ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(1); 15-27

Bireyselleştirilmiş Bilgisayarlı Sınıflama Testlerinde Sınıflama Doğruluğu, Test Uzunluğu ve Ölçme Kesinliğinin İncelenmesi *

Seda DEMİR ** Burcu ATAR ***

Öz
Bu araştırmada, Bireyselleştirilmiş Bilgisayarlı Sınıflama Testlerinde (BBST) iki kategorili puanlanan maddelerden oluşan tek boyutlu madde havuzu üzerinden Ardışık Olasılık Oran Testi (AOOT) ve Güven Aralığı (GA) sınıflama kriterleri ile Kestirilen Yetenek Temelinde Maksimum Fisher Bilgisi (MFB-KY) ve Kesim Noktası Temelinde Maksimum Fisher Bilgisi (MFB-KN) madde seçme yöntemlerinin iki, üç veya dört kategoride yapılan sınıflamalarda Ortalama Sınıflama Doğruluğu (OSD), Ortalama Test Uzunluğu (OTU) ve ölçme kesinliği bakımından karşılaştırılması amaçlanmıştır. Araştırmada içerik dengeleme ve madde kullanım sıklığı kontrolü de yapılmış olup yetenek kestirim yöntemi olarak Ağırlıklandırılmış Olabilirlik Kestirimi (AOK) tercih edilmiştir. İçerik dengelemede, Kısıtlanmış Bireyselleştirilmiş Bilgisayarlı Test (KBBT) ve Modifiye Edilmiş Multinominal Model (MMM), madde kullanım sıklığı kontrolünde ise Sympson-Hetter Yöntemi (SH) ve Madde Uygunluk Yöntemi (MU) incelenmiştir. Araştırma kapsamında, R yazılımı kullanılarak türetilen 500 madde ve 5000 bireyden oluşan veriye dayanan Monte Carlo (MC) simülasyonu için toplam 48 koşul oluşturulmuş ve elde edilen değerler her bir koşul için 30 tekrarın ortalaması alınarak hesaplanmıştır. Çalışma sonucunda, OTU bakımından GA'nın, OSD bakımından ise AOOT'nin daha başarılı performans gösterdiği; MFB-KY madde seçme yönteminin MFB-KN'den daha kullanışlı olduğu; korelasyon, yanlılık, RMSE ve Ortalama Mutlak Hata (OMH) değerleri bakımından AOOT'nin daha etkili olduğu görülmüştür. Bununla birlikte, istenen içerik dengesini sağlamada MMM yöntemi daha başarılıyken, test etkililiği (OTU ve OSD) bakımından KBBT'nin daha avantajlı olduğu; madde kullanım sıklığı kontrolünde ise MU yöntemi daha iyi performans gösterirken, test etkililiğinde SH'nin daha avantajlı olduğu belirlenmiştir. Artan kategori sayısının ise OTU'yu artırırken OSD'yi azalttığı; ayrıca korelasyon, yanlılık, RMSE ve OMH değerlerinde iyileşme sağladığı görülmüştür.

Anahtar Kelimeler: Bireyselleştirilmiş bilgisayarlı sınıflama testi, içerik dengeleme, madde kullanım sıklığı kontrolü, sınıflama kriterleri, madde seçme yöntemleri.

GİRİŞ

Eğitimde test etmenin birçok farklı amacı olabilir. Bu amaçlar arasında eğitimin etkililiğini arttırma, öğrencileri bireysel olarak değerlendirme, seçme veya yerleştirme kararları alma, belgelendirme, öğrenmedeki ilerlemeyi izleme ve tanılayıcı amaçlı test etme sıralanabilir. Belirtilen bu amaçların gerçekleştirilebilmesinde, öğrencinin sahip olduğu yetenek düzeyi hakkında zamanında ve doğru bilgiye erişebilmenin kritik bir öneme sahip olduğu söylenebilir. Bu noktada, bilgi ve iletişim teknolojilerindeki gelişmelerin eğitim alanındaki en büyük yansımalarından biri olan Bireyselleştirilmiş Bilgisayarlı Testler (BBT), giderek daha nitelikli ve etkili değerlendirmeler yapılmasına katkıda bulunmaktadır.

Geleneksel kağıt-kalem testlerinden farklı olarak, BBT'de, farklı yetenek düzeylerine sahip olan bireyler, kişiselleştirilmiş performanslarına dayalı olarak farklı test formlarıyla gerçek zamanlı olarak test edilmektedir (Bao, Shen, Wang, & Bradshaw, 2021). BBT'nin amacı, her bir bireyin gizil yeteneğini kestirmek ve bireyin mevcut performansına dayalı olarak madde havuzundan birey için en

* Bu çalışma, "Bireyselleştirilmiş Bilgisayarlı Sınıflama Testlerinde Sınıflama Doğruluğunun İncelenmesi" isimli doktora tezinden üretilmiştir.

** Dr. Öğr. Üyesi, Tokat Gaziosmanpaşa Üniversitesi, Eğitim Fakültesi, Tokat-Türkiye, seddadmr@gmail.com, ORCID ID: 0000-0003-4230-5593
*** Doç. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, burcua@hacettepe.edu.tr, ORCID ID: 0000- 0003-3527-686X

Bu makaleye atıfta bulunmak için:
Demir, S., & Atar, B. (2021). Bireyselleştirilmiş bilgisayarlı sınıflama testlerinde sınıflama doğruluğu, test uzunluğu ve ölçme kesinliğinin incelenmesi. *Journal of Measurement and Evaluation in Education and Psychology, 12*(1), 15-27. doi: 10.21031/epod.787865

Geliş Tarihi: 30.08.2021
Kabul Tarihi: 21.02.2021

uygun (en çok bilgi sağlayan) test maddelerini seçmektir (Eggen & Straetmans, 2000). Süreç sonunda ise BBT'lerin geleneksel testlerden daha az sayıda maddeyle daha güvenilir yetenek kestirimleri sağladığını söylemek mümkündür (Bao ve diğerleri, 2021; Fan, Wang, Chang, & Douglas, 2012; Thompson, 2009). BBT'nin sağladığı bu avantajlar, Graduate Management Admission Test (GMAT), Graduate Record Examination (GRE), the National Assessment of Educational Progress (NAEP) gibi büyük ölçekli BBT uygulamalarının tercih edilmesinin temel nedeni olarak görülebilir. Bazen bireyleri test etmenin temel amacı etkili bir yetenek kestirimine ek olarak, geçti-kaldı gibi iki veya daha fazla kategoride yapılacak sınıflamaların doğruluğu olabilir. Bu durumda Bireyselleştirilmiş Bilgisayarlı Sınıflama Testlerinin (BBST) kullanımı tercih edilir. Sınıflama sonucuna bağlı olarak alınabilecek önemli kararlar (örneğin, lise bitirme, meslek seçimi veya yüksek riskli testlere bağlı kararlar) nedeniyle etkili ve doğru sınıflamalar yapabilmek kritik öneme sahiptir (Thompson & Ro, 2007).

Bunların yanı sıra test etkililiği hem BBT'ler hem de BBST'ler için önemli bir kavramdır. Tek boyutlu madde havuzuna sahip BBT uygulamalarında yüksek test etkililiği, yetenek kestirimi için mümkün olduğunca az madde ve düşük standart hatalar anlamına gelmektedir (van der Linden & Hambleton, 1996 aktaran, Thompson, 2009). BBST'lerde ise mümkün olduğunca az madde kullanılırken, düşük sınıflama hataları hedeflenerek test etkililiği sağlanmaktadır (Thompson, 2009).

