BİREYSELLEŞTİRİLMİŞ BİLGİSAYARLI SINIFLAMA TESTİ KRİTERLERİNİN SINIFLAMA DOĞRULUĞU VE TEST UZUNLUĞU AÇISINDAN KARŞILAŞTIRILMASI

A COMPARISON OF COMPUTERIZED ADAPTIVE CLASSIFICATION TEST CRITERIA IN TERMS OF CLASSIFICATION ACCURACY AND TEST LENGTH

Ceylan GÜNDEĞER

Hacettepe Üniversitesi

Eğitim Bilimleri Anabilim Dalı, Eğitimde Ölçme ve Değerlendirme Bilim Dalı

Doktora Tezi

olarak hazırlanmıştır.

2017

KABUL ve ONAY

Eğitim Bilimleri Enstitüsü Müdürlüğü'ne,

Ceylan GÜNDEĞER'in hazırladığı "Bireyselleştirilmiş Bilgisayarlı Sınıflama Testi Kriterlerinin Sınıflama Doğruluğu Ve Test Uzunluğu Açısından Karşılaştırılması" başlıklı bu çalışma jürimiz tarafından **Eğitim Bilimleri Anabilim Dalı, Eğitimde Ölçme ve Değerlendirme Bilim Dalı'nda Doktora Tezi** olarak kabul edilmiştir.

Başkan	Prof. Dr. Selahattin GELBAL	
Üye (Danışman)	Prof. Dr. Nuri DOĞAN	
Üye	Prof. Dr. Hülya KELECİOĞLU	
Üye	Doç. Dr. Şeref TAN	
Üye	Yrd. Doç. Dr. Hamide Deniz GÜLLEROĞLU	

ONAY

Bu tez Hacettepe Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği'nin ilgili maddeleri uyarınca yukarıdaki jüri üyeleri tarafından 13 / 10 / 2017 tarihinde uygun görülmüş ve Enstitü Yönetim Kurulunca ..... / ..... / ..... tarihinde kabul edilmiştir.

Prof. Dr. Ali Ekber ŞAHİN
Eğitim Bilimleri Enstitüsü Müdürü

YAYIMLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.
Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

☐ **Tezimin/Raporumun tamamı dünya çapında erişime açılabilir ve bir kısmı veya tamamının fotokopisi alınabilir.**
(Bu seçenekle teziniz arama motorlarında indekslenebilecek, daha sonra tezinizin erişim statüsünün değiştirilmesini talep etseniz ve kütüphane bu talebinizi yerine getirse bile, teziniz arama motorlarının önbelleklerinde kalmaya devam edebilecektir)

☒ **Tezimin/Raporumun 13.10.2019 tarihine kadar erişime açılmasını ve fotokopi alınmasını (İç Kapak, Özet, İçindekiler ve Kaynakça hariç) istemiyorum.**
(Bu sürenin sonunda uzatma için başvuruda bulunmadığım takdirde, tezimin/raporumun tamamı her yerden erişime açılabilir, kaynak gösterilmek şartıyla bir kısmı veya tamamının fotokopisi alınabilir).

☐ **Tezimin/Raporumun ..... tarihine kadar erişime açılmasını istemiyorum ancak kaynak gösterilmek şartıyla bir kısmı veya tamamının fotokopisinin alınmasını onaylıyorum.**

☐**Serbest Seçenek/Yazarın Seçimi:** .....
.....

13 / 10 / 2017

Ceylan GÜNDEĞER

ETİK BEYANNAMESİ

Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

Ceylan GÜNDEĞER

TEŞEKKÜR

Yüksek lisans eğitimimden bu yana birlikte çalışma olanağı yakalayabildiğim için kendimi çok şanslı hissettiğim, akademik birikimi ve kişisel özelliklerini her zaman örnek aldığım ve alacağım, zor zamanlarda verdiği destekle çalışmaya motive olmamı sağlayan değerli danışmanım Prof. Dr. Nuri DOĞAN'a teşekkür ederim.

Derslerinin yanında özellikle danışmanımın yurt dışında olduğu bir yıl içerisinde vermiş olduğu destekle tez çalışmamın derinleşmesini sağlayan sevgili hocam Prof. Dr. Hülya KELECİOĞLU'na ve lisansüstü eğitimimde önemli rolü olan, tezime yaptığı katkılarla eksiklerimi görmemi sağlayan sevgili hocam Prof. Dr. Selahattin GELBAL'a çok teşekkür ederim.

Tezimin her kritik aşamasında kıymetli yorumlarını esirgememiş olan sayın hocam Doç. Dr. Şeref TAN'a; tez savunma sınavımdaki katkılarından ve içtenliğinden ötürü Yrd. Doç. Dr. Hamide Deniz GÜLLEROĞLU'na teşekkürlerimi sunarım.

Bilginin paylaştıkça çoğalan bir hazine olduğunu unuttuğumuz günlerde benimle kendi tez verisini paylaşarak araştırmamın önemli bir kısmını tamamlamamı sağlayan Yrd. Doç. Dr. Fatih KEZER'e; çalışmamın her aşamasında bilgisine ve deneyimine başvurduğum, yolumu bulmamda ciddi emeği olan canım arkadaşım Arş. Gör. Ezgi NEVRUZ'a; beyin fırtınasına ihtiyaç duyduğum zor anlarda bilgi ve yorumlarını esirgemeyen oda arkadaşlarım Arş. Gör. Dr. Çiğdem AKIN ARIKAN ve Arş. Gör. Sümeyra SOYSAL'a; özellikle tez teslim sürecindeki yardımları için arkadaşlarım Arş. Gör. Mine ZORLU ve Arş. Gör. Haydar KARAMAN'a çok teşekkür ederim.

Hayatım boyunca desteklerini ve bana olan güvenlerini hissettiğim canım annem Firdes GÜNDEĞER, canım babam İlhan GÜNDEĞER ve canım kardeşim Ceyhan GÜNDEĞER'e ***ne kadar teşekkür etsem azdır***… İyi ki varsınız, iyi ki benim ailemsiniz. Ailemi seçme şansım olsa yine sizi seçerdim.

Arkadaşlığın anlamını ve önemini sayelerinde öğrendiğim Meltem Gizem ŞATIR, Besen ŞATIR, Hazal NEVRUZ, Başaran KARABULUT, Mehmet Kemal GÜMÜŞ, Müge GÜMÜŞ ve Gülcan SARUGAN'a her daim yanımda oldukları ve bana güç verdikleri için çok teşekkür ederim.

BİREYSELLEŞTİRİLMİŞ BİLGİSAYARLI SINIFLAMA TESTİ KRİTERLERİNİN SINIFLAMA DOĞRULUĞU VE TEST UZUNLUĞU AÇISINDAN KARŞILAŞTIRILMASI

Ceylan GÜNDEĞER

ÖZ

Bireyselleştirilmiş Bilgisayarlı Sınıflama Testleri (BBST) bireyi, önceden belirlenen bir ya da birden fazla sayıda kesme noktasına göre en az sayıda maddeyle en yüksek sınıflama doğruluğunda sınıflamayı amaçlar. Bu sınıflamaların etkililiği, madde havuzlarına, sınıflama kriterlerine, madde seçme ve yetenek kestirim yöntemlerine göre değişkenlik göstermektedir. Buna göre BBST'de farklı desenlerin oluşturulması ve bu desenlerin Monte Carlo (MC) ve Post Hoc (PH) simülasyonlar altında incelenmesi gerçek uygulamalar için önem arz etmektedir.

Bu çalışmada BBST'de farklı sınıflama kriterleri, yetenek kestirim ve madde seçme yöntemleri hem MC hem de PH simülasyonları altında, sınıflama doğruluğu, test uzunluğu ve ölçme kesinliği bakımından karşılaştırılmıştır. Araştırmada sınıflama kriterlerinden Ardışık Olasılık Oran Testi (AOOT), Genelleştirilmiş Olabilirlik Oranı (GOO) ve Güven Aralığı (GA) yöntemleri; yetenek kestirim yöntemlerinden Beklenen Sonsal Dağılım (BSD) ve Ağırlıklandırılmış Olabilirlik Kestirimi (AOK) yöntemleri; madde seçme yöntemlerinden ise kesme noktasında (KN) ve kestirilen yetenek (KY) temelinde Maksimum Fisher Bilgisi (MFB) ve Kullback-Leibler Bilgisi (KLB) yöntemleri incelenmiştir. Bu amaçla MC simülasyonu için 3 PLM temel alınarak kesme noktası 1,0 ve etrafında yüksek bilgi verecek şekilde 500 maddelik bir havuz oluşturulmuş; PH simülasyonu için ise 80 maddelik gerçek veri setinden yararlanılmıştır. MC simülasyonunda birey yetenekleri normal dağılım yardımıyla (N(0,1)) toplam 3000 kişi üzerinden türetilmiştir. PH simülasyonunda ise veri setindeki 994 bireyin yetenek düzeyleri 3 PLM temelinde BSD ile kestirilmiştir. MC simülasyonunda bireylerin madde cevap örüntüleri R yazılımda rasgele türetilmiş; PH simülasyonda ise herhangi bir manüpülasyon olmaksızın gerçek madde cevap örüntüsü kullanılmıştır. Çalışmada PH ve MC simülasyonları için toplam 96 koşul incelenmiştir. BBST simülasyonu sonunda, ortalama test uzunluğu (OTU), ortalama sınıflama doğruluğu (OSD), bireylerin gerçek yetenek düzeyleriyle

kestirilen yetenek düzeyleri arasındaki korelasyon (r), yanlılık, RMSE ve ortalama mutlak hata (OMH) değerlerinin 25 tekrara ait ortalamaları hesaplanmıştır.

