereken madde sayısı tüm sınıflama kriterleri için artmaktadır. OTU için MFB-KY madde seçme yöntemi ile GA (%90) sınıflama kriteri birlikte kullanıldığında oluşturulan koşulun en az madde ile en etkili desen olduğu anlaşılmaktadır.

Tablo 2. *Yeteneğin AOO ile Kestirildiği Çok Kategorili Sınıflamada Koşullara Ait OTU, OSD, r, Yanlılık, RMSE, OHM Değerlerinin Ölçme Kesinliği ve Test Etkililiği Açısından Karşılaştırılması*

Koşullar			*Bağımlı Değişkenler*					
Madde Seçme Yöntemi	*Sınıflama Kriteri*	*SKS*	*OTU*	*OSD*	*r*	*Yanlılık*	*RMSE*	*OMH*
MFB-KY	AOOT (FB= 0,1)	İki	33,35	0,899	0,984	0,003	0,184	0,141
		Üç	38,81	0,897	0,985	0	0,181	0,138
		Dört	46,81	0,885	0,987	0,001	0,17	0,128
	GA (%90)	İki	11,35	0,88	0,92	0,002	0,406	0,304
		Üç	20,16	0,874	0,945	0,05	0,339	0,239
		Dört	28,04	0,871	0,97	0,001	0,254	0,182
	GOO (FB: 0,1)	İki	14,58	0,893	0,931	0,001	0,378	0,277
		Üç	25,09	0,892	0,955	0.009	0,308	0,214
		Dört	33,36	0,882	0,972	0,001	0,17	0,128
MFB-KN	AOOT (FB= 0,1)	İki	24,989	0,891	0,873	0,2	0,627	0,418
		Üç	36,36	0,895	0,941	0,1	0,392	0,238
		Dört	48,27	0,887	0,982	-0,003	0,198	0,141
	GA (%90)	İki	14,68	0,891	0,873	0,2	0,627	0,503
		Üç	24,35	0,884	0,93	0,1	0,422	0,272
		Dört	31,63	0,873	0,969	-0,015	0,255	0,183
	GOO (FB= 0,1)	Iki	14,06	0,888	0,825	0,3	0,627	0,418
		Üc	24,79	0,892	0,917	0,13	0,465	0,295
		Dört	33,55	0,881	0,965	0,005	0,275	0,186
KLB-KY	AOOT (FB: 0,1)	İki	33,65	0,895	0,985	0,001	0,406	0,304
		Üç	38,96	0,896	0,985	-0,001	0,181	0,137
		Dört	46,8	0,887	0,987	0	0,171	0,128
	GOO (FB= 0,1)	İki	14,67	0,892	0,934	0,001	0,369	0,271
		Üç	25,05	0,891	0,955	0,009	0,307	0,214
		Dört	33,25	0,881	0,972	-0,005	0,243	0,169
	GA (%90)	İki	11,29	0,874	0,921	-0,01	0,402	0,303
		Üç	20,16	0,874	0,947	0,005	0,331	0,235
		Dört	27,89	0,871	0,97	-0,001	0,254	0,183
KLB-KN	AOOT (FB= 0,1)	İki	24,56	0,891	0,874	0,2	0,624	0,415
		Üç	36,32	0,895	0,941	0,11	0,392	0,239
		Dört	48,33	0,887	0,981	-0,004	0,199	0,141
	GOO (FB= 0,1)	İki	14,0	0,888	0,83	0,3	0,746	0,521
		Üç	24,35	0,893	0,92	0,12	0,457	0,291
		Dört	33,54	0,881	0,965	0,004	0,275	0,187
	GA (%90)	İki	14,49	0,89	0,841	0,3	0,72	0,499
		Üç	24,20	0,885	0,93	0,10	0,422	0,272
		Dört	31,49	0,872	0,97	-0,015	0,254	0,182

SKS: Sınıflama kategori sayısı, OTU: Ortalama test uzunluğu, OSD: Ortalama sınıflama doğruluğu, OMH: Ortalama mutlak hata, r: Gerçek yeteneklerle kestirilen yetenekler arası korelasyon

KLB-KY madde seçme yöntemi ile sınıflama kriterleri ve sınıflama kategori sayısıyla oluşturulan koşullara ait OTU için iki, üç ve dört kategorili sınıflamada, AOOT (FB: 0,1) sınıflama kriteri ile birlikte kullanıldığında 33,65-46,8 aralığında madde sayısı ile, GOO (FB: 0,1) sınıflama kriteri ile birlikte 14,67-33,25 aralığında madde sayısı ile, GA (%90) sınıflama kriteri ile birlikte 11,29-27,89 aralığında madde sayısı ile sınıflama yapıldığı anlaşılmaktadır. En az madde ile testi sonlandıran sınıflama kriteri GA (%90) sınıflama kriteridir. GA sınıflama kriteri her maddeden sonra belirlenen yetenek düzeylerini belirlenen güven aralığını kesme puanı ile karşılaştırdığı için KY temelli madde seçme yöntemleri ile performansının daha yüksek olduğu yorumu yapılabilir. GA yöntemi madde seçme yöntemlerinin tümüyle birlikte en az madde ile iki, üç ve dört kategorili tüm sınıflamalarda en etkili yöntemdir.