Araştırmanın Amacı

BBST uygulamasına yönelik olarak alanyazın incelendiğinde, yapılan çalışmaların büyük çoğunluğunda, sınıflamanın sadece iki kategoride yapıldığı (örneğin, Gündeğer & Doğan, 2018a; Lau, 1996; Reckase, 1983; Spray & Reckase, 1996), içerik dengeleme ve madde kullanım sıklığı kontrolünün dikkate alınmadığı, genel olarak yalnızca sınıflama kriterlerinin (örneğin, Kingsbury & Weiss, 1980; Spray & Reckase, 1996; Thompson, 2009) veya madde seçme yöntemlerinin (örneğin, Gündeğer & Doğan, 2018b; Eggen, 1999; Lin & Spray, 2000) karşılaştırıldığı ya da farklı madde seçme yöntemleriyle çaprazlanan sınıflama kriterlerinin (örneğin, Eggen & Straetmans, 2000; Thompson & Ro, 2007) performanslarının incelendiği görülmektedir. Bunun yanı sıra sınıflama kriterinin performansının Ortalama Test Uzunluğu (OTU) ve Ortalama Sınıflama Doğruluğu (OSD) bakımından farklı madde kullanım sıklığı kontrol yöntemlerine göre karşılaştırıldığı (Huebner, 2012; Lau & Wang, 1999), Sympson-Hetter (SH) madde kullanım sıklığı kontrol yöntemiyle birlikte içerik dengeleme amacıyla spiral yöntemin kullanıldığı (Huebner & Li, 2012) az sayıda çalışmaya da rastlanmaktadır. Test sonuçlarına göre yapılacak doğru sınıflamaların bireyleri seçme, izleme veya yerleştirmede sağlayacağı katkılar da göz önüne alındığında, BBST çalışmalarına farklı araştırma desenleriyle yenilerinin katılmasının gerekli olduğu söylenebilir. Çalışmanın bu yönüyle alanyazına sağlayacağı katkı nedeniyle önemli olduğu düşünülmektedir.

Bu araştırmanın temel amacı, tek boyutlu ve iki kategorili maddelerden oluşan madde havuzundaki BBST uygulamalarında kullanılan farklı sınıflama kriterleri ve madde seçme yöntemlerinin performansının; yetenek kestirim yöntemi Ağırlıklandırılmış Olabilirlik Kestirimi (AOK) olduğunda, farklı içerik dengeleme yöntemleri, madde kullanım sıklığı kontrol yöntemleri ve sınıflama kategori sayılarına göre OSD, OTU, bireylerin gerçek ve kestirilen yetenek düzeyleri arasındaki korelasyon, yanlılık, Root Mean Squared Error (RMSE) ve Ortalama Mutlak Hata (OMH) değerleri bakımından nasıl değiştiğinin incelenmesidir. Araştırmanın alt problemleri aşağıda ifade edilmiştir.

Ardışık Olasılık Oran Testi (AOOT) sınıflama kriterinin farksızlık bölgesi (FB) sabiti δ: .20 düzeyi ve Güven Aralığı (GA) sınıflama kriterinin GA: %90 güven düzeyi için yetenek kestirim yöntemi AOK iken iki, üç veya dört kategorili sınıflamalarda;

1. Kestirilen Yetenek Temelinde Maksimum Fisher Bilgisi (MFB-KY) ve Kesim Noktası Temelinde Maksimum Fisher Bilgisi (MFB-KN) madde seçme yöntemlerinin,
2. MFB-KY, MFB-KN madde seçme yöntemleri ile Kısıtlanmış BBT (KBBT), Modifiye edilmiş Multinominal Model (MMM) içerik dengeleme yöntemleri ve Sympson-Hetter (SH), Madde Uygunluk (MU) madde kullanım sıklığı kontrol yöntemlerinin

bir arada ele alındığı koşullarda, OSD, OTU, bireylerin gerçek ve kestirilen yetenek düzeyleri arasındaki korelasyon, yanlılık, RMSE ve OMH değerleri nasıl değişmektedir?

Araştırmanın amacı doğrultusunda, ilerleyen bölümlerde simülasyon çalışmasının tasarımı, veri üretimi, BBST simülasyon koşulları ve verilerin analizi süreçleri açıklanmıştır. Ardından elde edilen araştırma sonuçları temel bulgular vurgulanarak özetlenmiş, OSD, OTU ve ölçme kesinliğinin gelecekteki araştırmalar için sonuçları tartışılarak önerilerde bulunulmuştur.

YÖNTEM

Bu çalışma, Monte Carlo (MC) simülasyonu olarak gerçekleştirilmiş ve BBST uygulama sonuçları simülatif veri setleri üzerinden karşılaştırılmıştır. Diğer araştırma yöntemleri, Ne oldu, nasıl oldu ve neden? sorularına yanıt veriyorken, simülasyon çalışmaları … olsa ne olur? sorusunun yanıtlanmasına yardımcı olmaktadır. Simülasyon çalışmalarında, geleceğe yönelik olası farklı koşullar oluşturulabildiği için daha karmaşık sistemlerin incelenebilmesi mümkündür (Dooley, 2002). Kullanılan veri setleri, araştırmada incelenen koşullara uygun şekilde R ortamında türetilmiştir (R Core Team, 2013). Araştırmanın bağımlı değişkenleri; OSD, OTU, bireylerin gerçek ve kestirilen yetenek düzeyleri arasındaki korelasyon (r), yanlılık, RMSE ve OMH değerleridir. Bağımsız değişkenleri ise; sınıflama kriterleri (AOOT, GA), madde seçme yöntemleri (MFB-KY, MFB-KN), içerik dengeleme yöntemleri (KBBT, MMM), madde kullanım sıklığı kontrol yöntemleri (SH, MU) ve sınıflama kategori sayısıdır (iki, üç, dört). Dolayısıyla araştırmada toplam, 2 sınıflama kriteri x 2 madde seçme yöntemi x 2 içerik dengeleme yöntemi x 2 madde kullanım sıklığı kontrol yöntemi x 3 sınıflama kategori sayısı= 48 adet koşul oluşturulmuştur.

Veri Üretimi

Bu çalışmada kullanılan veriler, belirli özelliklere uygun olarak simülasyonla üretilmiştir.

Monte Carlo (MC) simülasyonu için madde ve yetenek parametrelerinin türetilmesi

MC simülasyonu olarak gerçekleştirilen bu çalışmada, Thompson'ın (2011) çalışması dikkate alınmış ve madde havuzu, Madde Tepki Kuramı (MTK) modellerinden üç parametreli lojistik model (3 PLM) temel alınarak 30 tekrarın her biri için 500 maddeden oluşturulmuştur. Çalışmada hem KN hem de KY temelinde MFB'ye dayanan madde seçme yöntemleri kullanılarak iki, üç veya dört kategoride sınıflamalar yapıldığından madde havuzu, $\theta = 0$ kesme noktası ve etrafında yüksek bilgi veren ve yetenek düzeyi (-3, 3) aralığını kapsayan maddelerden oluşturulmuştur. Havuzdaki maddeler için a parametresi, Kingsbury ve Weiss'in (1980) çalışmaları dikkate alınarak, orta ve yüksek ayırıcılık düzeylerini temsil edebilmesi bakımından tek biçimli dağılımdan $U[0.5, 2.0]$ olacak şekilde; b parametresi, Thompson'ın (2009) ve Warm'ın (1989) çalışmalarında değinildiği gibi gerçek uygulamadaki değerlere yakın olabilmesi bakımından, normal dağılımdan $N(-0.5, 1.5)$ olacak şekilde; c parametresi ise Thompson'ın (2009) çalışması dikkate alınarak ve yine gerçek bir uygulamaya yakın olması düşünülerek, normal dağılımdan $N(0.20, 0.05)$ olacak şekilde türetilmiştir. Bunun yanı sıra bireylerin yetenek parametreleri (-3, +3) aralığında, ortalaması 0 ve standart sapması 1 olan normal dağılımdan ($N(0, 1)$) 30 tekrarın her biri için toplam 5000 birey olarak R yazılımında rastgele türetilmiştir.

BBST Simülasyon Koşulları

Bu çalışmada kullanılan BBST simülasyon koşulları alt başlıklar halinde detaylı olarak açıklanmıştır.

Başlama noktası

BBST uygulamalarında başlama noktası olarak bireyler hakkında mevcut olan ön bilgiler kullanılabilir (Weiss & Kingsbury, 1984; Yang, Poggio, & Glasnapp, 2006). Çok sık kullanılmamakla birlikte, popülasyon ortalaması da başlama noktası olarak belirlenebilir (Thompson, 2007b). Bu çalışmada tüm koşullar için başlama noktası $\theta = 0$ olarak belirlenmiştir.