Araştırma sonuçlarına göre hem MC hem de PH simülasyon çalışmasında test etkililiği bakımından GOO ve GA yöntemlerinin AOOT'ye kıyasla daha iyi performans gösterdiği; AOOT'nin yanlılık, RMSE ve OMH bakımından diğer iki yönteme kıyasla daha başarılı çalıştığı; sınıflama kriterlerinin farksızlık bölgesi genişledikçe veya hata düzeyi değeri küçüldükçe OTU'nun azaldığı ve test etkililiğinin arttığı görülmüştür. Bununla birlikte sınıflama kriterlerinin tümünün her koşulda oldukça yüksek düzeyde sınıflama doğruluğuna sahip oldukları; gerçek ve kestirilen yetenekler arasındaki korelasyonlar bakımından BSD ve AOK yetenek kestirim yöntemlerinin her ikisinin de başarılı kestirimlerde bulundukları ancak yanlılık, RMSE ve OMH bakımından BSD'nin AOK'tan göreli olarak daha iyi performans sergilediği belirlenmiştir. İncelenen madde seçme yöntemlerinin ise tümünün birbirine benzer çalıştığı; ancak MFB-KY'nin tüm bağımlı değişkenler açısından tüm koşullarda daha iyi performans gösterdiği görülmüştür.

Anahtar sözcükler: Bireyselleştirilmiş bilgisayarlı sınıflama testleri, sınıflama kriterleri, yetenek kestirim yöntemleri, madde seçme yöntemleri, post hoc simülasyon

Danışman: Prof. Dr. Nuri DOĞAN, Hacettepe Üniversitesi, Eğitim Bilimleri Anabilim Dalı, Eğitimde Ölçme ve Değerlendirme Bilim Dalı

A COMPARISON OF COMPUTERIZED ADAPTIVE CLASSIFICATION TEST CRITERIA IN TERMS OF CLASSIFICATION ACCURACY AND TEST LENGTH

Ceylan GÜNDEĞER

ABSTRACT

Computerized Adaptive Classification Testing (CACT) aims to classify the persons with the highest classification accuracy using the least number of items according to one or more predefined cut-points. The efficiency of these classifications varies by item pools, classification criteria, item selection methods and ability estimation methods. According to this, in the CACT, forming of different patterns and identification of these patterns under Monte Carlo (MC) and Post Hoc (PH) simulations are important for real applications.

In this study, different classification criteria, various methods for item selection and ability estimation in the CACT, are compared using classification accuracy, test length and precision of measurement under the simulations of both MC and PH. In our research, as classification criteria, Sequential Probability Ratio Test (SPRT), Generalized Likelihood Ratio (GLR) and Confidence Interval (CI) methods; as ability estimation methods, Expected a Posteriori (EAP) and Weighted Likelihood Estimation (WLE) methods; and as item selection methods, Maximum Fisher Information (MFI) and Kullback-Leibler Information (KLI) methods on the basis of cut-point (CP) and estimated ability (EA) have been examined. For this aim, for the MC simulation, a pool of 500 items, which is based on 3 PLM and informs at the cut-point (theta=1,0) and around, has been generated; for the PH simulation, a real data set including 80 items has been used. In the MC simulation, individual abilities have been generated using normal distribution (N(0,1)) for 3000 individuals. In the PH simulation, the ability level of the 994 individuals in the data set have been estimated by EAP on the basis of 3 PLM. The item response patterns have been generated randomly in R software in the MC simulation, whereas, the real item response pattern has been used without any manipulation in PH simulation. In our study, 96 conditions have been investigated for the MC and the PH simulations. At the end of the CACT simulations, the mean values of Average Test Length (ATL), Average Classification Accuracy (ACA), correlation

between the real thetas and estimated thetas (r), bias, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for 25 replications have been calculated.

According to results of the study, it has been observed in both the MC and the PH simulation results that the GLR and the CI classification criteria perform better compared to the SPRT in terms of test efficiency, however the SPRT works better compared to the other two methods in terms of bias, RMSE and MAE. It has also been deduced that the ATL decreases and test efficiency increases as the indifference region of classification criteria expands or the error value decreases. In addition, it has been concluded that all classification criteria have considerably high level of the classification accuracy in all conditions; and both ability estimation methods, the EAP and the WLE, have successful estimation results in terms of the correlation between real and estimated thetas (r); wheras the EAP relatively performs better than the WLE in terms of the bias, RMSE and MAE. It has also been observed that, all of the item selection methods work similarly to each other however the MFI-EA performs better for all conditions in terms of all dependent variables.

Keywords: Computerized adaptive classification testing, classification criteria, ability estimating methods, item selection methods, post hoc simulation

Advisor: Prof. Dr. Nuri DOĞAN, Hacettepe University, Department of Educational Sciences, Division of Measurement and Evaluation in Education

İÇİNDEKİLER

TABLOLAR DİZİNİ

ŞEKİLLER DİZİNİ

SİMGELER VE KISALTMALAR DİZİNİ

BBT: Bireyselleştirilmiş Bilgisayarlı Test

BBST: Bireyselleştirilmiş Bilgisayarlı Sınıflama Testi

MTK: Madde Tepki Kuramı

AOOT: Ardışık Olasılık Oran Testi

FB: Farksızlık Bölgesi

GOO: Genelleştirilmiş Olabilirlik Oranı

GA: Güven Aralığı

BSD: Beklenen Sonsal Dağılım

AOK: Ağırlıklandırılmış Olabilirlik Kestirimi

MFB-KY: Kestirilen Yeteneği Temel Alan Maksimum Fisher Bilgisi

MFB-KN: Kesme Noktasını Temel Alan Maksimum Fisher Bilgisi

KLB-KY: Kestirilen Yeteneği Temel Alan Kullback-Leibler Bilgisi

KLB-KN: Kesme Noktasını Temel Alan Kullback-Leibler Bilgisi

BUT: Bireyselleştirilmiş Uzmanlık Testi

BKK: Bayesci Karar Kuramı

OTU: Ortalama Test Uzunluğu

OSD: Ortalama Sınıflama Doğruluğu

OMH: Ortalama Mutlaka Hata

1. GİRİŞ

Bu bölümde araştırmanın temelini oluşturan problem durumuna, amacı ve önemine, problem cümlesi ve alt problemlerine, sınırlılıklarına, tanımlara ve kuramsal temeline yer verilmiştir.

1.1. Problem Durumu

Bilgi ve iletişim teknolojilerinde yaşanan gelişmeler, bilgiye ulaşmada ve eğitim uygulamalarında sıklıkla kendini göstermektedir. Bu gelişmeler sayesinde öğrencilerin öğrenme sürecinde, yeteneklerinin-becerilerinin ölçülmesinde ve değerlendirilmesinde birçok değişiklik meydana gelmektedir. Özellikle bilgisayar destekli değerlendirme (Computer Assisted Assissment) ile ölçme ve değerlendirme süreci yenilenmekte ve güçlenmektedir. Bilgisayar destekli değerlendirme türlerinden günümüzde en popüleri ise Bireyselleştirilmiş Bilgisayarlı Test (BBT; Computerized Adaptive Testing: CAT) uygulamalarıdır.

Bilgisayarlı ve Bireyselleştirilmiş kelimelerinden oluşuyor olması BBT'nin iki temel özelliğine işaret etmektedir. Bunlardan biri testin bilgisayar ortamında uygulanıyor oluşu iken diğeri testin bireyselleştirilmiş olmasıdır. Bu iki özellikten ilki, öğrencinin bilgisayar ekranında gördüğü maddeyi klavye vb. yardımıyla cevaplaması anlamına gelirken; ikincisi, testin öğrencinin yetenek düzeyine göre ayarlanmış olmasına işaret etmektedir (McBride, 1985).

Geleneksel ölçme araçları incelendiğinde, testi alan tüm öğrencilerin testte yer alan tüm maddeleri cevapladığı görülmektedir. Bir başka deyişle çoktan seçmeli testler, doğru-yanlış testleri vb. geleneksel kâğıt-kalem testlerinde farklı yetenek düzeylerindeki öğrencilere aynı maddeler uygulanmaktadır. Öğrencilerin yetenek düzeylerindeki farklılık sebebiyle kâğıt-kalem testlerinde çok sayıda madde kullanılmaktadır. Çok sayıda madde kullanımı sebebiyle ölçmenin etkililiği ve kullanışlılığı azalmaktadır.