Kesme Noktası Temelli KLB ve MFB madde seçme yöntemlerinde iki, üç ve dört kategorili sınıflamada KLB madde seçme yöntemi kullanıldığında OTU için GA (%90) sınıflama kriteri 14,49-31,49 aralığında madde ile, GOO (FB: 0,1) sınıflama kriteri 14,0 -33,54 aralığında madde ile, AOOT (FB: 0,1) sınıflama kriteri ise 24,56-48,33 aralığında madde sayısı ile sınıflama yaptığı görülmektedir. En uygun yöntem en az madde ile sınıflama yapan GA yöntemidir. Sınıflama kategori sayısı arttıkça testi sonlandırmak için kullanılan madde sayısı artmıştır. Kesme noktası temelli MFB madde seçme yöntemi OTU için GA (%90) sınıflama kriteri ile 14,68-31,63 aralığında, GOO (FB: 0,1) sınıflama kriteri ile 14,06-33,55 aralığında, AOOT (FB: 0,1) sınıflama kriteri ile ise 24,98-48,27 aralığında madde ile sınıflama yaptığı tablodan anlaşılmaktadır. En az madde ile sınıflama yapan GA yöntemidir. Kesme noktası temelli madde seçme yöntemi sınıflama kriteri ve sınıflama kategori sayısı ile çaprazlandığında MFB yöntemi KLB ile OTU bakımından benzer sonuçlar gösterse de KLB yönteminin daha az madde ile sınıflama yaptığı anlaşılmıştır.

OSD değerleri iki, üç ve dört kategorili sınıflamada 0,871 ile 0,899 aralığındadır. AOOT sınıflama kriteri KY ve KN temelli madde seçme yöntemleri ve sınıflama kategori sayısı ile çaprazlandığında OSD 0,899-0,885 aralığındadır. Sınıflama kategori sayısı arttıkça OSD değeri düşmüştür. GA sınıflama kriteri KY ve KN temelli madde seçme yöntemleri ile 0,88-0,871 aralığında OSD ile sınıflama yapıldığı görülmektedir. AOOT sınıflama kriterine göre nispeten daha düşük OSD değeri hesaplanmıştır. GOO sınıflama kriteri kullanıldığında OSD için 0,893-0,882 aralığında değerler hesaplanmıştır. İki, üç ve dört kategorili sınıflamada kesme noktası ve kestirilen yetenek temelli MFB ve KLB madde seçme yöntemleri sınıflama kriterleri ve sınıflama kategori sayısı ile çaprazlandığında OSD için GOO ve AOOT kriterlerinin sınıflamanın doğruluğu olarak benzer, GA sınıflama kriteri yöntemi nispeten daha düşük değerler oluşturmuştur. OTU olarak GA yöntemi ile daha düşük madde sayısı ile sınıflama yaptığı görülmüştür. Sınıflama kategori sayısındaki artış OTU' nu artırırken OSD' nu azaltmıştır. GOO sınıflama kriterinin çok kategorili sınıflama için literatürde performansını destekleyecek araştırma örneklerine rastlanmamaktadır.

Simülasyon öncesi oluşturulan gerçek yetenek düzeyleri ile simülasyon sonrası kestirilen yetenek düzeyleri arasındaki korelasyon (r), MFB ve KLB kestirilen yetenek ve kesme noktası temelli madde seçme yöntemlerinin sınıflama yöntemleri ve sınıflama kategori sayısı ile çaprazlandığı araştırmada MFB-KY ve KLB-KY için 0,987-0,934 aralığında ve yüksektir. Sınıflama kategori sayısındaki artışa göre de korelasyon değeri artmıştır. MFB-KN ve KLB-KN için korelasyon değeri (r) 0,982-0,841 aralığında değişmektedir. KY temelli madde seçme yönteminin KN temelli madde seçme yöntemine göre kestirilen ve gerçek yetenek düzeyleri arasındaki korelasyon olarak daha etkili performans gösterdiği söylenebilir.

Yanlılık, kestirimin standart hatasını gösteren RMSE ve OMH değerleri için MFB-KY ve MFB-KN madde seçme yöntemlerinin AOOT sınıflama kriteri ile KY temelli MFB yöntemi kullanıldığında daha düşük yanlılık ve kestirimin standart hatası değerleri elde edilmiştir. Sınıflama kategori sayısı arttıkça

RMSE ve OMH değerleri düşmüş, yanlılık fazla etkilenmemiştir. GA ve GOO sınıflama kriterleri için de benzer durum olduğu söylenebilir. Ölçme kesinliği olarak kestirimin standart hatası, yanlılığın düşük, kestirilen yetenekler ile gerçek yetenekler arasındaki korelasyonun yüksek olması, test etkililiği olarak az madde ile sınıflama doğruluğunun yüksek olması BBST için beklenilen durumdur. Bu durumda KY temelli madde seçme yöntemlerinin ölçme kesinliği için daha yüksek performans gösterdiği söylenebilir.

3.3 Üçüncü Araştırma Sorusuna Ait Bulgular

Araştırmanın üçüncü alt probleminde çok kategorili sınıflamada sınıflama kategori sayısına göre sınıflama kriterleri, madde seçme yöntemleri ve yetenek kestirim yöntemleri çaprazlama olmadan kullanıldığında oluşturulan ölçme kesinliği ve test etkililiği değerleri Tablo 3 de gösterilmektedir.

Tablo 3'te OTU için iki, üç ve dört kategorili sınıflamada GA sınıflama kriterinin iki kategorili sınıflamada 11 madde ile sınıflama yaptığı anlaşılmaktadır. GOO sınıflama kriterinin 14 madde ile ikinci sırada, AOOT sınıflama kriterinin 33 madde ile üçüncü sırada sınıflama yaptığı görülmektedir. Sınıflama kategori sayısı arttıkça testin sonlanabilmesi için ihtiyaç duyulan madde sayısı artmaktadır.