Madde seçimi

Bilgisayarın havuzdaki kullanılmamış maddeleri değerlendirdiği ve sonraki madde olarak uygulanmasına karar verdiği en iyi maddenin seçildiği akıllı madde madde seçme yöntemleri genel olarak kesme puanı temelli ve kestirim temelli olmak üzere iki türde sınıflandırılmaktadır (Thompson, 2007b). Kullanılan psikometrik model MTK olduğunda tercih edilebilecek kesme puanı temelli yöntemler; MFB, Maksimum Kullback Liebler Bilgisi (KLB) ve Log-Odds Oranıdır (Lin & Spray, 2000). AOOT ile geleneksel olarak, kesme noktasında Fisher bilgisini en üst düzeye çıkaran madde seçme yöntemi kullanılmaktadır. Özellikle de farksızlık bölgesi genişledikçe, AOOT'nin daha iyi sonuçlar üretmesi beklenmektedir (Eggen, 1999). Bu çalışmada da MFB-KY ve MFB-KN madde seçme yöntemleri tercih edilmiştir.

Yetenek kestirimi

Alanyazın incelendiğinde, ikili puanlanan (1-0) ve tek boyutlu MTK'ya dayanan farklı yetenek kestirim yöntemlerinin bulunduğu görülmektedir. Bu yöntemlerden en temel kabul edilen ve sıklıkla rastlananlar; Maksimum Olabilirlik Kestirim yöntemi (MOK), Marjinal Maksimum Olabilirlik Kestirim yöntemi (MMOK), Ağırlıklandırılmış Olabilirlik Kestirim yöntemi (AOK) ve Bayesçi yetenek kestirim yöntemlerinden; Owen'ın Ardışık Bayesci kestirim yöntemi, Maksimum Sonsal Dağılım yöntemi (MSD) ile Beklenen Sonsal Dağılım yöntemidir (BSD). Warm (1989), tüm bu yöntemlerin bir miktar yanlı kestirimler yapabildiğini ifade etmiştir. Yanlılık, sistematik bir şekilde sınıflama kararlarının doğruluğunu etkilemektedir (Wang & Wang, 2001). Bununla birlikte, Warm (1989) yaptığı çalışmada, özellikle sabit uzunluklu testlerde, AOK yöntemiyle yapılan kestirimlerin, MOK ve MSD yöntemleriyle yapılan kestirimlere kıyasla daha az yanlılığa sahip olduğu sonucuna ulaşmıştır. Değişken uzunluktaki bireyselleştirilmiş testlerde ise testin MSD ile benzer ama MOK'tan daha az sayıda maddeyle sonlandığını ifade ederek yetenek kestirimi için MOK'un modifiye edilmiş bir hali olan AOK yöntemini önermiştir. Madde kullanım sıklığını düşürebilecek ve test süresini azaltarak testin kullanışlılığını artırabilecek bu durum BBT ve BBST uygulamaları için AOK yönteminin kullanımına yönelik bir avantaj olarak değerlendirilebilir. AOK, yanlılığı azaltan ve madde parametreleri ile yetenek düzeyine özgü ağırlıklandırma fonksiyonu temelinde çalışan bir yöntemdir (Warm, 1989). Alanyazındaki BBST uygulamalarında da çoğunlukla AOK'un tercih edildiği görülmektedir (Eggen & Straetmans, 2000; Nydick, Nozawa, & Zhu, 2012; Wouda, & Eggen, 2009; Yang ve diğerleri, 2006). Sağladığı avantajlar ve sınıflama literatüründeki yeri de göz önüne alınarak, bu çalışmada yetenek kestirim yöntemi olarak AOK ele alınmıştır. AOK yetenek kestirim yöntemi, simülasyonlarda sabit kalan bir koşuldur.

Sınıflama kriteri

BBST'de MTK'ya dayalı üç temel sınıflama kriterinden söz edilebilmektedir. Bunlar; AOOT, GA ve Bayesçi Karar Kuramıdır (BKK). Her üç sınıflama kriteri de sabit formlu geleneksel testlerden daha az madde gerektirmekle birlikte benzer düzeyde sınıflama doğruluğu sağlamaktadır (Kingsbury, & Weiss, 1983). Yapılan bazı araştırmalar, madde seçimi KY temelinde iken GA yönteminin, madde seçimi KN temelinde iken ise AOOT yönteminin daha etkili olduğunu göstermektedir (Eggen & Straetmans, 2000; Spray & Reckase, 1996; Thompson, 2009). Bunun yanı sıra AOOT yönteminin özellikle sınıflama doğruluğu bakımından, GA yönteminden daha etkili olduğunu gösteren çalışmalara da rastlanmaktadır (Eggen & Straetmans, 2000). Ayrıca Thompson'ın (2009) da dikkat çektiği gibi alanyazındaki BBST çalışmalarında en sık kullanılan sınıflama kriterinin AOOT olduğu

görülmektedir. Yapılan açıklamalar ışığında, bu çalışmanın odağındaki sınıflama kriterleri AOOT (δ: .20) ve GA (%90) olarak belirlenmiştir.

İçerik dengeleme

İçerik dengelemesi yapılmış olan BBT uygulamalarında, bireyler, içerik alanlarının her birini mümkün olduğu kadar uygun şekilde temsil eden, geçerliği daha yüksek bir testle karşılaşmaktadır. Alanyazın incelendiğinde, BBST çalışmalarında en sık kullanılan içerik dengeleme yöntemlerinin spiral yöntem (spiraling method; Kingsbury & Zara, 1989) (örneğin, Finkelman, 2008; Huebner, 2012) ve KBBT yöntemi (örneğin, Eggen & Straetmans, 2000; Huebner & Li, 2012) olduğu görülmektedir. Lin (2011) ise yaptığı çalışmada MMM içerik dengeleme yöntemini kullanmıştır. Ancak alanyazında, KBBT ve MMM yöntemlerini karşılaştıran bir BBST çalışmasına rastlanmamıştır. Bu nedenle bu çalışmada alanyazındaki çalışmalardan farklı olarak; KBBT ve MMM olmak üzere iki farklı içerik dengeleme yöntemi ele alınmıştır. İçerik dengeleme koşullarını sağlamak için testi sonlandırmadan önce kullanılacak minimum madde sayısı 10, maksimum madde sayısı ise 70 olarak belirlenmiştir. KBBT ve MMM yöntemlerinin araştırma koşullarına dahil edildiği durumlarda, R ortamında türetilen 500 maddeyle oluşturulmuş madde havuzu, rastgele madde atamasıyla dört içerik alanına ayrılmıştır. Ardından araştırmacı tarafından yazılan fonksiyon ve döngüler yardımıyla madde seçiminin bu içerik alanları dikkate alınarak yapılması sağlanmıştır. Belirlenen içerik alanlarının istenen oranları ise sırasıyla %40, %30, %20 ve %10 olacak şekilde ayarlanmıştır.

Madde kullanım sıklığı

Madde kullanım sıklığının kontol edilmediği BBT uygulamalarında, maddelerin yalnızca maksimum bilgiye göre seçilmesi, bazı maddelerin çoğu katılımcıya uygulanması sonucu çok fazla kullanılmasına neden olabilir. Buna karşılık, madde kullanım sıklığı kontrol teknikleri uygulandığında ise ölçüm hassasiyeti korunurken hem test güvenliği hem de madde havuzunun daha dengeli kullanımı sağlanır (Leroux ve diğerleri, 2019). Alanyazın incelendiğinde, BBST uygulamalarında en sık kullanılan madde kullanım sıklığı kontrol yöntemlerinin, rastgelelik stratejilerine dayanan tesadüfi madde seçimi ve koşullu seçim stratejisine dayanan SH yöntemi (Sympson & Hetter, 1985) olduğu görülmektedir. Rastgelelik stratejilerinin gerçekçi test koşullarında çok fazla etkili olmayacağı düşünüldüğünden, bu çalışmada SH yöntemi ve SH yöntemiyle aynı yaklaşımı temel alan MU yöntemine (van der Linden & Veldkamp, 2004) odaklanılmıştır. Madde kullanım sıklığı kontrolünde kullanılacak SH ve MU yöntemleri için maksimum madde kullanım sıklığı oranı ise Huebner (2012) ile Huebner ve Li'nin (2012) çalışmaları dikkate alınarak yaygın ve sıklıkla seçilen bir değer olan r_{max} = .20 (Leung, Chang, & Hau, 2002) olarak alınmıştır.