BBT'nin yapısı ve uygulanışı ise, geleneksel kâğıt-kalem testlerinin aksine, bilgisayara dayalı ve bireyselleştirilmiştir (Eggen, 2004). BBT uygulamaları ile, Madde Tepki Kuramı'nın (MTK) avantajları sayesinde öğrencilere yetenek düzeylerine uygun maddeler sunulabilmekte; öğrencinin aldığı test öğrencinin yetenek düzeyine göre ayarlanarak bireyselleştirilebilmektedir. Böylece BBT ile

geleneksel testlere kıyasla daha kısa zamanda, daha az sayıda maddeyle ve yüksek güvenirlik düzeyinde yetenek kestirimi elde edilebilmektedir (Wainer, 2000). Ayrıca BBT ile hızlı puanlama yapılabilmekte ve öğrenciler sınav sonucunu uygulama sonunda direkt öğrenebilmektedir. Bu nedenlerden dolayı özellikle yurt dışında, GRE (Graduate Record Examination), GMAT (Graduate Management Admission Test) gibi sınavlarda BBT'nin tercih edildiği görülmektedir.

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

Klasik BBT uygulamalarında, genellikle bireyin son yetenek kestiriminde en yüksek bilgiyi veren maddenin seçimi söz konusudur ve uygulama yeteneğin etkili bir biçimde kestirilmesiyle sonlanır. BBST uygulamalarında ise madde seçimi, bireyin kesme noktasının hangi tarafına düşeceği hakkında bilgiyi veren maddelerin seçimine ve bireyi sınıflama çabasının sonuç vermesine dayanır (Nydick, 2013). Kısaca BBST uygulamalarında temel amaç öğrencileri, belirlenen kesme noktasına göre daha az sayıda maddeyle etkili bir biçimde sınıflara ayırmaktır.

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

BBST uygulamalarında cevaplanması gereken ilk soru hangi modelin kullanılacağıdır. MTK kapsamında model, çoklu puanlanan maddelere dayanan Ardışık Tepki Modeli, Kısmi Kredi Modeli vb. olabildiği gibi ikili puanlanan maddeleri temel alan 1 PLM, 2 PLM veya 3 PLM olabilmektedir. 1 PLM'de tüm

maddelerin ayırt edicilik parametreleri eşitken, 2 PLM'de ayırt edicilikler değişebilmekte ve 3 PLM'de ise ayırt ediciliklerin yanında şans parametresi de modele dâhil olmaktadır.

Madde havuzu, BBT ve BBST uygulamalarının en önemli bileşenlerinden biridir. Maddelerin, MTK'ya dayalı uygun modelle havuza kalibre edilmesi BBT ve BBST'nin ilk aşamasıdır. Genellikle başarı testlerinde kullanılan madde havuzu çok zor ve çok kolay maddeler içerir. Madde güçlükleri ise tek biçimli (uniform) dağılıma sahiptir. Ölçüt referanslı testlerde ise madde havuzundaki maddelerin kesme noktası etrafında en yüksek bilgiyi verebilecek madde güçlük değerlerine sahip olması beklenir (Boyd, 2003). Flaugher'a (2000) göre madde havuzunun kalitesi ne kadar iyiyse bireyselleştirilmiş test algoritması da o kadar başarılı performans gösterecektir.

Başlama kuralı, BBST'nin nasıl bir maddeyle başlayacağına işaret etmektedir. Başlangıç maddesi veya maddeleri testlerde genellikle 0 (sıfır) yetenek düzeyindeki maddelerden seçilmektedir. Bunun yanında (-1,1) yetenek düzeyleri aralığı veya varsa bireylerin önceden kestirilmiş yetenek düzeyleri de kullanılabilmektedir.

Madde seçme yöntemleri incelendiğinde BBT'de, Maksimum Fisher Bilgisi (MFB; Maximum Fisher Information: MFI), Kullback-Leibler Bilgisi (KLB; Kullback-Leibler Information: KLI), a-tabakalama vb. birçok yöntemin tanımlanmış ve çalışılmış olduğu görülmektedir. Klasik BBT uygulamalarında madde seçilirken, bireyin kestirilen geçici (interim) yetenek düzeyinde en yüksek bilgiyi veren maddenin seçimi; BBST uygulamalarında ise bireyin kestirilen geçici yetenek düzeyinde ve bununla birlikte BBT'den farklı olarak kesme noktasında en yüksek bilgiyi veren maddenin seçimi söz konusudur.

BBT ve BBST uygulamalarında birçok yetenek kestirim yöntemi ele alınabilmektedir. Bunlardan en sık kullanılanları Maksimum Olabilirlik Kestirimi (MOK; Maximum Likelihood Estimation: MLE), Beklenen Sonsal Dağılım (BSD; Expected a Posteriori: EAP) ve Maksimum Sonsal Dağılım (MSD; Maximum a Posteriori: MAP) yöntemleridir. Bunların dışında Ağırlıklandırılmış Olabilirlik Kestirimi (AOK; Weighted Likelihood Estimation: WLE) ve Owen'ın Ardışık Bayesci Kestirim yöntemi nadiren de olsa çalışmalarda yer almıştır. Alanyazın

incelendiğinde BBST araştırmalarında yetenek kestirim yöntemlerinin pek çalışılmadığı; değişken uzunluklu bu testlerde yöntemlerin birbirlerine kıyasla nasıl performans gösterdiklerinin henüz fazla bilinmediği görülmektedir.

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

1.2. Araştırmanın Amacı ve Önemi:

Bu araştırma, özellikle son yıllarda önem kazanan BBT çalışmalarının bir alt grubu olan BBST uygulamasını içermektedir. Cheng ve Liou'ya (2000) göre başarılı bir BBT veya BBST uygulamasında i) yetenek kestirim yönteminin uygunluğu ve ii) madde seçme yönteminin etkililiği oldukça önemlidir. Bu çalışmada hem türetilmiş veriye dayanan Monte Carlo (MC) hem de gerçek veriye dayanan Post Hoc (PH) simülasyonu ile, alanyazında tanımlanan farklı sınıflama kriterlerinin yanında, Cheng ve Liou (2000) tarafından öneminin altı çizilen yetenek kestirim yöntemleri ve madde seçme yöntemleri ele alınmış; oluşturulan koşulların sınıflama doğruluğu, test uzunluğu ve ölçme kesinliği bakımından karşılaştırılması amaçlanmıştır. Bir başka deyişle bu çalışmada, iki kategorili maddelerden oluşan simülatif ve gerçek veri setleri üzerinden uygulanan BBST simülasyonu sonucunda sınıflama doğruluğunun, test uzunluğunun ve ölçme kesinliğinin farklı sınıflama kriterlerine, farklı yetenek kestirim yöntemlerine ve farklı madde seçme yöntemlerine göre nasıl değiştiğinin incelenmesi amaçlanmıştır.

Geçmişten günümüze ülkemizde yapılan çalışmalar incelendiğinde BBT uygulamasıyla kağıt-kalem testleri arasındaki ilişkinin incelendiği (Köklü, 1990; Kaptan, 1993; Yaşar, 1999; İşeri, 2002; Kalender, 2011; Bulut ve Kan, 2012;

Kezer, 2013) birçok çalışmanın olduğu görülmektedir. Özellikle Tıp alanında gerçek BBT uygulamasına yönelik madde havuzu oluşturma çabaları da (Öztuna, 2008; Kaskatı, 2011; Altuğ Koşan, 2013) dikkati çeken diğer bir durumdur. Bunların yanında alanyazında BBT uygulamasında yetenek kestirim yöntemlerinin (Kalender, 2011); farklı teste başlama, devam etme ve testi sonlandırma kurallarından oluşturulan koşulların (Gökçe, 2012); madde seçme yöntemlerinin (Sulak, 2013); sonlandırma kurallarının (Eroğlu, 2013); madde kullanım sıklığının (Boztunç Öztürk, 2014) incelendiği çalışmalar da yer almaktadır. Ancak yurt içi alanyazında BBT uygulamasının sınıflama amacıyla kullanılmasına; bir başka deyişle BBST uygulamasına dair herhangi bir araştırmaya rastlanmamıştır. Çalışmanın ülkemizde yapılacak ilk çalışma olması nedeniyle önemli olduğu; yurt içi alanyazına katkı sağlayacağı ve son zamanlarda ülkemizdeki büyük ölçekli sınavlarda yapılmak istenen değişiklikler düşünüldüğünde de uygulayıcılara - kurumlara fikir sağlayacağı düşünülmektedir.