Ortalama sınıflama doğruluğu için sınıflama kategori sayısı arttıkça sınıflamanın doğruluğu azalmaktadır. GA ve GOO yöntemlerinin sınıflama kategori sayısı düştükçe test etkililiği için uygun sınıflama kriterleri olduğu görülmektedir. Araştırmanın bu bulguları Nydick ve diğerleri (2012) ve Thompson (2011), Demir (2019) araştırma bulguları ile uyumludur. Yapılan bu araştırmalarda, GA ile yapılan sınıflamaların AOOT ile yapılan sınıflamalara kıyasla daha düşük OTU ve OSD ile sonlandığı ifade edilmiştir. Gündeğer (2017) ile iki kategorili sınıflamada nispeten uyumludur. Gündeğer (2017) yaptığı çalışmada, GA yönteminin AOOT'ye kıyasla OTU bakımından daha iyi OSD bakımından ise benzer performans gösterdiği sonucuna ulaşmıştır

Bireylerin gerçek yetenek düzeyleri ile kestirilen yetenek düzeyleri arasındaki korelasyon (r) sınıflama kategori sayısı arttıkça artmıştır. Korelasyon açısından en uygun değerin AOOT sınıflama kriteri ile yapılan sınıflamada hesaplandığı görülmektedir.

Yanlılık RMSE ve OMH bakımından sınıflama kategori sayısı arttıkça hata ve yanlılık düşmüştür. BBST de amaç az sayıda madde ile doğru olarak sınıflamaktır ve test etkililiği için GOO sınıflama kriterinin sınıflama kategori sayısı azaldıkça daha az sayıda madde ile daha düşük hata ve yanlılıkla daha doğru sınıflama yaptığı görülmektedir.

Madde seçme yöntemlerine göre kesme noktası temelli madde seçme yöntemleri kestirilen yetenek temelli madde seçme yöntemlerine göre sınıflama kategori sayısı azaldıkça daha az madde ile daha doğru sınıflama yaptıkları görülmüştür. KLB-KN ve MFB-KN temelli madde seçme yöntemleri ile OTU ve OSD için sınıflama kategori sayısı düştükçe daha etkili ve benzer sonuçlar hesaplanmıştır. Çok kategorili sınıflama için fazla çalışma literatürde bulunmadığından ortak bir görüş oluşturulamamıştır. MFB-KY ve KLB-KY madde seçme yöntemlerinin yanlılık RMSEA ve OMH değerleri için etkili oldukları görülmektedir.

Yetenek kestirim yöntemleri Beklenen sonsal dağılım (BSD), Ağırlıklı olabilirlik kestirimi (AOK) ve Bayes yetenek kestirimi (BYK) yöntemleri sınıflama kategori sayısına göre karşılaştırıldığında OTU ve OSD bakımından AOK ve BSD yöntemleri ile yapılan sınıflamada bezer değerlerin hesaplandığı görülmüştür. Yanlılık RMSE ve OMH açısından BSD daha iyi performans gösterdiği, AOK' un sınıflama kategori sayısı azaldıkça daha az madde ile sonuca ulaştığı görülmektedir. BYK yönteminin diğer iki yönteme oranla hata değeri yüksek OSD değeri daha düşük olduğu görülmektedir.

Tablo 3. *Çok kategorili sınıflamada Madde seçme yöntemleri, Yetenek Kestirim Yöntemleri, Sınıflama kriterlerine göre OTU, OSD ,r, yanlılık ,RMSE, OHM değerleri*

Koşullar		*Bağımlı Değişkenler*						
		SKS	*OTU*	*OSD*	*r*	*Yanlılık*	*RMSE*	*OMH*
Sınıflama Kriterleri	AOOT (FB= 0,1)	İki	33.241	0.896	0.984	-0.001	0.181	0.14
		Üç	37.929	0.885	0.985	-0.006	0.18	0.137
		Dört	46.699	0.874	0.987	0	0.169	0.127
	GA (%90)	İki	11.595	0.876	0.924	-0.005	0.396	0.297
		Üç	17.983	0.876	0.944	-0.007	0.342	0.252
		Dört	28.207	0.863	0.969	-0.001	0.256	0.183
	GOO (FB= 0,1)	İki	14.592	0.892	0.935	-0.001	0.367	0.271
		Üç	22.586	0.88	0.985	-0.006	0.297	0.216
		Dört	33.24	0.871	0.971	-0.003	0.246	0.17
Madde Seçme Yöntemleri	MFB-KY	İki	19.755	0.889	0.947	-0.001	0.332	0.238
		Üç	26.16	0.878	0.962	-0.003	0.281	0.202
		Dört	36.092	0.869	0.976	-0.001	0.227	0.16
	MFB-KN	İki	17.462	0.89	0.827	0.169	0.615	0.428
		Üç	27.99	0.891	0.927	0.062	0.405	0.262
		Dört	37.207	0.87	0.971	-0.003	0.251	0.173
	KLB-KY	İki	19.864	0.888	0.949	-0.003	0.326	0.234
		Üç	26.171	0.877	0.962	-0.007	0.281	0.201
		Dört	36.006	0.87	0.976	-0.002	0.227	0.16
	KLB-KN	İki	17.290	0.889	0.83	0.165	0.609	0.424
		Üç	27.959	0.891	0.928	0.062	0.402	0.261
		Dört	37.178	0.87	0.971	-0.004	0.251	0.173
Yetenek Kestirim Yöntemleri	BSD(EAP)	İki	19.889	0.889	0.951	-0.001	0.322	0.233
		Üç	26.166	0.878	0.963	-0.002	0.279	0.201
		Dört	36.287	0.871	0.976	0	0.225	0.159
	AOK (WLE)	İki	19.817	0.889	0.951	-0.001	0.322	0.233
		Üç	26.289	0.877	0.963	-0.004	0.28	0.201
		Dört	36.029	0.869	0.976	-0.001	0.226	0.16
	BYK (BKY)	İki	19.772	0.886	0.949	-0.005	0.33	0.236
		Üç	26.042	0.877	0.962	-0.008	0.284	0.203
		Dört	35.83	0.869	0.975	-0.004	0.231	0.161