Sınıflama kategorilerinin sayısı

Alanyazındaki BBST çalışmalarında genellikle, geçti-kaldı gibi sadece iki kategorinin ve tek bir kesme noktasının kullanıldığı görülmüştür. Huebner (2012), Lin ve Spray (2000), Reckase (1983), Sie, Finkelman, Riley, and Smits (2015), Thompson (2009), van Groen, Eggen ve Veldkamp'ın (2016) yaptıkları çalışmalarda bireyleri geçti-kaldı gibi iki kategoride; Eggen (1999), Thompson (2007a) iki ve üç kategoride; Eggen ve Straetmans (2000) üç kategoride; Nydick ve diğerleri (2012) üç ve beş kategoride; Yang ve diğerleri (2006) ise beş kategoride sınıflamayı amaçlamışlardır. Bu çalışmada ise değişimleri karşılaştırabilmek amacıyla sınıflamalar iki, üç ve dört kategoride yapılmıştır. Yapılan sınıflamalar için kesme noktalarının belirlenmesinde, bireyler için R'da üretilen yetenek parametrelerinden faydalanılmıştır. Türetilen yetenek parametreleri, düşük yetenek düzeyinden yüksek yetenek düzeyine doğru sıralanmış ve iki kategoride yapılacak sınıflamalar için bir, üç kategoride yapılacak sınıflamalar için iki, dört kategoride yapılacak sınıflamalar içinse üç kesme noktası Eggen ve Straetmans'ın (2000) çalışmalarında kullanılan yöntemle belirlenmiştir. İki kategoride yapılan sınıflamalarda, düşükten yükseğe doğru sıralanmış yetenek düzeylerinin ilk yarısı düzey 1 ve ikinci yarısı düzey 2 olarak kodlanmış, ardından, düzey 1'deki en yüksek yetenek düzeyinin %70'i alınarak kesme noktası (KN = 0.00) belirlenmiştir. Benzer şekilde üç kategoride yapılacak

sınıflamalarda, sıralanmış yetenek düzeyleri düzey 1, düzey 2, düzey 3 şeklinde kodlanmış ve KN1 = -0.29, KN2 = 0.31 olarak, dört kategoride yapılacak sınıflamalarda ise sıralanmış yetenek düzeyleri düzey 1, düzey 2, düzey 3, düzey 4 şeklinde kodlanmış ve KN1 = -0.47, KN2 = -0.01, KN3 = 0.48 olarak belirlenmiştir.

Verilerin Analizi

Araştırma kapsamında oluşturulan 48 simülasyon koşulunun her biri için 30 tekrar yapılmış ve araştırmanın bağımlı değişkenlerine ait değerler, yapılan tekrarların ortalaması alınarak hesaplanmıştır. Gerçek ve kestirilen yetenek düzeyleri arasındaki ilişki Pearson Momentler Çarpımı Korelasyon Katsayısı olarak, yanlılık, RMSE ve OMH değerleri ise R ortamında yazılan fonksiyonlar yardımıyla elde edilmiştir.

Yanlılık, bireylerin kestirilen son yetenek düzeylerinin ($\widehat{\theta_i}$) gerçek yetenek düzeylerinden (θ_i) farkları toplamının birey sayısına (n) oranı alınarak hesaplanır (Miller & Miller, 2004):

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

RMSE, bireylerin kestirilen son yetenek düzeylerinin ($\widehat{\theta_i}$) gerçek yetenek düzeylerinden (θ_i) farklarının kareleri toplamının birey sayısına (n) oranının karekökü eşittir.

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

OMH ise bireylerin kestirilen son yetenek düzeylerinin ($\widehat{\theta_i}$) gerçek yetenek düzeylerinden (θ_i) farklarının mutlak değerleri toplamının birey sayısına (n) oranına eşittir.

$$\text{OMH} = \frac{\sum_{i=1}^{n}|\widehat{\theta_i} - \theta_i|}{n}$$

Bununla birlikte içerik dengeleme ve madde kullanım sıklığı kontrolü için de R ortamında madde seçme yöntemine ek fonksiyon ve döngüler yazılmıştır.

BULGULAR

Çalışmanın her bir alt problemine yönelik elde edilen bulgular alt başlıklar halinde sunulmuştur.

Birinci Alt Probleme İlişkin Bulgular

Araştırmanın birinci alt problemine ilişkin her bir koşul için yapılan 30 tekrarın ortalaması alınarak hesaplanan değerler Tablo 1'de sunulmuştur.

Tablo 1 incelendiğinde, iki, üç veya dört kategoride yapılan sınıflamalarda, AOOT sınıflama kriteri kullanıldığında elde edilen OSD'nin .82 ile .94 aralığında ve oldukça yüksek, OTU'nun ise 22.95 ile 42.88 aralığında olduğu, buna karşılık GA sınıflama kriteri kullanıldığında elde edilen OSD'nin .71 ile .90 aralığında nispeten daha düşük, OTU'nun ise 11.33 ile 13.82 aralığında olduğu görülmektedir. Buna göre, BBST simülasyonunda yetenek kestirim yöntemi AOK iken, OTU bakımından GA yöntemi, OSD bakımından ise AOOT yöntemi daha iyi sonuçlar vermiştir. Bunun yanı sıra MFB-KY veya MFB-KN madde seçme yöntemi aynı sınıflama kriteriyle birlikte kullanıldığında test etkililiği bakımından benzer sonuçlar elde edilmiştir. Ayrıca sınıflama kategori sayısındaki artış her iki sınıflama kriteri için de test etkililiğini düşürmüş, başka bir ifadeyle OTU'yu artırırken OSD'yi azaltmıştır.

Tablo 1. Sınıflama Kriterlerinin (SK) ve Madde Seçme Yöntemlerinin (MSY) Sınıflama Kategori Sayısı (SKS) İki, Üç veya Dört Olduğunda Ortalama Test Uzunluğu (OTU), Ortalama Sınıflama Doğruluğu (OSD) ve Korelasyon (r), Yanlılık, RMSE, Ortalama Mutlak Hata (OMH) Değerleri İçin Ölçme Kesinliği Açısından Karşılaştırılması

SK	MSY	SKS	OTU	OSD	*r*	Yanlılık	RMSE	OMH
AOOT (δ = .20)	MFB-KY	İki	24.72	.94	.94	-0.011	0.35	0.27
		Üç	34.08	.88	.96	-0.012	0.32	0.24
		Dört	41.34	.82	.96	-0.014	0.29	0.22
	MFB-KN	İki	22.95	.94	.90	0.019	0.44	0.32
		Ü	33.93	.89	.92	0.015	0.38	0.28
		Dört	42.88	.82	.93	0.012	0.35	0.26
GA (%90)	MFB-KY	İki	11.33	.89	.90	0.016	0.46	0.35
		Üç	12.52	.79	.91	0.015	0.45	0.35
		Dört	13.81	.71	.91	0.016	0.44	0.34
	MFB-KN	İki	11.55	.90	.87	0.019	0.49	0.38
		Üç	12.62	.80	.87	0.017	0.48	0.37
		Dört	13.82	.71	.88	0.020	0.47	0.36

Not. AOOT= Ardışık olasılık oran testi, GA = güven aralığı, MFI-KY = kestirilen yetenek temelinde maksimum Fisher bilgisi, MFI-KN = kesim noktası temelinde maksimum Fisher bilgisi.