Yurt dışında yapılan çalışmalar incelendiğinde ise 1980'lerden bugüne BBST ile ilgili oldukça fazla sayıda çalışmaya rastlanmaktadır. Çalışmalar incelendiğinde sadece sınıflama kriterlerinin karşılaştırılmış olduğu araştırmaların yanında (Kingsbury ve Weiss, 1980; Reckase, 1983; Spray ve Reckase, 1996; Jiao ve Lau, 2003; Thompson ve Ro, 2007; Wouda ve Eggen, 2009; Thompson, 2011; Nydick, Nozawa ve Zhu, 2012; Huebner, 2012; Nydick, 2013) sınıflama kriterlerinin madde seçme yöntemleriyle çaprazlanarak ele alındığı çalışmaların da olduğu göze çarpmaktadır (Spray ve Reckase, 1994; Lau ve Wang, 1998; Eggen, 1999; Lau ve Wang, 1999; Eggen ve Straetmans, 2000; Lin ve Spray, 2000; Thompson, 2007a, 2009). Alanyazın incelendiğinde tek boyutlu ve iki kategorili madde havuzu üzerinden tek kesme puanı bulunması, bir başka deyişle sınıf sayısının iki olması durumunda, farklı sınıflama kriterlerinin, farklı yetenek kestirimi yöntemleriyle ve farklı madde seçme yöntemleriyle çaprazlanarak sonuçların sınıflama doğruluğu, test uzunluğu ve ölçme kesinliği bakımından karşılaştırıldığı bir çalışmanın alanyazında yer almadığı görülmektedir. Bu açıdan çalışmanın uluslararası alanyazına da katkı getireceği düşünülmektedir.

Ayrıca bu çalışmada, MC simülasyonunun yanı sıra, gerçek bir veri setinin de ele alındığı PH simülasyon çalışmasına da yer verilmiştir. Araştırmanın birinci basamağı olan MC simülasyonunda veri setinin tamamının yazılımda türetilmesi

söz konusu iken; çalışmanın ikinci basamağı olan PH simülasyon çalışmasındaki veri seti gerçek bir uygulama sonucunu içermektedir. Veri setleriyle ilgili detaylı bilgiye yöntem kısmında yer verilmiş olsa da kısaca belirtmek gerekirse MC'de madde parametreleri, bireylerin yetenek parametreleri ve madde cevap örüntüsü yazılımda rasgele oluşturulmuş; PH'de ise bu parametreler ve madde cevap örüntüsü gerçek bir uygulama sonucundan elde edilmiştir. Böylece çalışmada oluşturulan tüm koşullar için MC simülasyon sonuçlarının yanında PH simülasyon sonuçları da incelenmiştir.

BBST uygulamalarının yukarıda bahsedilen altı bileşeni ve alanyazın dikkate alındığında, hem simülatif veri seti hem de gerçek bir veri seti üzerinden, farklı madde seçme yöntemleri, yetenek kestirim yöntemleri ve sınıflama kriterlerinin birbirleriyle çaprazlanarak çalışılması hem ulusal hem de uluslararası düzeyde önem arz etmektedir. Bununla birlikte BBST hakkında araştırmalar tasarlamanın ve yapmanın gerekli olduğu; bu araştırma sonuçlarının ölçme ve değerlendirme alanyazını açısından önemli sonuçlar doğurabileceği düşünülmektedir.

Bu çalışmanın temel amacı, iki kategorili maddelerden oluşan simülatif veri seti üzerinden Monte Carlo (MC) ve gerçek veri seti üzerinden Post-Hoc (PH) simülasyonunda, kesme noktasının bir (sınıf sayısının iki) olması durumunda BBST uygulamalarının sınıflama doğruluğunun, test uzunluğunun ve ölçme kesinliğinin, farklı sınıflama kriterlerine (ve bu kriterlerin farklı farksızlık bölgesi (FB) veya güven aralığı değerlerine), farklı yetenek kestirim yöntemlerine ve farklı madde seçme yöntemlerine göre nasıl değiştiğinin incelenmesidir. Problem durumunun çözümlenmesi için ele alınan veri setleri, sınıflama kriterleri, yetenek kestirim yöntemleri ve madde seçme yöntemleri aşağıda Tablo 1.1'de özetlenmiştir.

Tablo 1.1: Araştırma Kapsamında İncelenen Değişkenler

Veri Seti	*Sınıflama Kriterleri*	*Yetenek Kestirim Yöntemleri*	*Madde Seçim Yöntemleri*
Simülatif Veri Seti (Monte Carlo)	AOOT (FB: 0,05)	BSD	MFB-KY
Gerçek Veri Seti (Post-Hoc)	AOOT (FB: 0,10)	AOK	MFB-KN
	GOO (FB: 0,05)		KLB-KY
	GOO (FB: 0,10)		KLB-KN
	GA (%70)		
	GA (%90)		

Tablo 1.1'de görüldüğü üzere çalışmada sınıflama kriterlerinden AOOT, GOO ve GA; yetenek kestirim yöntemlerinden BSD ve AOK; madde seçme yöntemlerinden ise MFB (kestirilen yetenekte: KY ve kesme noktasında: KN) ve KLB (KY ve KN) ele alınmıştır. AOOT'de ve GOO'da ikişer farksızlık bölgesi düzeyi (FB: 0,05 ve 0,10); GA'da ise iki güven aralığı değeri (%70 ve %90) olmak üzere toplam altı sınıflama kriteri; MFB'ye ve KLB'ye ait ikişer olmak üzere toplam dört madde seçme yöntemi çaprazlanmıştır. Buna göre araştırmada toplam, ***2 veri seti x 6 sınıflama kriteri x 2 yetenek kestirim yöntemi x 4 madde seçme yöntemi = 96 adet koşul*** oluşturulmuştur.

1.3. Problem Cümlesi:

İki kategorili maddelerin kalibre edildiği tek boyutlu madde havuzu üzerinde yapılan Monte Carlo ve Post-Hoc çalışmalarında, kesme puanının bir (sınıf sayısının iki) olması durumunda, BBST uygulamalarının sınıflama doğruluğu, test uzunluğu ve ölçme kesinliği, sınıflama kriterlerine, yetenek kestirim yöntemlerine ve madde seçme yöntemlerine göre nasıl değişmektedir?

1.3.1. Alt Problemler:

Çalışma 1: Monte Carlo simülasyon çalışmasında:

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

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

3. Sınıflama kriterlerine göre ortalama test uzunluğu, ortalama sınıflama doğruluğu, korelasyon, yanlılık, RMSE ve OMH değerleri nasıl değişmektedir?

4. Madde seçme yöntemlerine göre ortalama test uzunluğu, ortalama sınıflama doğruluğu, korelasyon, yanlılık, RMSE ve OMH değerleri nasıl değişmektedir?

5. Yetenek kestirim yöntemlerine göre ortalama test uzunluğu, ortalama sınıflama doğruluğu, korelasyon, yanlılık, RMSE ve OMH değerleri nasıl değişmektedir?

Çalışma 2: Post Hoc simülasyon çalışmasında:

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

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

8. Sınıflama kriterlerine göre ortalama test uzunluğu, ortalama sınıflama doğruluğu, korelasyon, yanlılık, RMSE ve OMH değerleri nasıl değişmektedir?

9. Madde seçme yöntemlerine göre ortalama test uzunluğu, ortalama sınıflama doğruluğu, korelasyon, yanlılık, RMSE ve OMH değerleri nasıl değişmektedir?

10. Yetenek kestirim yöntemlerine göre ortalama test uzunluğu, ortalama sınıflama doğruluğu, korelasyon, yanlılık, RMSE ve OMH değerleri nasıl değişmektedir?

1.4. Sınırlılıklar:

Çalışmada,

1. Tek boyutlu MTK modellerinden ikili puanlamaya dayalı 3 PLM kullanılmıştır.
2. Tüm koşullardaki başlama kuralı $\theta = 0$ olarak belirlenmiştir.
3. İçerik dengeleme ve madde kullanım sıklığı dikkate alınmamıştır.

1.5. Araştırmanın Kuramsal Temeli

BBT, bireylere kağıt-kalem testi yerine bilgisayar ile uygulanan bir ölçme aracıdır. BBT'nin en önemli avantajlarından biri MTK'nın değişmezlik özelliğini kullanarak testi alan her birey için "iyi" ölçme sağlayacak algoritmayı oluşturmasıdır. Bir başka deyişle test, MTK sayesinde çok kolay ya da çok zor olmayan, bireyin yetenek düzeyine uygun maddelerle bireyin yetenek düzeyine göre ayarlanabilmektedir. Böylece test kağıt-kalem uygulamalarına kıyasla daha kısa zamanda sonlanabilmektedir (Embretson ve Reise, 2000).