SKS: Sınıflama kategori sayısı, OTU: Ortalama test uzunluğu, OSD: Ortalama sınıflama doğruluğu, OMH: Ortalama mutlak hata, r: Gerçek yeteneklerle kestirilen yetenekler arası korelasyon

4. Sonuç Tartışma ve Öneriler

Bu araştırmada BBST uygulaması için iki, üç ve dört kategorili sınıflamada gerçek uygulamaya en yakın sonuçları elde edebilmek amacıyla 25 tekrarın ortalaması alınarak test etkililiği ve ölçme kesinliği için en uygun sınıflama kriterleri, madde seçme yöntemleri ve yetenek kestirim yöntemlerini belirlemek amaçlanmıştır. Araştırma sonucunda en az madde ile en doğru sınıflama yapılarak test etkililiğini artıracak yöntemler belirlenmiştir. İki, üç ve dört kategorili sınıflamada Standart hata değerlerinin düşük olduğu, simülasyon sonunda bireylerin kestirilen yetenekleri ile gerçek yeteneklerinin arasındaki korelasyonun yüksek olduğu ölçme kesinliğini artıran yöntemler belirlenmiştir.

Ortalama test uzunluğu için GA sınıflama kriterinin iki, üç ve dört kategorili sınıflamada GOO ve AOOT sınıflama kriterlerine göre daha az madde ile sınıflama yaptığı sonucuna ulaşılmıştır. GA sınıflama kriteri her maddeden sonra belirlenen yetenek düzeylerini belirlenen güven aralığını kesme puanı ile karşılaştırdığı için KY temelli madde seçme yöntemleri ile performansının daha yüksek olduğu yorumu yapılabilir. Bu sonuçlar iki kategorili sınıflama için Nydick (2012) sonuçları ile benzerlik göstermektedir. Nydick (2012) de GA ile yapılan sınıflamaların AOOT ile yapılan sınıflamalara kıyasla daha düşük OTU ve OSD ile sonlandığı ifade edilmiştir. Sınıflama kategori sayısı artıkça test uzunluğu artmıştır.

Ortalama sınıflama doğruluğu için AOOT ve GOO yöntemlerinin GA yöntemine göre sınıflama kategori sayısı düştükçe daha etkili oldukları sonucuna ulaşılmıştır. Bu sonuçlar birlikte değerlendirildiğinde Test etkililiği için OTU düşük OSD yüksek olması beklenir. GOO sınıflama kriteri AOOT sınıflama kriterinin modifiye edilmiş halidir (Haring, 2014). Gerçek uygulamalarda GA ve GOO sınıflama kriterinin daha az madde ile daha yüksek doğrulukta sınıflama yapmaları nedeniyle uygulamada kullanımının tercih edilmesi önerilmektedir. Demir (2019) araştırmasında iki kategorili sınıflama için OTU açısından GA yönteminin, OSD için AOOT sınıflama kriterinin performansının başarılı olduğu sonucuna ulaşmıştır. Thompson (2009) GA ile yapılan sınıflamaların AOOT ile yapılan sınıflamalara kıyasla daha düşük OTU ve OSD ile sonlandığını ifade etmiştir. Araştırmanın bu sonucu Demir (2019) ve Thompson (2009) araştırma sonuçlarıyla nispeten uyumludur.

Kestirilen yetenekler ile gerçek yetenekler arasındaki korelasyonun sınıflama kategori sayısı arttıkça arttığı yanlılık RMSE ve OHM standart hataların düştüğü sonucuna ulaşılmıştır. Çok kategorili sınıflamanın daha hassas ölçümle yapıldığı hata değerlerinin düştüğü yorumu yapılabilir.

Madde seçme yöntemlerinin sınıflama kriterleriyle çaprazlandığı sonuçlarda kestirilen yetenek temelli MFB madde seçme yöntemi test etkililiği açısından daha az madde ile yüksek doğrulukta sınıflama için GA ve GOO yöntemi ile birlikte iki kategorili sınıflamalarda uygun bir desen olduğu sonucuna ulaşılmıştır. Üç, dört kategorili sınıflamalarda hata değeri düştüğü için daha hassas ölçme yapıldığı ölçme kesinliğinin arttığı söylenebilir. Nydick ve diğerleri (2012), Thompson (2009) araştırma sonuçlarıyla bu sonuçlar benzerlik göstermektedir. Thompson (2009) yaptığı çalışmada, GA sınıflama kriteri için KY temelindeki madde seçiminin, AOOT sınıflama kriteri için ise KN temelindeki madde seçiminin daha uygun olduğu sonucuna ulaşmıştır. Sonuçlar birlikte değerlendirildiğinde Özellikle sonuçları yüksek önem gösteren çok kategorili sınıflamalarda BBST' nin kullanılması hata değerleri daha düşük sınıflama yapıldığı için uygulamada önerilmektedir MFB madde seçme yöntemi ile GA ve GOO sınıflama kriterlerinin birlikte kullanımı test etkililiği yüksek sonuçlar oluşturduğu için uygulamada kullanılması önerilmektedir.