Bireylerin gerçek ve kestirilen yetenek düzeyleri arasındaki korelasyon (*r*) değerleri; sınıflama kriterlerinden AOOT için .90 ile .96 aralığında, GA için ise .87 ile .91 aralığında hesaplanmıştır. Sınıflama kriterlerinin madde seçme yöntemleriyle çaprazlandığı koşullar incelendiğinde ise her iki sınıflama kriteri için de MFB-KN'ye kıyasla MFB-KY madde seçme yönteminin kullanıldığı koşullarda daha yüksek korelasyonlar hesaplanmıştır. Bununla birlikte sınıflama kategori sayısındaki artışa karşılık benzer korelasyon değerleri elde edilmiştir. AOOT ile MFB-KY yöntemleri birlikte kullanıldığında hesaplanan yanlılığın (-0.014, -0.011 aralığında), AOOT ile MFB-KN'nin birlikte kullanıldığı (0.012, 0.019 aralığında) koşullara kıyasla daha düşük olduğu görülmektedir. Benzer şekilde GA ile birlikte MFB-KY kullanıldığında hesaplanan yanlılık (0.015, 0.016 aralığında) GA ile birlikte MFB-KN'nin kullanıldığı (0.017, 0.020 aralığında) koşullara kıyasla daha düşüktür. Yanlılıkla birlikte kestirimin standart hatasını da dikkate alan RMSE değeri ve OMH değeri için de durum benzerdir. Buna göre, AOOT sınıflama kriteri veya MFB-KY madde seçme yöntemi kullanıldığında, daha düşük yanlılık, RMSE ve OMH değerlerinin hesaplandığı söylenebilir. Ayrıca kategori sayısındaki artış yanlılığı çok fazla etkilemezken RMSE ve OMH değerlerini kısmen düşürmüştür.

İkinci Alt Probleme İlişkin Bulgular

Araştırma koşullarına içerik dengeleme için KBBT veya MMM, madde kullanım sıklığı kontrolü için ise SH veya MU yöntemlerinin dahil edildiği ikinci alt probleme ilişkin her bir koşul için yapılan 30 tekrarın ortalaması alınarak hesaplanan değerler Tablo 2'de sunulmuştur.

Tablo 2'ye göre MMM içerik dengeleme yönteminin kullanıldığı tüm koşullarda; uygulanan içerik oranlarının istenen içerik oranlarını (sırasıyla %40, %30, %20 ve %10) sağladığı görülmektedir. KBBT içerik dengeleme yönteminin kullanıldığı koşullarda ise uygulanan içerik oranlarının istenen içerik oranlarının kısmen üstünde veya altında değerler aldığı görülmektedir. Örneğin Tablo 2'de AOOT ile MFB-KN yöntemlerinin birlikte kullanıldığı, madde kullanım sıklığının MU ile kontrol edildiği ve sınıflamanın dört kategoride yapıldığı koşulda KBBT ile uygulanan içerik oranları sırasıyla yaklaşık %32, %28, %23 ve %16 olarak hesaplanmıştır. Ayrıca MU madde kullanım sıklığı kontrol yönteminin kullanıldığı koşullarda; Madde Kullanım Sıklığını Aşan Maddelerin Yüzdesi (MKSAMY) daha düşük olup Madde Kullanım Sıklığını Aşan Maddelerin Madde Kullanım Sıklıklarının Ortalamasının (MKSAMO) istenen r_{max} = .20 değerini sağladığı söylenebilir. Buna karşılık SH yönteminin kullanıldığı koşullarda MKSAMY değerlerinin daha yüksek olduğu ve MKSAMO değerlerinin ise istenen r_{max} = .20 değerinin oldukça üstünde olduğu görülmektedir. Örneğin Tablo 2'de AOOT ile MFB-KY yöntemlerinin birlikte kullanıldığı, içerik dengelemenin KBBT ile yapıldığı ve sınıflamanın dört kategoride yapıldığı koşulda SH ile kontrol edilen madde kullanım sıklığı için

hesaplanan MKSAMY değeri yaklaşık .25 iken MKSAMO değeri .29 olarak hesaplanmıştır. Başka bir ifadeyle maddelerin yaklaşık %25'i maksimum madde kullanım sıklığı oranını (r_{max} = .20) aşmış ve bu maddelerin madde kullanım sıklığı ortalaması yaklaşık .29 olarak hesaplanmıştır.

Tablo 2. Araştırma Koşullarına İçerik Dengeleme Yöntemleri (İDY) ve Madde Kullanım Sıklığı Kontrol Yöntemleri (MKSKY) Dahil Edildiğinde, Sınıflama Kriterlerinin (SK) ve Madde Seçme Yöntemlerinin (MSY) Sınıflama Kategori Sayısı (SKS) İki, Üç veya Dört Olduğunda Ortalama Test Uzunluğu (OTU), Ortalama Sınıflama Doğruluğu (OSD) ve Korelasyon (r), Yanlılık, RMSE, Ortalama Mutlak Hata (OMH) Değerleri İçin Ölçme Kesinliği Açısından Karşılaştırılması

SK	MSY	İDY	MKSKY	SKS	Uygulanan İçerik Oranı				MKS AMY	MKS AMO	OTU	OSD	*r*	Yanlılık	RMSE	OMH
AOOT	MFB-KY	KBBT	SH	İki	36.27	29.61	21.78	12.34	.14	.28	26.91	.94	.94	-0.014	0.36	0.28
(δ= .20)			r_{max} = .20	Üç	34.34	28.95	22.56	14.15	.21	.29	37.50	.87	.95	-0.015	0.32	0.24
				Dört	33.32	28.65	22.94	15.09	.25	.29	44.69	.81	.96	-0.017	0.30	0.22
			MU	İki	35.56	29.38	22.09	12.97	.09	.20	29.87	.93	.94	-0.017	0.37	0.28
			r_{max} = .20	Üç	33.71	28.77	22.79	14.74	.15	.20	41.71	.86	.95	-0.018	0.33	0.25
				Dört	32.81	28.47	23.15	15.57	.18	.20	48.50	.78	.96	-0.018	0.31	0.23
		MMM	SH	İki	39.82	29.98	20.09	10.10	.14	.28	27.42	.94	.94	-0.015	0.37	0.28
			r_{max} = .20	Üç	39.91	29.98	20.04	10.07	.21	.29	37.86	.87	.95	-0.015	0.33	0.25
				Dört	39.92	29.99	20.05	10.04	.26	.29	45.35	.80	.96	-0.017	0.30	0.23
			MU	İki	39.80	30.03	20.06	10.11	.10	.20	30.82	.93	.94	-0.015	0.37	0.28
			r_{max} = .20	Üç	39.84	30.01	20.08	10.08	.15	.20	42.27	.85	.95	-0.016	0.33	0.25
				Dört	39.86	30.01	20.07	10.07	.17	.20	49.01	.77	.96	-0.018	0.32	0.24
	MFB-KN	KBBT	SH	İki	37.00	29.71	21.46	11.83	.13	.33	25.70	.94	.90	0.009	0.44	0.33
			r_{max} = .20	Üç	34.41	28.98	22.49	14.12	.21	.34	38.35	.87	.92	0.009	0.39	0.28
				Dört	32.96	28.56	23.07	15.41	.25	.36	47.02	.79	.93	0.006	0.36	0.26
			MU	İki	35.83	29.40	21.96	12.81	.11	.20	30.55	.93	.90	0.002	0.44	0.33
			r_{max} = .20	Üç	33.46	28.74	22.90	14.91	.19	.20	43.94	.84	.92	0.003	0.40	0.30
				Dört	32.40	28.39	23.30	15.92	.23	.20	51.14	.75	.93	0.000	0.38	0.28
		MMM	SH	İki	39.81	29.96	20.12	10.11	.13	.33	26.18	.94	.90	0.009	0.43	0.33
			r_{max} = .20	Üç	39.88	30.00	20.08	10.04	.21	.34	38.61	.87	.92	0.008	0.39	0.29
				Dört	39.92	30.01	20.03	10.05	.25	.36	47.36	.79	.93	0.006	0.36	0.26
			MU	İki	39.75	29.95	20.16	10.13	.12	.20	31.01	.93	.90	0.005	0.44	0.33
			r_{max} = .20	Üç	39.83	30.00	20.07	10.10	.19	.20	44.85	.84	.92	0.003	0.40	0.30
				Dört	39.84	30.01	20.05	10.09	.23	.20	51.93	.75	.93	0.006	0.39	0.29