BBT uygulaması altı önemli yapıdan oluşmaktadır. Bunlar (i) Cevaplama Modeli, (ii) Madde Havuzu, (iii) Başlama Kuralı, (iv) Madde Seçme Kuralı, (v) Yetenek Kestirim Yöntemi ve (vi) Sonlandırma Kuralı'dır (Weiss ve Kingsbury, 1984). Bunların yanında, özellikle uygulamada fayda sağlaması adına, içerik dengeleme (content balancing) ve madde kullanım sıklığı kontrolünün (item exposure control) de ek bileşenler olduğu söylenebilir. BBST'nin BBT'den farkı ve odak noktası, bu yapılardan sonuncusu olan sonlandırma kuralı yerine geçen sınıflama kriteridir. Aşağıda BBT'nin sınıflama içeren bir alt boyutu gibi düşünülebilen BBST hakkında detaylı kuramsal bilgiye yer verilmiştir.

1.5.1. Bireyselleştirilmiş Bilgisayarlı Sınıflama Testi Uygulamaları (BBST)

BBST uygulamaları, sınıflama hatalarını azaltarak olabilecek en az sayıda maddeyle öğrencileri gruplara ayırmayı amaçlayan etkili bir test yöntemidir (Thompson, 2007b). BBST'de öğrenciye uygulanan madde sayısının az olması önemlidir. Böylece testin uygulama zamanı kısalacak, test için az sayıda madde geliştirmek zorunda kalınacak ve güvenlik problemleri azalacaktır (Finkelman, 2008).

BBST, öğrencileri bilgisayar ile sınıflamayı amaçlayan değişken uzunluklu bilgisayar testleri anlamında da kullanılmaktadır. Değişken uzunluklu testlerde, belirlenen amaca ulaşılıncaya kadar her öğrenci farklı uzunlukta test alır. BBST'de test, öğrenciyi sınıflayabileceği noktada biter. BBT'de ise test belirlenen yetenek kestirimi güvenirliği elde edilince ya da önceden belirlenen sayıda madde alınınca sonlanmaktadır (Thompson, 2007b). Buna göre BBST değişken uzunluklu testlerden iken; BBT hem değişken hem de sabit uzunluklu test uygulaması olabilmektedir.

Thompson'a göre (2007b) BBST *(i)* Psikometrik model, *(ii)* Kalibre edilmiş madde havuzu, *(iii)* Başlama noktası, (iv) Madde seçme algoritması ve *(v)* Sonlandırma (Sınıflama) kriteri olmak üzere beş temel bileşenden oluşmaktadır. Madde kullanım sıklığı ve içerik dengeleme ise yine ele alınabilecek ek bileşenlerdir. BBST, BBT'den yeteneğin nokta kestirimi bakımından farklılaşmakta; BBST'de test öğrenci sınıflandığında sonlanmaktadır.

Diao ve Reckase'e göre (2009) ise herhangi bir bireyselleştirilmiş test için karar verilmesi gereken beş anahtar soru vardır: *i)* Hangi model kullanılacak? *ii)* İlk madde nasıl seçilecek? *iii)* Yetenek nasıl puanlanacak? *iv)* Sonraki madde nasıl seçilecek? *v)* Test nasıl sonlanacak? Bu anahtar sorulara göre bireyselleştirilmiş testlerde madde seçimi ve yetenek kestiriminin oldukça önemli olduğu söylenebilir. Bu iki çalışmaya dayanarak BBST bileşenleri aşağıda Tablo 1.2'deki gibi özetlenebilir.

Tablo 1.2: BBST Uygulamalarının Bileşenleri

BBST'nin Bileşenleri	***Uygulanabilen Seçenekler***
1. Psikometrik Model	Klasik Test Kuramı veya Madde Tepki Kuramı
2. Madde Havuzu	Sivri veya Basık
3. Başlama Noktası	θ = 0 veya ön bilgi
4. Madde Seçimi	Kestirim Temelli, Kesme Noktası Temelli, Global
5. Yetenek Kestirimi	MOK, AOK, BSD, MSD, Owen
6. Sonlandırma Kriteri	AOOT, GOO, GA, BUT, BKK

Tablo 1.2'ye göre BBST uygulamalarında psikometrik model olarak MTK veya Klasik Test Kuramı (KTK) temel alınabilmekte; madde havuzu sivri veya basık belirlenebilmekte; teste başlama kuralında tüm yetenek düzeyleri sıfır kabul edilebilmekte veya (varsa) bireylere ait ön bilgiler (örneğin bireylerin bir önceki testten aldıkları puanlar) kullanılabilmekte; kestirim temelli, kesme noktası temelli veya global (ortak) madde seçme yöntemleri ele alınabilmekte; MOK, AOK vb. yetenek kestirim yöntemleri kullanılabilmekte ve sonlandırma kriterleri ise AOOT, GOO, BUT vb. olabilmektedir.

Tablo 1.2.'den görüldüğü üzere BBST'nin bileşenlerinin yanında bu bileşenlerin her birinin de farklı opsiyonları bulunmaktadır. Bu nedenle BBST uygulamalarında farklı desenlerin oluşturulması, karşılaştırılması ve koşullara uygun desenin belirlenebilmesi mümkün ve önemlidir.

1.5.1.1. Psikometrik Model

BBST'de ilk adım, diğer BBST bileşenleri için temel oluşturacak psikometrik modelin seçimidir. BBST'de hem KTK hem de MTK kullanılabilmektedir. Bunların her ikisi de madde parametrelerinin kalibrasyonu için öğrenci örneklemine gereksinim duyar. MTK modelleri BBST uygulamasına oldukça uygundur (Wainer ve Mislevy, 2000). MTK'nın avantajı; maddelerin, öğrenci yeteneklerinin ve kesme puanının/puanlarının aynı ölçekte yer almasıdır. Bu durum, diğer bileşenler için de önemli bir özellik olarak karşımıza çıkmaktadır (Thompson, 2007b). Bu çalışmada sadece MTK temel alınacaktır. Bu nedenle aşağıda MTK hakkında detaylı bilgi verilmiştir.

1.5.1.1.1. Madde Tepki Kuramı (MTK)

BBT ve BBST uygulamalarında bireylerin madde havuzundan belirli yöntemlerle seçilen farklı maddeleri cevaplaması söz konusudur. Bu durum testi alan tüm bireylerin aynı maddeleri cevapladığı geleneksel kağıt-kalem testlerinde yaşanmamaktadır (Wainer, 2000). Geleneksel testlerde temele alınan KTK'da madde istatistikleri testin uygulandığı gruba ve öğrenci yeteneği de testteki maddelere bağlıdır. Ayrıca bireylerin tümü için ortalama bir hata elde edilmektedir. MTK'ya göre ise, KTK'nın aksine, madde istatistikleri gruptan bağımsız (madde parametrelerinin değişmezlik özelliği) ve öğrenci yetenekleri de maddelerden bağımsız kestirilebilmekte (yetenek parametrelerinin değişmezlik özelliği); her öğrencinin kestirilen yetenek düzeyine ilişkin kestirimin standart hatası hesaplanabilmektedir (Hambleton ve Swaminathan, 1985).

MTK, bireyin test performansı altında yatan gözlenemeyen yetenek düzeyi ile gözlenen performansı arasındaki ilişkiyi matematiksel modellerle belirleyen bir kuramdır. MTK'da, madde karakteristik eğrisi (MKE) yardımıyla belli bir yetenek düzeyindeki öğrencinin maddeyi doğru cevaplama olasılığı kestirilir. MKE, yetenek ölçeğindeki farklı noktalar için maddenin doğru cevaplanma olasılığını verir. Bu eğri bireyin örtük özelliği ile madde performansı arasında monoton artan bir fonksiyondur. Buna göre yüksek yetenek düzeyindeki bireyin maddeyi doğru cevaplama olasılığı, düşük yetenek düzeyindeki bireyin maddeyi doğru cevaplama olasılığından daha yüksektir (Hambleton ve Swaminathan, 1985).

MTK modelleri için, testin boyutluluk durumuna ve testteki maddelerin kategori sayısına göre farklı sınıflamalar yapılabilmektedir. Testin boyutluluk durumuna

göre tek boyutlu ve çok boyutlu MTK modelleri yer alırken; testteki maddelerin kategori sayısına göre ikili ve çoklu puanlanan veya süreklilik özelliği gösteren MTK modelleri söz konusudur. Bu çalışmada tek boyutlu ve ikili puanlanan maddeler kullanıldığından bu modellerin varsayımlarına ve modellerin matematiksel temeline yer verilmiştir.

1.5.1.1.2. MTK Modellerinin Varsayımları

MTK modellerinin varsayımları tek boyutluluk, yerel bağımsızlık ve testin hız testi olmaması şeklinde sıralanabilmektedir (Hambleton ve Swaminathan, 1985). Tek boyutluluk, maddelerin ölçtüğü ve bireylerin cevaplama performanslarının altında yatan tek bir örtük özellik olması anlamına gelmektedir. Bir başka deyişle madde cevapları arasındaki varyansın tek bir örtük özellik tarafından açıklanmasıdır. Elbette kişilik, motivasyon, test alma becerisi gibi faktörler de bireyin test performansında etkilidir. Ancak tek boyutlulukla kastedilen maddelerin ölçtüğü başat bir boyutun olmasıdır (Hambleton ve Swaminathan, 1985). Tek boyutluluğun sağlanamadığı durumlarda çok boyutlu MTK modellerinden yararlanılabilmektedir.