Üç, dört kategorili sınıflamalarda hata değeri düştüğü için daha hassas ölçme yapıldığı ölçme kesinliğinin arttığı sonucuna ulaşılmıştır. OTU sınıflama kategori sayısı arttıkça artmıştır. Nydick ve diğerleri (2012) ise üç ve beş kategoride yapılan sınıflamaları karşılaştırmışlar ve bireyler daha fazla

kategoride sınıflandığında OTU artarken OSD'nin azaldığı, diğer bir ifadeyle test etkililiğinin düştüğü sonucuna ulaşmışlardır. Nydick ve diğerleri (2012), araştırma sonuçlarıyla bu araştırmanın sonuçları benzerlik göstermektedir. Sonuçlar birlikte değerlendirildiğinde MFB madde seçme yöntemi ile GA ve GOO sınıflama kriterlerinin birlikte kullanılması test etkililiği yüksek sonuçlar oluşturduğu için uygulamada kullanılması önerilmektedir. OTU için MFB-KY madde seçme yöntemi tüm kategorilerde GA sınıflama kriteri ile en az madde ile sınıflama yaptığı sonucuna ulaşılmıştır. GA sınıflama kriteri madde seçme yöntemlerinin ikisi ile birlikte en az madde ile iki, üç ve dört kategorili tüm sınıflamalarda en etkili yöntemdir.

Araştırmada yetenek AOK ile kestirildiğinde KY temelli madde seçme yöntemleri kullanıldığında MFB madde seçme yönteminin performansının KLB madde seçme yönteminin performansından yüksek olduğu sonuçlanmıştır. Kesme noktası temelli madde seçme yöntemleri kullanıldığında MFB madde seçme yöntemi KLB ile OTU bakımından nispeten benzer sonuçlar gösterse de KLB yönteminin daha az madde ile sınıflama yaptığı sonucuna ulaşılmıştır.

Araştırmanın başka bir sonucu kestirilen yetenek ile gerçek yetenekler arasındaki korelasyon (r) madde seçme yöntemleri, sınıflama kriterleri ve sınıflama kategorileri çaprazlandığında sınıflama kategori sayısı arttıkça artmıştır. Kestirilen yetenek temelli MFB madde seçme yöntemi GOO sınıflama kriteri ile test etkililiği ve ölçme kesinliği açısından daha iyi performans gösterdiği sonucuna ulaşılmıştır. Uygulayıcıların test etkililiği ve ölçme kesinliği için MFB madde seçme yöntemi ile birlikte GOO sınıflama kriterini kullanması uygun olacaktır. GOO sınıflama kriterinin çok kategorili sınıflama için literatürde performansını destekleyecek araştırma örneklerine rastlanmamaktadır. GOO sınıflama kriterinin uygulamada OTU ve OSD için başka bir ifade ile ölçme kesinliği ve test etkililiği için uygulayıcılara avantaj sağlayacağı yorumu yapılabilir. GOO sınıflama kriteri AOOT sınıflama kriterinin modifiye edilmiş halidir. Güçlük parametresi kesme puanı ile eşleşmeyen daha uzaktaki maddeleri de verimli hale getiren özelliğinden dolayı madde havuzu daha etkin kullanıldığından AOOT sınıflama kriterine göre daha az madde ile sınıflama yaptığı yorumu yapılabilir. Uygulamada AOOT den daha avantajlı olabileceği söylenebilir.

MFB-KY temelli madde seçme yöntemi AOOT sınıflama kriteri ile birlikte kullanıldığında yanlılık, kestirimin standart hata değerlerinin daha düşük olduğu ölçme kesinliğinin yüksek olduğu sonucuna ulaşılmıştır. Uygulamada MFB-KY temelli madde seçme yöntemi test etkililiği için uygulayıcılara önerilebilir.

Araştırmada iki, üç ve dört kategorili sınıflama için madde seçme yöntemleri, yetenek kestirim yöntemleri ve sınıflama kriterleri ayrı ayrı incelendiğinde, MFB-KN temelli madde seçme yöntemi diğer yöntemlere göre daha az madde ile daha doğru sınıflama yapmaktadır.

Sınıflama kategori sayısı arttıkça tüm madde seçme yöntemleri için test etkililiği azalmıştır. Kestirilen yetenek ile gerçek yetenek düzeyi arasındaki korelasyon, yanlılık, RMSE, OMH olarak MFB-KY yöntemi daha iyi performans göstermiştir. Test etkililiği için MFB-KN yöntemi, ölçme kesinliği için MFB-KY temelli madde seçme yöntemi uygulayıcılara önerilmektedir. Spray ve Reckase (1994) araştırmasında da kesme noktasında en yüksek bilgiyi veren madde seçme yönteminin diğer yöntemlere kıyasla daha iyi performans sağlandığı görülmüştür. Bu sonuçlar Spray ve Reckase (1994) ile Lau ve Wang'ın (1998) sonuçlarıyla iki kategorili sınıflama için çok uyumlu değildir. Çok kategorili sınıflama için fazla çalışma literatürde bulunmadığından ortak bir görüş oluşturulamamıştır.