(devam ediyor)

Tablo 2 (devam)

SK	MSY	İDY	MKSKY	SKS	Uygulanan İçerik Oranı				MKS AMY	MKS AMO	OTU	OSD	*r*	Yanlılık	RMSE	OMH
GA	MFB-KY	KBBT	SH	İki	42.13	30.95	19.09	7.83	.04	.27	11.13	.88	.89	0.011	0.48	0.37
%90			r_{max} = .20	Üç	41.60	30.85	19.33	8.22	.05	.27	12.24	.78	.90	0.014	0.48	0.37
				Dört	41.01	30.69	19.64	8.66	.05	.27	13.37	.70	.90	0.013	0.47	0.36
			MU	İki	42.04	30.90	19.17	7.90	.03	.20	11.19	.88	.89	0.016	0.49	0.38
			r_{max} = .20	Üç	41.56	30.80	19.40	8.24	.03	.20	12.25	.78	.89	0.013	0.49	0.37
				Dört	41.03	30.70	19.64	8.63	.03	.20	13.39	.69	.90	0.014	0.48	0.37
		MMM	SH	İki	39.97	30.04	19.96	10.04	.04	.27	11.18	.88	.89	0.015	0.48	0.37
			r_{max} = .20	Üç	39.95	30.01	20.01	10.02	.05	.27	12.28	.78	.90	0.011	0.47	0.36
				Dört	40.07	29.93	20.00	9.99	.05	.27	13.45	.70	.90	0.012	0.47	0.36
			MU	İki	40.06	29.94	20.01	9.99	.03	.20	11.21	.88	.89	0.012	0.49	0.38
			r_{max} = .20	Üç	40.05	30.01	19.95	9.98	.03	.20	12.31	.78	.90	0.012	0.48	0.37
				Dört	40.01	30.00	19.99	9.99	.03	.20	13.51	.69	.90	0.013	0.48	0.37
	MFB-KN	KBBT	SH	İki	41.96	30.93	19.19	7.92	.05	.36	11.36	.89	.86	0.012	0.50	0.39
			r_{max} = .20	Üç	41.49	30.77	19.40	8.34	.06	.36	12.46	.79	.87	0.01	0.49	0.39
				Dört	40.88	30.64	19.68	8.80	.06	.36	13.69	.70	.87	0.01	0.48	0.38
			MU	İki	42.02	30.90	19.13	7.96	.05	.20	11.42	.88	.85	0.009	0.52	0.41
			r_{max} = .20	Üç	41.43	30.75	19.46	8.37	.05	.20	12.56	.77	.86	0.004	0.51	0.40
				Dört	40.55	30.55	19.86	9.03	.05	.20	14.60	.66	.86	0.006	0.51	0.39
		MMM	SH	İki	39.97	30.01	20.00	10.01	.06	.35	11.37	.89	.86	0.012	0.50	0.39
			r_{max} = .20	Üç	39.91	30.02	20.07	9.99	.06	.35	12.56	.79	.87	0.009	0.49	0.38
				Dört	39.97	30.02	20.00	10.01	.07	.35	13.65	.69	.87	0.013	0.49	0.38
			MU	İki	40.00	30.02	19.99	9.98	.05	.20	11.49	.88	.86	0.006	0.52	0.41
			r_{max} = .20	Üç	40.04	29.98	20.01	9.97	.05	.20	12.74	.78	.86	0.008	0.51	0.40
				Dört	39.99	29.98	20.02	10.00	.05	.20	14.42	.67	.87	0.009	0.50	0.39

Not: AOOT= Ardışık Olasılık Oran Testi, GA= Güven Aralığı, MFI-KY= Kestirilen Yetenek Temelinde Maksimum Fisher Bilgisi, MFI-KN= Kesim Noktası Temelinde Maksimum Fisher Bilgisi, KBBT= Kısıtlanmış BBT, MMM= Modifiye Edilmiş Multinominal Model, SH= Sympson-Hetter Yöntemi, MU= Madde Uygunluk Yöntemi, rmax= Maksimum Madde Kullanım Sıklığı Oranı, MKSAMY= Madde Kullanım Sıklığını Aşan Maddelerin Yüzdesi, MKSAMO= Madde Kullanım Sıklığını Aşan Maddelerin Madde Kullanım Sıklıklarının Ortalaması

Tablo 2'de, aynı sınıflama kriteri ve madde seçme yöntemi üzerinden bir karşılaştırma yapıldığında, test etkililiği bakımından çok küçük farklarla KBBT içerik dengeleme yöntemi daha iyi performans

göstermiş olsa da genel olarak MMM ile benzer sonuçlar verdikleri görülmektedir. Bunun yanı sıra SH madde kullanım sıklığı kontrol yöntemi de MU'ya kıyasla daha iyi bir performans göstermiştir. Genel olarak, OTU bakımından elde edilen en iyi sonucun (OTU = 11.13 ve OSD = .88) GA ve MFB-KY'nin KBBT ve SH ile birlikte kullanıldığı ve sınıflamanın iki kategoride yapıldığı koşulda, en kötü sonucun ise (OTU = 51.93 ve OSD = .75) ise AOOT ve MFB-KN'nin, MMM ve MU ile birlikte kullanıldığı ve sınıflamanın dört kategoride yapıldığı koşulda gözlendiği görülmektedir. Başka bir ifadeyle, en iyi ve en kötü sonuçlar arasında OTU'nun yaklaşık 5 katına çıktığı, OSD'nin oldukça düştüğü söylenebilir.

AOOT'nin kullanıldığı koşullar için hesaplanan korelasyon (*r*) değerleri .90 ile .96 aralığında iken GA'nın kullanıldığı koşullar için .85 ile .90 aralığında hesaplanmıştır. Yanlılık değerlerinin ise AOOT'nin kullanıldığı koşullarda (-0.018, 0.009) aralığında değişirken GA'nın kullanıldığı koşullarda (0.004, 0.016) aralığında değiştiği görülmektedir. Bunun yanı sıra en yüksek RMSE değeri (0.52) ve en yüksek OMH değeri (0.41) GA, MFB-KN, KBBT (veya MMM), MU yöntemlerinin birlikte kullanıldığı ve sınıflamanın iki kategoride yapıldığı koşulda gözlenirken en düşük RMSE değeri (0.30) AOOT, MFB-KY, KBBT (veya MMM), SH yöntemlerinin birlikte kullanıldığı, en düşük OMH değeri (0.22) ise AOOT, MFB-KY, KBBT, SH yöntemlerinin birlikte kullanıldığı ve sınıflamanın dört kategoride yapıldığı koşulda gözlenmiştir.

Özetle, Tablo 1'den elde edilen bulgulara paralel olarak; OTU bakımından GA'nın, OSD bakımından AOOT'nin daha iyi performans gösterdiği, sınıflama kategori sayısı arttıkça OTU'nun arttığı, OSD'nin ise azaldığı, korelasyon (*r*), yanlılık, RMSE ve OMH değerleri bakımından ise AOOT'nin GA'dan, MFB-KY'nin ise MFB-KN'den daha başarılı bir performans gösterdiği görülmektedir. Ayrıca artan kategori sayısına karşılık, korelasyon ve yanlılık değerleri benzer sonuçlar verirken RMSE ve OMH değerleri nispeten düşmüştür.