Yerel bağımsızlık ise farklı maddelere verilen öğrenci cevaplarının istatistiksel olarak birbirinden bağımsız olması olarak düşünülebilir. Bir öğrencinin bir maddedeki performansı, testteki diğer maddeleri cevaplamasını iyi ya da kötü etkilememelidir. Bir madde diğer bir maddenin cevabına ipucu olmamalıdır. Bir başka deyişle sabit bir yetenek düzeyinde iki maddenin birbirinden bağımsız olması gerekmektedir (Hambleton ve Swaminathan, 1985). Buna göre yetenek düzeyi sabitlendiğinde maddeler arası kovaryansın sıfır olması yerel bağımsızlık varsayımının sağlandığı şeklinde yorumlanabilir. Buna yerel bağımsızlığın zayıf formu da denilmektedir.

Ortak köklü maddeler içeren testlerde, birbirini temsil eden maddelerin yer aldığı kişilik testlerinde, performans değerlendirmelerinde ve hız testlerinde yerel bağımsızlık varsayımının sağlanamaması olasıdır (Embretson ve Reise, 2000). Bu gibi durumlarda, devreye ikinci bir boyut gireceğinden yerel bağımsızlıkla birlikte tek boyutluluk varsayımı da ihlal edilmiş olur.

Yerel bağımsızlık varsayımı çeşitli istatistiksel yöntemlerle test edilebilmektedir. McDonald'a (1967) göre tek boyutluluğun anlamlı bir açıklığı, yerel bağımsızlık ilkesini temel alır ve eğer benzer/aynı yetenek düzeyindeki öğrenciler için

maddeler arasındaki kovaryans sıfıra eşitse test tek boyutludur (Akt: Hambleton ve Swaminathan, 1985). Bir başka deyişle eğer bir test tek boyutluluk özelliği gösteriyorsa bu testte yer alan maddelerin yerel bağımsızlık özelliğine de sahip olduğu söylenebilir.

Testin hız testi olup olmaması cevaplanmamış madde sayısı, testi bitirememiş birey sayısının yanında paralel test uygulaması veya hızsızlık testi gibi yöntemlerle kontrol edilebilmektedir (Hambleton ve Swaminathan, 1985). Testin hız testi olması durumunda öğrencilerin maddeleri cevaplarken başka bir boyut olan zamandan ya da hızdan etkilenmesi söz konusu olacaktır. Bu da testin tek boyutluluk ve yerel bağımsızlık varsayımlarını ihlal etmek anlamına gelecektir. Buna göre öğrencilere maddeleri cevaplamada yeterli zaman verilmesi, testin hız testi olmadığı ve bu varsayımın karşılandığı anlamına gelebilir.

1.5.1.1.3. MTK Modelleri

Öğrenci cevapları az önce de belirtildiği üzere, çoktan seçmeli testlerde, doğru-yanlış maddelerinde, kısa cevaplı sorularda cevap doğruysa 1 ve yanlışsa 0 (sıfır) olmak üzere iki kategorili puanlanmaktadır. Bunun yanında Likert tipi ölçeklerde her madde için çeşitli düzeyler yer almakta ve maddeler öğrencinin cevabına göre çoklu puanlanabilmektedir (Hambleton ve Swaminathan, 1985). Bu çalışmada ikili puanlanan maddeler üzerinde çalışıldığından ikili puanlanan MTK modellerine yer verilmiştir. Bu noktada 1 Parametreli Lojistik Model (PLM), 2 PLM, 3 PLM ve 4 PLM olmak üzere dört farklı MTK modelinden bahsedilebilir. Bu modeller MKE'lerin formülasyonu bakımından birbirinden farklılık göstermektedir.

1.5.1.1.3.1. 1 Parametreli Lojistik Model

1 PLM'de tüm maddeler madde güçlükleri (b parametresi) bakımından farklılaşırken ayırıcılıklarının (a parametresi) birbirine/bir sabite eşit olması söz konusudur. 1 PLM'nin özel bir hali olan Rasch Model'de ise a parametresi tüm maddeler için 1 değerine sabitlenmektedir. 1 PLM için MKE'nin matematiksel formülü şu şekildedir (Hambleton ve Swaminathan, 1985):

$$Pi(\theta) = \frac{e^{D\bar{a}(\theta - bi)}}{1+e^{D\bar{a}(\theta - bi)}} \quad (1)$$

Formülde ā tüm maddeler için sabit bir değer olan a parametresi; D ise 1,7 değerindeki ölçekleme sabitidir. Modelde b parametresi (-2,2) veya (-3,3)

aralığında yer alır. Modelde b parametresi 0,50 olasılıkla maddeyi doğru cevaplayabilecek yetenek düzeyini göstermektedir. 1 PLM'de ayırıcılıklar tüm maddeler için aynı olduğundan MKE'ler çakışmaz. 1 PLM, az sayıda parametre içermesi sebebiyle çalışılması ve anlaşılması kolay olan bir modeldir. Eğer amaç sadece öğrenci yeteneklerini MTK'den yararlanarak oran ölçeğine yakın bir düzeyde kestirmekse 1 PLM önerilebilir.

1.5.1.1.3.2. 2 Parametreli Lojistik Model

2 PLM'de maddelerin b parametrelerinin yanında a parametreleri de değişkenlik göstermektedir. Bir başka deyişle maddeler örtük özellikle aynı derecede ilişkiye sahip değildir ve a parametresi de öğrencinin test performansını etkilemektedir. Örneğin, iki öğrencinin 20 maddelik bir testte farklı beş maddeyi doğru cevaplamış olduğu düşünülürse öğrenciler 1 PLM dikkate alındığında öğrenciler aynı yetenek düzeyinde iken; 2 PLM'ye göre ayırıcılık parametresi yüksek maddeleri doğru cevaplayan öğrenci diğerine kıyasla daha yüksek yetenek düzeyindedir. 2 PLM için MKE'nin formülü şu şekildedir (Hambleton ve Swaminathan, 1985):

$$Pi(\theta) = \frac{e^{Dai(\theta - bi)}}{1 + e^{Dai(\theta - bi)}} \quad (2)$$

2 PLM'de a parametreleri maddeler arası değişkenlik gösterdiğinden MKE'ler çakışabilmektedir. a parametresi (0,2) aralığında değer almaktadır ve madde ayırıcılığı arttıkça MKE sivrileşecektir. Bu modelde de b parametresi 0,50 olasılıkla maddeyi doğru cevaplayabilecek yetenek düzeyini göstermektedir.

1.5.1.1.3.3. 3 Parametreli Lojistik Model

2 PLM'de tahmin parametresi (c) yer almazken 3 PLM'de maddelerin a ve b parametrelerinin yanında c parametresi de bulunmaktadır. c parametresi aslında MKE'nin düşük asimptotudur. Bir başka deyişle c parametresi, en düşük yetenek düzeyinde maddenin doğru cevaplanma olasılığıdır. c parametresinin eklenmesiyle MKE aşağıdaki gibi formüle edilebilmektedir (Hambleton ve Swaminathan, 1985):

$$Pi(\theta) = ci + (1 - ci)\frac{e^{Dai(\theta - bi)}}{1 + e^{Dai(\theta - bi)}} \quad (3)$$

c parametresi için kriter olarak belli bir aralık tanımlanmamıştır ancak dört seçenekli bir maddenin tahminle doğru cevaplandırılma olasılığı 0,25'tir. Bu değer

dikkate alınarak c parametresi yorumlanabilmektedir. Modelde b parametresi 1 ve 2 PLM'den farklı olarak, 0,50 olasılıkla maddeyi doğru cevaplayabilecek yetenek düzeyine değil; c parametresi arttıkça bu düzeyin üstüne denk gelmektedir. 3 PLM tahmin faktörünün bulunduğu çoktan seçmeli testlerde sıklıkla tercih edilen bir modeldir. Ayrıca alanyazın incelendiğinde BBST çalışmalarında diğer modellere kıyasla 3 PLM'in tercih edildiği durumlarda test uzunluğunun kısaldığı ve sınıflama doğruluğunun arttığı görülmüştür (Reckase, 1983; Lau, 1996; Jiao ve Lau, 2003). Bu sebeplerden dolayı çalışmada 3 PLM temel alınmıştır.