Araştırmada yetenek kestirim yöntemleri AOK, BSD, BYK ortalama test uzunluğu ve ortalama sınıflama doğruluğu açısından benzer performans göstermişlerdir. Standart hata değerleri, gerçek yetenekler ile kestirilen yetenekler arasındaki korelasyon açısından BSD' ın ölçme kesinliği daha fazladır. BSD yetenek kestiriminin performansının diğer yöntemlerden yüksek olduğu uygulamalarda

kullanılmasının uygun olacağı düşünülmektedir. Sınıflama kategori sayısı artıkça ölçme kesinliği ve sınıflama doğruluğu azalmıştır. Uygulayıcılar için BSD yetenek kestirimi daha az kategorili sınıflamalarda önerilebilir.

İki, üç ve dört kategorili sınıflamalarda sınıflama kriterleri diğer yöntemlerle çaprazlanmadan araştırıldığında GA yöntemi OTU bakımından en uygun yöntemdir, AOOT yöntemi OSD bakımından uygundur. Haring' e 2014 göre AOOT (FB=0,1) sınıflama kriteri farksızlık bölgesi ne kadar küçük olursa daha yüksek sınıflama doğruluğu ile daha fazla madde ile sınıflama yapmaktadır. Bu özelliğinden dolayı daha fazla madde ile yüksek doğrulukta sınıflama yaptığı söylenebilir. GOO yöntemi OTU için GA benzer OSD için AOOT yöntemine benzer performans göstermiştir. Test etkililiği için sınıflamada gereken madde sayısının az olması BBST' de beklenen özelliktir. GOO yöntemi uygulayıcılara az kategorili sınıflama için önerilmektedir.

Tüm araştırma koşullarında sınıflama kategori sayısının artması OTU 'nu arttırmış, OSD 'nu azaltmış test etkililiği düşmüştür. Korelasyon yanlılık değerleri sonuçları benzerken RMSE ve OMH değerlerinin düştüğü ölçme kesinliğinin arttığı söylenebilir. Bu sonuca dayanarak çok kategorili sınıflamanın daha hassas ölçme ile yapıldığı için yüksek risk taşıyan sınıflama alanlarında kullanılmasının uygun olacağı test etkililiğinin düşük olduğu sonucuna ulaşılmıştır.

Bu çalışma simülasyon verisi ile yapılmıştır. Gerçek veri ile araştırma tekrarlanabilir. Araştırmada pratik kısıtlamalar ve içerik dengeleme kullanılmamıştır. Pratik kısıtlamaların ve içerik dengelemenin sınıflamaya etkisi de araştırılabilir. Daha fazla koşul çaprazlanarak alanyazına katkı sağlayan araştırmalar yapılabilir.

Kaynaklar

Arce-Ferrer, A., Frisbie, D. A., & Kolen, M. J. (2002). Standard errors of proportions used in Reporting changes in school performance with achievement levels. *Educational Assessment, 8*(1), 59-75.

Demir, S. (2019). *Bireyselleştirilmiş bilgisayarlı sınıflama testlerinde sınıflama doğruluğunun incelenmesi [Investigation of classification accuracy in individualied computerized classification tests]* (Yayın No. 600532) [Doktora tezi, Hacettepe Üniversitesi]. YÖK. https://tez.yok.gov.tr/UlusalTezMerkezi/

Eckes, T. (2017). Rater effects: Advances in item response modeling of human ratings–Part I. *Psychological Test and Assessment Modeling*, *59*(4), 443-452.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23(3),* 249-261. https://doi.org/10.1177/01466219922031365

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60(5),* 713-734. https://doi.org/10.1177/00131640021970862

Gündeğer, C. (2017*). Bireyselleştirilmiş bilgisayarlı sınıflama testi kriterlerinin sınıflama doğruluğu ve test uzunluğu açısından karşılaştırılması [Comparison of adaptive computerized classification test criteria in terms of classification accuracy and test length]* (Yayın No. 483376) [Doktora tezi, Hacettepe Üniversitesi]. YÖK. https://tez.yok.gov.tr/UlusalTezMerkezi/

Haring, S. H. (2014). A comparison of three statistical testing procedures for computerized classification testing with multiple cutscores and item selection methods. (Doctoral dissertation, University of Texas at Austin). http://hdl.handle.net/2152/24838

Kaptan, S. (1995). *Bilimsel araştırma teknikleri ve istatistik teknikleri.* Rehber Yayınevi.

Kingsbury, G. G., & Weiss, D.J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing,* (pp. 237-254). Academic Press.

Lau, C. A. (1996). *Robustness of a unidimensional computerized testing mastery procedure with multidimensional testing data*. (Doctoral Dissertation, The University of Iowa).

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

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

Nydick, S. W., Nozawa, Y., & Zhu, R. (2012, Nisan). *Accuracy and efficiency in classifying examinees using computerized adaptive tests: An application to a large scale test.* The National Council on Measurement in Education (NCME) toplantısında sunulan bildiri, Vancouver, BritishColumbia, Canada. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.476.3381&rep=re p1&type=pdf

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

Nydick, S. W. (2014). *catirt: An R Package for Simulating IRT-Based Computerized Adaptive Tests.* https://cran.rproject.org/web/packages/catIrt/catIrt.pdf

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

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. & Reckase, M. D. (1994). The Selection of Test Items for Decision Making with a Computer Adaptive Test. *The Annual Meeting of the National Council on Measurement in Education. NewOrleans,* LA, 5-7 April 1994. https://eric.ed.gov/?id=ED372078

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics,* 21(4), 405-414. https://doi.org/10.3102/10769986021004405