SONUÇLAR ve TARTIŞMA

Bu araştırmanın temel odak noktası sınıflama doğruluğu olduğundan farklı koşullarda OSD için hesaplanan değerler, sunulan bulguların yorumlanmasında büyük öneme sahiptir. Tüm araştırma koşullarında OSD bakımından yüksek değerler elde edilmesinin yanı sıra AOOT sınıflama kriterinin GA'ya kıyasla daha iyi performans gösterdiği ve bireyleri doğru kategorilere sınıflama oranının daha yüksek olduğu sonucuna ulaşılmıştır. Buna karşılık OTU bakımından ise tüm araştırma koşullarında GA sınıflama kriterinin AOOT'ye kıyasla daha iyi performans gösterdiği ve bireyleri sınıflamada daha az sayıda madde gerektirdiği belirlenmiştir. Araştırmanın bu bulgusu Gündeğer ve Doğan (2018a), Nydick ve diğerleri (2012), Thompson (2009) ile Thompson ve Ro'nun (2007) araştırma sonuçlarıyla da benzerlik göstermektedir. Yapılan bu araştırmalarda genel olarak, GA ile yapılan sınıflamaların AOOT ile yapılan sınıflamalara kıyasla daha düşük OTU ve OSD ile sonlandığı ifade edilmiştir. Dolayısıyla araştırmada ele alınan AOOT ve GA sınıflama kriterleri, sınıflama doğruluğu bakımından değerlendirildiğinde, daha yüksek OSD'lerin hesaplandığı AOOT yönteminin tercih edilmesi önerilebilir. Buna karşılık OTU bakımından yapılan değerlendirme sonucu bireylerin sınıflanıp testin sonlanabilmesi için daha az sayıda madde gerektiren GA sınıflama kriterinin tercih edilmesi söz konusu olabilir. Ancak yüksek riskli testler (örneğin, Tıp alanında uygulanan ve doğrudan insan hayatıyla ilgili olan testler) dikkate alındığında artan madde sayısına rağmen daha yüksek sınıflama doğruluğu sağlayan yöntemin tercih edilmesinin çok daha önemli olduğu unutulmamalıdır. Bunların yanı sıra BBST uygulamalarında, test etkililiği adı altında OTU ve OSD genellikle birlikte değerlendirilmektedir. Dolayısıyla test etkililiği bakımından en başarılı sınıflama kriterinin tercih edilmesi gerektiğinde, her iki sınıflama kriterinin de iyi düzeyde sınıflama doğruluğu sağladığı koşullar için, GA sınıflama kriterinin kullanılması önerilebilir.

Araştırma sonucunda, ölçme kesinliği bakımından; AOOT sınıflama kriterinin GA'dan, MFB-KY madde seçme yönteminin ise MFB-KN'den daha iyi performans gösterdiği tespit edilmiştir. Öyle ki, AOOT sınıflama kriterinin veya MFB-KY madde seçme yönteminin kullanıldığı koşullarda; bireylerin gerçek ve kestirilen yetenek düzeyleri arasında daha yüksek korelasyon değerleri, daha düşük yanlılık, RMSE ve OMH değerleri hesaplanmıştır. Buna göre, sınıflama kriteri AOOT olduğunda veya madde

seçme yöntemi MFB-KY olduğunda bireylerin son yetenek düzeylerinin daha hassas ve gerçek yetenek düzeylerine daha yakın kestirildiği yorumu yapılabilir. Bu durumun kaynağı olarak, madde havuzunun $\theta = 0$ kesme noktası ve etrafında yüksek bilgi veren maddelerden oluşturulmuş olması görülebilir. Ayrıca test etkililiği bakımından da MFB-KY madde seçme yöntemi MFB-KN yöntemine kıyasla nispeten daha başarılı sonuçlar vermiştir. Başka bir ifadeyle MFB-KY madde seçme yöntemi kullanıldığında, daha düşük OTU ve benzer OSD oranları elde edilmiştir. Ayrıca bireylerin gerçek ve kestirilen yetenek düzeyleri arasındaki korelasyon değerlerinin oldukça yüksek olduğu, özellikle AOK yetenek kestirim yönteminin AOOT sınıflama kriteri ve MFB-KY madde seçme yöntemiyle birlikte kullanıldığı koşullarda daha yüksek korelasyonlar hesaplandığı görülmüştür. Buradan yola çıkarak, yetenek kestiriminde kullanılan AOK yönteminin başarılı bir performans gösterdiği söylenebilir.

Tablo 1 ve Tablo 2'de sunulan bulgular karşılaştırıldığında; içerik dengeleme ve madde kullanım sıklığı kontrolü araştırma koşullarında dahil edildiğinde beklentiye uygun olarak kısmen daha yüksek OTU ve daha düşük OSD elde edildiği görülmüştür. Thompson'a (2007b) göre ise içerik dengeleme ve madde kullanım sıklığı kısıtlamaları genellikle yalnızca OTU'da artışa neden olmaktadır. BBST uygulamasında içerik dengeleme ve madde kullanım sıklığı kontrolü yapıldığında OTU'da gözlenen artış ve OSD'de gözlenen düşüşün, uygulanan içerik alanında birey hakkında yeterince bilgi sağlayan ve madde kullanım sıklığı oranını aşmamış bir maddenin bulunmamasından kaynaklanmış olabileceği yorumu yapılabilir. Bu sorunun çözümü için her bir içerik alanında, maksimum madde kullanım sıklığını aşan maddelerin yoğunlaştığı yetenek ranjındaki madde sayısı arttırılarak madde havuzu genişletilebilir. Araştırma koşullarına dahil edilen içerik dengeleme ve madde kullanım sıklığı kontrol yöntemleri, bireylerin gerçek ve kestirilen yetenek düzeyleri arasındaki korelasyonu değiştirmezken, yanlılıkta düşüşe, RMSE ve OMH değerlerinde ise bir miktar artışa sebep olmuştur. Ayrıca, GA sınıflama kriteri ile elde edilen sonuçların çok az etkilendiği görülmüştür. Bu durum GA'nın sağladığı bir avantaj olarak yorumlanabilir.

İçerik dengeleme yöntemlerinden MMM'in, istenen içerik oranını sağlamada KBBT'ye kıyasla daha iyi performans gösterdiği, buna karşılık test etkililiği söz konusu olduğunda, GA sınıflama kriterinin kullanıldığı koşullarda çok küçük değişiklikler olsa da AOOT'nin kullanıldığı koşullarda özellikle OTU bakımından KBBT'nin daha iyi performans gösterdiği sonucuna ulaşılmıştır. Araştırmanın bu bulgusu Lin'in (2011) çalışmasıyla da desteklenmektedir. Lin (2011), KBBT'nin BBST'lerde en sık tercih edilen içerik dengeleme yöntemlerinden olmasına rağmen alanyazında daha çok BBT'lerde kullanılmış olan MMM yönteminin istenen içerik dengesini sağlamada daha başarılı olduğunu ifade etmiştir. Dolayısıyla BBST uygulamasında, yüksek riskli testlerde olduğu gibi içerik dengeleme kritik öneme sahipse MMM yönteminin kullanımı tercih edilebilir. Buna karşılık test etkililiği daha kritik öneme sahipse de KBBT yönteminin kullanımı önerilebilir.

Madde kullanım sıklığını, MU yönteminin SH yöntemine kıyasla daha iyi kontrol ettiği görülmüştür. Araştırmanın bu bulgusu Huebner'in (2012) araştırma sonuçlarıyla da örtüşmektedir. Huebner (2012) çalışmasında, madde kullanım sıklığı kontrolü bakımından MU yönteminin SH'den daha etkili olduğu sonucuna ulaşmıştır. Test etkililiği bakımından ise özellikle AOOT sınıflama kriterinin kullanıldığı koşullarda, SH'nin performansının daha iyi olduğu görülmüştür. Öyle ki, SH yöntemi kullanıldığında daha düşük OTU ve daha yüksek OSD hesaplanmıştır. Bu sonuca dayanarak, BBST uygulamasında madde kullanım sıklığı kontrolü, başka bir ifadeyle sınavın/madde havuzunun güvenliği, kritik öneme sahipse MU yönteminin kullanımı önerilebilir. Buna karşılık test etkililiği daha kritik öneme sahipse de SH yönteminin kullanımı tercih edilebilir.