1.5.1.1.3.4. 4 Parametreli Lojistik Model

Yüksek yetenek düzeyindeki öğrenciler maddeyi her zaman doğru cevaplayamayabilir. Öğrencileri bazen dikkatsizlikten bazen de testi hazırlayanın umduğundan daha fazla bilgiye sahip olmaları sebebiyle cevap anahtarındaki doğru seçeneği işaretleyememektedir. Bu durumu çözebilmek amacıyla McDonald (1967) ile Barton ve Lord (1981) yeni bir model önermiştir. 4 PLM için MKE'nin matematiksel formülü aşağıdaki gibidir (Akt: Hambleton ve Swaminathan, 1985):

$$Pi(\theta) = ci + (yi - ci)\frac{e^{Dai(\theta - bi)}}{1+e^{Dai(\theta - bi)}} \quad (4)$$

Modelin 3 PLM'den tek farkı 1'den düşük bir değer alan y parametresidir ancak modelin sadece teorik bir model olduğu söylenebilir.

1.5.1.1.4. Değişmezlik Özelliği

MTK'nın temel varsayımları olan tek boyutluluk ve yerel bağımsızlık test edildikten sonra uygun modelin seçimi için model varsayımlarının sınanması gerekmektedir. Modeller için tanımlanan MKE formülleri dikkate alındığında 1 PLM'de maddelerin ayırıcılık indekslerinin eşit olması ve 1 ve 2 PLM'de öğrenci cevaplarında tahmin/şans faktörünün olmaması gerekmektedir. Bu noktada modellerin kendilerine has özelliklerinin incelenmesi; üç modelle de parametrelerin kestirilmesi ve modellere ait log-olabilirlik (-2LL) değerlerinin ki-kare dağılımı yardımıyla karşılaştırılması önemlidir. Bunun için örneğin X2 = $-2LL_{1PLM}$ – ($-2LL_{2PLM}$) hesaplanır ve bu değer serbestlik derecesi modele eklenen parametre sayısı olmak üzere tablodan elde edilen kritik X^2 ile karşılaştırılır. Hesaplanan değer, kritik değerden büyükse verinin 2 PLM'ye daha uygun olduğu söylenebilir.

Uygun modelin seçimiyle modelden beklenen iki özellik olan madde ve birey parametrelerinin değişmezliği gündeme gelmektedir. MTK'nın KTK'dan belki de en üstün özelliği olan madde parametrelerinin gruptan ve birey parametrelerinin maddelerden bağımsızlığı, BBT ve BBST uygulamaları için önemli bir temel teşkil etmektedir. MTK sayesinde BBT'de (ve BBST'de) bireyin yeteneği cevaplamış olduğu maddelerden bağımsız kestirilebilmektedir. Böylece bireyleri, bireyler farklı maddeleri cevaplamış olsalar bile karşılaştırabilmek mümkün olmaktadır (Hambleton ve Swaminathan, 1985).

1.5.1.2. Kalibre Edilmiş Madde Havuzu

BBST uygulamalarının ikinci bileşeni kalibre edilmiş madde havuzudur. Kullanılan madde havuzunun özellikleri diğer bileşenler tarafından belirlenmektedir. Kalibrasyon süreci seçilen psikometrik modele ve madde havuzunun yapısına bağlıdır. İstenen madde istatistiklerinin ranjı ise madde seçme algoritmasına bağlıdır. Örneğin madde seçimi, kesme puanına yakın güçlükteki maddenin seçimine dayanıyorsa bu bölgede güçlük değerine sahip birçok maddeye gereksinim duyulacaktır. Eğer algoritma öğrencinin kestirilen yetenek düzeyinde uygun güçlükteki maddenin seçimiyse güçlük parametresi ranjı yetenek ölçeği boyunca geniş olmalıdır. Ancak ne yazık ki madde havuzu hakkında detaylı bilgi, özellikle de sivrilik-basıklık vb. özellikler çalışmalarda raporlanmamaktadır (Thompson, 2007b). Bu çalışmada hem kesme noktasında (KN) hem de kestirilen yetenekte (KY) maddenin seçimine dayanan MFB ve KLB yöntemleri ele alındığından madde havuzu hem kesme noktasında yüksek bilgi veren hem de yetenek ölçeğini kapsayacak şekilde madde güçlüğü içeren maddelerden oluşturulmuştur.

Madde havuzu geliştirmede en önemli soru ne kadar maddenin gerekli olduğudur. Bu sorunun cevabı ise birçok duruma bağlıdır. Eğer test önemli sonuçları olan bir sınavsa (high stakes) ve sınıflama hatasının sadece küçük bir değeri tolere edilebiliyorsa, diğer testlere kıyasla daha çok madde gerektirmektedir. Eğer psikometrik model olarak MTK kullanılacaksa, yüksek ayırıcılığa sahip ve yüksek bilgi sağlayan maddeler gerekecektir (Thompson, 2007b). Bu çalışmanın birinci aşamasındaki madde havuzu MTK'ya dayalı 3 PLM temel alınarak türetilmiş; ikinci aşamasındaki gerçek veri setine ait madde parametreleri uygun modelin

belirlenmesi sonucu yine 3 PLM ile kestirilmiş ve simülasyon çalışmaları 3 PLM temelinde gerçekleştirilmiştir.

1.5.1.3. Başlama Noktası

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

1.5.1.4. Madde Seçimi

BBST'de ilk madde seçimi başlama noktasına bağlı olduğu gibi, sonraki madde seçimleri bireyin geçici yetenek düzeylerine göre ayarlanmaktadır. Alanyazında birçok madde seçme yöntemi betimlenmiştir (Eggen, 1999; Thompson, 2009). Madde seçme yöntemi, testin etkililiğini (madde sayısını) ve doğruluğunu belirler (Thompson, 2009). Örneğin Spray ve Reckase (1994) çalışmasında kesme noktasında en çok bilgiyi veren madde seçme yöntemiyle daha kısa test oluştuğunu gösterirken; Thompson (2007, 2009) araştırmalarında tam aksini, geçici yetenek düzeyinde en yüksek bilgi veren maddenin seçilmesi durumunda testin kısaldığını söylemektedir. Buna dayanarak madde seçiminin testin uzunluğu ve doğruluğu bakımından oldukça önemli olduğu düşünülebilir.

BBST için en basit madde seçim algoritması tesadüfi madde seçimidir (Kingsbury & Weiss, 1983; Akt: Thompson, 2007b). Testin her noktasında havuzdan madde rasgele olarak seçilir. Ne yazık ki bu yöntem, ne maddeler ne de öğrenciler hakkında bilgi kullanmamaktadır ve bu nedenle de etkili değildir. Daha uygun bir yaklaşım, zeki madde seçimi (intelligent item selection) yaklaşımıdır. Yaklaşımda bilgisayar uygulanmamış maddeleri değerlendirir ve "en iyi" maddeyi belirler. Bu noktada "en iyi"yi belirleyen birçok yöntem söz konusudur. Zeki madde seçim yöntemleri genel olarak iki kategori altında sınıflandırılabilmektedir: Kesme puanı temelli ve kestirim temelli (Thompson, 2007b). Kesme puanı temelli yöntemlerde kesme noktasında en yüksek bilgiyi sağlayan madde seçilirken; kestirim temelli yöntemlerde kesme puanı dikkate alınmaksızın bireyin kestirilen geçici yetenek düzeyinde maksimum bilgiyi veren madde seçilmektedir.

Psikometrik model KTK olarak belirlenmiş ise üç kesme puanı temelli yöntemden söz edilebilmektedir (Rudner, 2002; Akt: Thompson, 2007b). Bunlar maksimum ayırıcılık, bilgi kazanımı ve minimum beklenen değer yöntemleridir. Psikometrik model MTK olarak seçilmiş ise de üç kesme puanı temelli yöntem söz konusudur (Lin ve Spray, 2000). Bunlar Maksimum Fisher Bilgisi (MFB), Kullback Leibler Bilgisi (KLB) ve log-odds ratio yöntemleridir. MFB, doğru cevaplama olasılığı P ve yanlış cevaplama olasılığı Q olmak üzere, tek bir noktada bilginin maksimize edilmesini sağlar (Embretson ve Reise, 2000):

$$I_i(\theta) = (\partial P_i(\theta) | \partial\theta)^2 \; / \; P_i(\theta)Q_i(\theta) \tag{5}$$

KLB ise kesme puanı çevresindeki θ0'dan θ1'e kadar olan bölgedeki bilgiyi değerlendirir (Eggen, 1999):

$$K_i(\theta_1 \parallel \theta_0) = P_i(\theta_1)\log\frac{P_i(\theta_1)}{P_i(\theta_0)} + Q_i(\theta_1)\log\frac{Q_i(\theta_1)}{Q_i(\theta_0)} \tag{6}$$

Lin ve Spray'e (2000) göre kesme puanında yüksek bilgi veren maddeyle kesme puanı çevresindeki bölgede yüksek bilgi veren madde benzerdir ve bu nedenle de karşılaştırılabilirdir.