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

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://doi.org/10.7275/wq8m-zk25

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

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

Extended Abstract

Introduction

Computerized Adaptive Tests (CAT) allow students to evaluate their ability effectively in a computer environment with fewer items in a shorter time. Computerized Adaptıve Classification Test (CACT) can be used to classify individuals according to a certain feature more accurately with fewer items. The CACT psychometric model consists of five components: starting point, item selection method, ability estimation method, and classification criteria. All classification criteria classify with fewer items than traditional tests (Kingsbury, & Weiss, 1983). The current study employed ability estimation methods Weighted Likelihood Estimation (WLE), The Expected a Posteriori (EAP), Bayesian Ability Estimation (BAE), Estimated Ability (EA) and Cut-off Point Based Maximum Fisher Information (MFI) and Kulback Laibler Information (KLI) as the item selection method, Sequential Probability Ratio Test (SPRT), Confidence Interval (CI) and Generalized Likelihood Ratio (GLR). The study aimed to determine certain methods that maximize the efficiency of CACT according to the number of classification categories in multi-category classification with CACT simulation and to provide evidence about classification accuracy.

The sub-problems of the research are as follows.

1. How do the mean classification accuracy, mean test length, correlation (r), bias, RMSE, and Mean Absolute Error (MAE) values of the conditions in which the classification criteria are crossed with item selection methods change when the ability is estimated with the WLE ability estimation method in two, three, and four-category classification with CACT simulation?

2. How do the mean classification accuracy, mean test length, correlation (r), bias, RMSE, MAE values of the conditions in which the item selection methods are crossed with the classification criteria change when the ability is estimated with EA method in the two, three and four-category classification with CACT simulation?

3. How do the mean classification accuracy, mean test length, correlation (r), bias, RMSE, MAE values of item selection methods, ability estimation methods and classification criteria change in measurement accuracy and test effectiveness in two, three and four-category classification?

Method

The current research adopts a descriptive design and Monte Carlo simulation study. In the study, the researchers created three classification criteria, four item selection methods, three classification categories (3x4x3) and 36 conditions. Following Thompson (2009, 2011), the researchers created 500 items with a three-parameter logistics model. As in the study by Weiss (1980), the parameters of the items were derived from the U (0.5-1.5) distribution of parameter a and the N (0-1) distribution of parameter b. In line with Thompson (2009), 500 items were derived from the normal distribution as parameter c N (0-0.3). The ability parameter was derived in R environment with 25 repetitions for 1000 individuals with an average of 0 and a standard deviation of 1 (R Core Team, 2013). For the classification criteria, the researchers determined the SPRT and GLR methods with 0.1 indifference region (IR) and the CI methods with 90% confidence level (CI) according to Nydick (2013) and Eggen and Straetmans (2000). The RI and CI used in the classification criteria indicate the tolerable level of error. The starting point was set to zero for all conditions. The researchers determined breakpoints, using ability parameters for two, three, and four-category classification. As in Eggen and Straetmans (2000), the study also divided the skill level into two and determined the cut-off point, taking 70% of each level as the first part 1st level and the other part 2nd level. Also, the research

utilized CatIrt (Nydick, 2014) package. The study investigated RMSE, MAE, bias, correlation between ability levels (r), average classification accuracy (ACA), average test length (ATL) values.

Result

The first sub-problem of the study, which estimated the ability with EAP, showed that the classification criterion requiring the fewest item to classify individuals under conditions where the classification criteria were crossed with item selection methods was the CI (90%) classification criterion at 12.02 and 12.58 values with two-category classification. In the two-category classification, the SPRT (IR=0.1) classification criterion with a value of 0.897 for ACA showed the best performance. The researchers observed that the SPRT (IR =0.1) classification criterion had the least error for the correlation between error values, bias and real abilities and estimated abilities. As the number of classification categories increased, the number of items required to end the test increased. The three-category classification revealed that the CI (90%) classification criterion for ATL was the condition classified with the minimum number of items (18.03) when the Kullback Leibler Information (KLI)-Estimated Ability (EA) item selection method were used together. In the classification with the Generalized Likelihood Ratio (GLR) classification criterion, the researchers obtained findings similar to the SPRT (IR = 0.1) classification criterion in CI (90%) and ACA in ATL. For the four-category classification, the SPRT (IR = 0.1) classification criterion with the fewest items as the CI classification criterion ATL showed the best performance for the classification as ACA.

Results revealed that the KLI method, one of the Estimated Ability-Based item selection methods, created close results for the ATL of the MFI method. The CI method was the most effective method in all classifications with at least two, three and four categories with the EA-based item selection methods. When the cut-off point based item selection method was crossed with the classification criteria and the number of classification categories, although the MFI method showed similar results as KLI and ATL, the MFI item selection method classified with fewer items. As the number of classification categories increased, RMSE and MAE values decreased, and bias was not affected. It showed that SPRT classification criteria were effective for measurement accuracy and CI classification criteria were effective for test effectiveness.

The second question of the study, when the ability was estimated with WLE, showed that the classification criterion that requires the fewest item to end the test and classify the individuals for the Estimated Ability-Based and Cut-off Point-Based item selection methods according to the values obtained under the conditions where the item selection methods and classification criteria were crossed was the two-category classification and the CI (90%) classification criterion in the value ranged between 11.29 and 14.68. The three-category classification, when the CI (90%) classification criterion for ATL was used with the KLI-EA and MFI-EA item selection methods with a value of 20.16, revealed that conditions were classified with the fewest item. In the three-category classification for ACA, SPRT (IR = 0.1) showed the best performance with a classification criterion of 0.897. The three-category classification for the correlation between bias, RMSE, MAE, and estimated abilities and actual abilities, showed that the SPRT (IR = 0.1) classification criterion created the highest correlation values with EA-based item selection methods. It also revealed that the SPRT classification criterion as bias and error values showed the best performance with all item selection methods. Thus, EA-based item selection methods performed higher for measurement accuracy.