Tüm araştırma koşullarında; artan kategori sayısının OTU'yu artırırken OSD'yi azalttığı, başka bir ifadeyle test etkililiğini düşürdüğü görülmüştür. Araştırmanın bu bulgusu Eggen (1999) ile Nydick ve diğerlerinin (2012) araştırma sonuçlarıyla da desteklenmektedir. Eggen (1999) çalışmasında iki ve üç kategoride yapılan sınıflamaları, 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. Dolayısıyla test etkililiği bakımından düşünüldüğünde, sınıflama kategori sayısının mümkün olduğunca az belirlenmesi önerilebilir. Ayrıca sınıflama kategori sayısındaki artışa rağmen korelasyon ve yanlılık değerleri

benzer sonuçlar verirken RMSE ve OMH değerlerinin nispeten düşmüş olması araştırmanın önemli bulguları arasındadır. Bu bulguya göre, sınıflama kategori sayısı arttıkça testin sonlanması için gereken madde sayısının da artması sebebiyle, bireylerin son yetenek düzeylerinin daha hassas kestirildiği yorumu yapılabilir. O halde korelasyon, yanlılık, RMSE ve OMH değerleri de dikkate alınarak sınıflama kategori sayısının daha ideal şekilde belirlenebileceği söylenebilir.

Araştırma sonuçları genel bir çerçeveden değerlendirildiğinde, uygulayıcılara yönelik olarak BBST uygulamasının odak noktası OSD iken içerik dengeleme ve madde kullanım sıklığı kontrolü de kritik öneme sahipse; korelasyon, yanlılık, RMSE ve OMH değerleri bakımından da daha iyi performans gösteren AOOT sınıflama kriterinin, MFB-KY madde seçme yöntemi, MMM içerik dengeleme yöntemi ve MU madde kullanım sıklığı kontrol yöntemiyle birlikte kullanılması önerilebilir. Buna karşılık BBST uygulamasının odak noktası OTU iken içerik dengeleme ve madde kullanım sıklığı kontrolü de yapılacaksa GA sınıflama kriterinin, MFB-KY, MMM ve MU yöntemleriyle birlikte kullanılması önerilebilir. Araştırmacılara yönelik olarak ise yapılacak benzer BBST çalışmalarında çok boyutlu veya sahip olduğu madde sayısı bakımından farklı büyüklükteki, sivri, basık vb. özellikler taşıyan madde havuzlarının kullanımı önerilebilir. Ayrıca simülasyon çalışmasıyla sınırlı tutulan bu çalışmadan farklı olarak ileride yürütülecek çalışmalarda, temel BBST bileşenlerinin performansları, gerçek veri üzerinden de karşılaştırılabilir.

KAYNAKÇA

Bao, Y., Shen, Y., Wang, S., & Bradshaw, L. (2021). Flexible computerized adaptive tests to detect misconceptions and estimate ability simultaneously. *Applied Psychological Measurement, 45*(1), 3-21. doi: 10.1177/0146621620965730

Dooley, K. (2002). Simulation research methods. In J. Baum (Ed.), *Companion to organizations* (pp. 829-848). London: Blackwell.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*(3), 249-261. doi: 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. doi: 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. doi: 10.3102/1076998611422912

Finkelman, M. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics, 33*(4), 442-463. doi: 10.3102/1076998607308573

Gündeğer, C., & Doğan, N. (2018a). A comparison of computerized adaptive classification test criteria in terms of test efficiency and measurement precision. *Journal of Measurement and Evaluation in Education and Psychology, 9*(2), 161-177. doi: 10.21031/epod.401077

Gündeğer, C., & Doğan, N. (2018b). The effects of item pool characteristics on test length and classification accuracy in computerized adaptive classification testings. *Hacettepe University Journal of Education, 33*(4), 888-896. doi: 10.16986/HUJE.2016024284

Huebner, A. (2012). Item overexposure in computerized classification tests using sequential item selection. *Practical Assessment, Research & Evaluation, 17*(12), 1-9. Retrieved from https://pareonline.net/getvn.asp?v=17&n=12

Huebner, A., & Li, Z. (2012). A stochastic method for balancing item exposure rates in computerized classification tests. *Applied Psychological Measurement, 36*(3), 181-188. doi: 10.1177/0146621612439932

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: MN. Retrieved from http://iacat.org/sites/default/files/biblio/ki80-04.pdf

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). New York: Academic Press.

Kingsbury, G. G., & Zara, A.R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359-375. doi: 10.1207/s15324818ame0204_6

Lau, C. A. (1996). *Robustness of a unidimensional computerized testing mastery procedure with multidimensional testing data* (Unpublished doctoral dissertation). University of Iowa, Iowa City IA.

Lau, C. A., & Wang, T. (1999, April). *Computerized classification testing under practical constraints with a polytomous model*. Paper presented at the annual meeting of the American Educational Research Association (AERA), Montreal, Canada. Retrieved from http://iacat.org/sites/default/files/biblio/la99-01.pdf

Leroux, A. J., Waid-Ebbs, J. K., Wen, P-S., Helmer, D. A., Graham, D. P., O'Connor, M. K, & Ray, K. (2019). An investigation of exposure control methods with variable-length cat using the partial credit model. *Applied Psychological Measurement, 43*(8),624-638. doi: 10.1177/0146621618824856

Leung, C.-K., Chang, H. H., & Hau, K. T. (2002). Item selection in computerized adaptive testing: Improving the a-stratified design with the Sympson–Hetter algorithm. *Applied Psychological Measurement, 26*(4), 376-392. doi: 10.1177/014662102237795

Lin, C. (2011). Item selection criteria with practical constraints for computerized classification testing. *Applied Psychological Measurement 71*(1), 20-36. doi: 10.1177/0013164410387336

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. Retrieved from https://eric.ed.gov/?id=ED445066

Miller, I., & Miller, M. (2004). *John E. Freund's mathematical statistics with applications*. (7th Ed.). New Jersey: Prentice Hall.

Nydick, S. W., Nozawa, Y., & Zhu, R. (2012, April). *Accuracy and efficiency in classifying examinees using computerized adaptive tests: An application to a large-scale test.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Vancouver, British Columbia, Canada. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.476.3381&rep=rep1&type=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.R-project.org/

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: latent trait theory and computerized adaptive testing*, (pp. 237-254). New York: Academic Press.

Sie, H., Finkelman, M. D., Riley, B., & Smits, N. (2015). Utilizing response times in computerized classification testing. *Applied Psychological Measurement, 39*(5), 389-405. doi: 10.1177/0146621615569504

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. doi: 10.3102/10769986021004405

Sympson, J. B., & Hetter, R. D. (1985, October). Controlling item exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 937-977). San Diego, CA: Navy Personnel Research and Development Center. Retrieved from http://www.iacat.org/content/controlling-item-exposure-rates-computerized-adaptive-testing

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

Thompson, N. A. (2007b). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment Research & Evaluation, 12*(1), 1-13. Retrieved from http://www.iacat.org/sites/default/files/biblio/th07-01.pdf

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

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

Thompson, N. A., & Ro, S. (2007). Computerized classification testing with composite hypotheses. In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC conference on computerized adaptive testing*. Retrieved from http://www.iacat.org/sites/default/files/biblio/cat07nthompson.pdf

Van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29*(3), 273-291. doi: 10.3102/10769986029003273

Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2016). Multidimensional computerized adaptive testing for classifying examinees with within-dimensionality. *Applied Psychological Measurement, 40*(6), 387-404. doi: 10.1177/0146621616648931

Wang, S., & Wang, T. (2001). Precision of warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement, 25*(4), 317–331. doi: 10.1177/01466210122032163

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427-450. doi: 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. Retrieved from https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Wouda, J. T., & Eggen, T. J. H. M. (2009). Computerized classification testing in more than two categories by using stochastic curtailment. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC conference on computerized adaptive testing*. Retrieved from http://iacat.org/sites/default/files/biblio/cat09wouda.pdf

Yang, X., Poggio, J. C., & Glasnapp, D. R. (2006). Effects of estimation bias on multiple category classification with an IRT-based adaptive classification procedure. *Educational and Psychological Measurement, 66*(4), 545-564. doi: 10.1177/0013164405284031