The third sub-problem of the study showed that CI and GLR methods had appropriate classification criteria for test effectiveness as the number of classification categories decreased. In terms of correlation, the most appropriate value was in the classification with the SPRT classification

criterion. MFI-EA and KLI-EA material selection methods were effective for bias RMSE and MAE values. As the number of classification categories of WLE decreased, it caused the result with fewer items. The BAE method had a higher error value and a lower ACA value compared to the other two methods.

Conclusion and Discussion

The study concluded that the CI classification criterion for the average test length was classified with fewer items in the two-, three- and four-category classification compared to the GLR and SPRT classification criteria. Since the CI classification criterion compared the ability levels determined after each item with the cut- score of the determined confidence interval, its performance was higher with EA-based item selection methods. These results were similar to the results by Nydick (2012) for two-category classification. Nydick (2012) stated that the classifications with CI ended with lower ATL and ACA compared to the classifications with SPRT.

For the ACA, the SPRT and GLR methods were more effective as the number of classification categories decreased compared to the CI method. Considering these results, low ATL and high ACA were expected for test effectiveness. The GLR classification criterion is a modified version of the SPRT classification criterion (Haring, 2014). In real applications, it is recommended to use the CI and GLR classification criteria because they classify with higher accuracy with fewer items. Demir (2019) concluded that the performance of the CI method in ATL for two-category classification and the SPRT classification criterion for ACA was successful. Thompson (2009) stated that the classifications made with CI ended with lower ATL and ACA compared to the classifications with SPRT. This result was relatively consistent with those of Demir (2019) and Thompson (2009).

The research concluded that the correlation between the estimated abilities and the actual abilities increased as the number of classification categories increased, and the bias RMSE and MAE standard errors decreased. The multi-category classification was with more precise measurement and the error values decreased.

The results in which the item selection methods were crossed with the classification criteria showed that the EA-based MFI item selection method was appropriate in two-category classifications with the CI and GLR methods for high accuracy with fewer items in test effectiveness. Since the error value decreased in three- and four-category classifications, measurement accuracy increased. These results were similar to the results by Nydick et al. (2012) and Thompson (2009). Thompson (2009) concluded that the EA-based item selection was appropriate for the CI classification criterion and the Cut-off Point based (CB) item selection was appropriate for the SPRT classification criterion. Considering the results, the CACT is recommended, especially in multi-category classifications whose results are crucial, since the error values are classified lower.

Since the error value decreased in three- and four-category classifications, measurement accuracy increased. Nydick et al. (2012) compared the classifications in three and five categories and concluded that while ATL increased when individuals were classified in more categories, ACA and test effectiveness decreased. The research results were similar to the results by Nydick et al. (2012). Thus, it is recommended to use the MFI item selection method together with the CI and GLR classification criteria since the test effectiveness is high.

Estimation of ability with WLE showed that the performance of the MFI item selection method was higher than the performance of the KLI item selection method used with EA-based item selection methods. The research also showed that the EA-based MFI item selection method performed better in test effectiveness and measurement accuracy with the GLR classification criterion. It would be appropriate for practitioners to use the GLR classification criterion with the MFI item selection method

for test effectiveness and measurement accuracy. Thus, the GLR classification criterion will provide an advantage for ATL and ACA, for measurement accuracy and test effectiveness.

Considering the item selection methods, ability estimation methods, and classification criteria for two, three, and four-category classification, the MFI-EA method for test effectiveness and the MFI-EA-based item selection method for measurement accuracy are recommended to practitioners. Spray and Reckase (1994) observed that the item selection method, which gave the highest information at the cut-off point, provided better performance compared to other methods. These results were not consistent with the results by Spray and Reckase (1994) for two-category classification.

Considering the classification criteria in the two, three and four-category classifications without crossing with other methods, the CI method was the most appropriate method in ATL, and the SPRT method was in ACA. According to Haring, 2014, the smaller the SPRT (IR=0.1) classification criterion IR, the higher the classification accuracy and the more items were classified. Hence, it showed more items with high accuracy.

The current study employed simulation data. Future research could be replicated with real data. The study did not employ practical limitations and content balancing, so further research could also investigate the impact of practical limitations and content balancing on classification. Researchers can contribute to the literature by crossing more conditions.

Yayın Etiği Beyanı

Bu araştırma simülasyon verisi kullanılarak gerçekleştirildiği için etik kurul iznine gerek yoktur. Bu araştırmanın planlanmasından, uygulanmasına, verilerin toplanmasından verilerin analizine kadar olan tüm süreçte "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. Bu araştırmanın yazım sürecinde bilimsel, etik ve alıntı kurallarına uyulmuş; toplanan veriler üzerinde herhangi bir tahrifat yapılmamıştır. Bu çalışma herhangi başka bir akademik yayın ortamına değerlendirme için gönderilmemiştir.

Araştırmacıların Katkı Oranı Beyanı

Birinci yazar %70, ikinci yazar %30 oranında katkı sağlamıştır

Çatışma Beyanı

Araştırmanın yazarları olarak herhangi bir çıkar/çatışma beyanımız olmadığını ifade ederiz.