MFB ve KLB kestirim temelli madde seçiminde de kullanılabilmektedir (Reckase, 1983; Spray ve Reckase, 1994; Eggen, 1999). İki yöntemin eşitlikleri de kesme puanı temelli eşitliklerle aynıdır ancak kesme puanı yerine kestirilen geçici yetenek düzeyleri hesaplamada dikkate alınmaktadır. İki uygulama da bir öncekine benzer şekilde yakın ilişki göstermektedir: Kestirilen yetenek düzeyinde maksimum bilgi sağlayan maddeyle kestirilen yetenek düzeyinin çevresindeki bölgede maksimum bilgi sağlayan maddenin seçimi benzerdir. Kestirim temeli madde seçim yöntemleri bireyselleştirilmiş madde seçimleri olarak düşünülebilir çünkü bireysel olarak öğrencinin kestirilen yetenek düzeyini dikkate almakta ve öğrenci cevabının bir vektörü olmaktadır. Böylece bireysel olarak öğrencinin yetenek düzeyine uygun madde seçilmektedir.

Ayrıca ortak (mutual) bilgi madde seçim yöntemi de madde bilgisini yetenek düzeyi ranjı arasında değerlendirmektedir. Bu yöntem yetenek düzeyinin geniş bir ranja sahip olduğu ve öğrenci yetenekleri hakkında bilgi sahibi olunmadığı test başlama durumlarında veya birden fazla kesme puanı kullanılması durumunda oldukça kullanışlıdır. Bu çalışmada tek bir kesme noktası belirlendiğinden madde seçme

yöntemlerinden MFB-kestirilen yetenek yöntemi (MFB-KY), MFB-kesme noktası yöntemi (MFB-KN), KLB-kestirilen yetenek yöntemi (KLB-KY) ve KLB-kesme noktası (KLB-KN) yöntemleri ele alınmıştır.

1.5.1.5. Yeteneğin Kestirilmesi

BBST'de, madde havuzunun özellikleri, madde seçme yöntemi, testi sonlandırma kriteri ve yetenek kestirimi performansı etkileyen faktörlerdir. Bunlardan özellikle yetenek kestirimi, son sınıflama kararlarının etkililiği ve uygunluğu bakımından oldukça önemli bir değişkendir (Yang, Poggio ve Glasnapp, 2006). Bu değişken, raporlanan son yetenek kestirimini etkilediği gibi madde seçimi ve test sonlanmasını da etkilemektedir (Wang & Wang, 2001). BBST'de madde seçme yöntemi gibi yetenek kestirim yönteminin de testin etkililiğini ve doğruluğunu belirleyen önemli faktörlerden olduğu söylenebilir. Özellikle bireyin kestirilen geçici yetenek düzeyine göre madde seçilmesi durumunda, hem yetenek kestirimi hem de madde seçimi önem kazanmaktadır. BBST uygulamasının aşamaları düşünüldüğünde, yetenek kestiriminin doğruluğu arttıkça uygun maddenin seçimi kolaylaşacak ve test daha az sayıda maddeyle daha doğru bir sınıflama yaparak sonlanacaktır. BBST'nin amacı da tam olarak az sayıda maddeyle yüksek doğrulukta sınıflama yapmaktır (Thompson, 2009). Buna göre yetenek kestirim yöntemlerinin performansının BBST için oldukça önemli olduğu söylenebilir.

Alanyazında ikili kodlanan tek boyutlu MTK modellerine dayanan birçok yetenek kestirim yöntemi tanımlanmıştır. Bunlardan en sık kullanılanları Maksimum Olabilirlik Kestirim yöntemi (MOK; Maximum Likelihood Estimation: MLE; Birnbaum, 1968), Ağırlıklandırılmış Olabilirlik Kestirim yöntemi (AOK; Weighted Likelihood Estimation: WLE; Warm, 1989), Marjinal Maksimum Olabilirlik Kestirim yöntemi (MMOK; Marginal Maximum Likelihood Estimation: MMLE; Bock ve Aitkin, 1981) ve Bayesci yöntemlerden Beklenen Sonsal Dağılım yöntemi (BSD; Expected a Posteriori: EAP; Bock ve Aitkin, 1981), Maksimum Sonsal Dağılım yöntemi (MSD; Maximum a Posteriori: MAP; Samejima, 1969) ile Owen'ın Ardışık Bayesci kestirim yöntemidir (Owen, 1975). Bu kestirim yöntemlerinin tümü bir düzeye kadar yanlı kestirimler yapmaktadır (Warm, 1989).

Günümüze kadar yapılan BBT araştırmaları, bazı sebeplerden dolayı yetenek kestirim yöntemlerinden hangisinin daha iyi çalıştığına dair tamamlanmış bir resim

sunamamaktadır. Bu sebeplerden ilki, çalışmalarda bir ya da iki yetenek kestirim yönteminin ele alınması ve farklı desenlerin kullanılmasıdır. Bu nedenle çalışmalar arasında sonuçların karşılaştırılması oldukça zorlaşmaktadır. İkinci sebep olarak araştırmaların sadece tek bir test uzunluğuyla sınırlı olması gösterilebilir. Üçüncü sebep az sayıda araştırmanın hem sabit uzunluklu hem de değişken uzunluklu testler içermesidir. Dördüncü ve son sebep ise madde havuzu özelliğinin yetenek kestirim yöntemleri üzerindeki etkisinin incelenmemiş oluşudur. Bu nedenlerden dolayı BBT (ve BBST) çalışmalarında ya da uygulamalarında, belirlenen amaca uygun olarak hangi yetenek kestirim yönteminin en iyi performansı göstereceği hakkında sınırlı sayıda bilgi yer almaktadır (Wang ve Vispoel, 1998, s.110).

MTK, yukarıda değişmezlik özelliği bölümünde de bahsedildiği gibi parametrelerin değişmezliği üzerine kuruludur. Warm'a göre (1989) kestirilen parametrelerin değişmezliğinin sağlanabilmesi parametrelerin yansız kestirimine bağlıdır ancak tüm yetenek kestirim yöntemleri bir dereceye kadar yanlıdır. Uygulamalarda istenen, kestirimin yansız olmasıdır. Kestirimin yanlılığı birçok konuda önem taşımaktadır. Bunlardan ilki kâğıt-kalem testleriyle BBT sonuçlarının karşılaştırılması durumunda yansız kestirimler yapılırsa testleri eşitleme ihtiyacının ortadan kalkacağıdır. İkincisi ise sertifika-lisans sınavlarında kesme puanının geçerliğinin sağlanmasıdır: Öğrencileri sıralamada bu kadar önemli olmayan yanlılık sistematik bir şekilde geçme puanının güvenirliğini-hassaslığını (precision), buna bağlı olarak da sınıflama kararlarının geçerliğini-doğruluğunu etkilemektedir (Wang ve Wang, 2001).

Yetenek kestirim yöntemlerinden belki de en sık kullanılan yöntem olan MOK, bireyin madde cevap örüntüsünün olabilirliğini maksimize eden yetenek düzeyini (theta) bulma sürecidir. Bir başka deyişle MOK'un, bireyin 1-0 kodlanan madde cevap örüntüsüne ve madde parametrelerine dayalı olarak bireyin yetenek ölçeğindeki en uygun yerini bulmaya odaklandığı söylenebilir. Bu süreç iteratif (yinelemeli) olduğu gibi olabilirlik fonksiyonunu temele alır. MOK, fazla sayıda maddeye sahip olunduğunda yansız ve etkili kestirim yapan bir yöntem olsa da tam doğru veya tam yanlış cevap örüntüsü durumlarında hesaplama yapamaması sebebiyle zayıf bir kestirim yöntemidir (Embretson ve Reise, 2000). Bireyin tüm maddeleri doğru (pozitif sonsuz) veya yanlış (negatif sonsuz) cevaplaması sınırsız bir yetenek kestirimi ortaya çıkarır ve bu durum MOK'un temel aldığı olabilirlik

eşitliğinin çözümünde sorun teşkil eder (Wang ve Vispoel, 1998, s.111). Bu noktada tüm cevapların doğru ya da yanlış olması durumunda da kestirim yapabilen Bayesci kestirim yöntemleri devreye girmektedir. Bu çalışmada incelenen Bayesci kestirim yönteminin BSD olması sebebiyle aşağıda sadece BSD hakkında bilgiye yer verilmesi uygun görülmüş; ardından MOK'un modifiye edilmiş bir hali olan AOK yöntemi açıklanmıştır.

1.5.1.5.1. Beklenen Sonsal Dağılım Kestirim Yöntemi

BSD yöntemi MOK'un aksine iteratif olmayan birikimli bir süreç içermekte ve buna dayanarak da daha hızlı yetenek kestirimi yapabilmektedir ki bu durum BBT ve BBST uygulamalarında istenen bir durumdur. Ayrıca yöntem, pozitif sonsuz veya negatif sonsuz cevap örüntülerinde ve çoklu puanlanan maddelerde de sonuç vermektedir (Embretson ve Reise, 2000). BSD'nin bu açılardan MOK'a üstünlük sağlaması, çalışmada tercih edilen yetenek kestirim yönteminin BSD olmasına neden olmuştur. Çalışmada BSD'nin seçilmesinin bir diğer nedeni de, Bayesci kestirim yöntemlerinden olan MSD'nin odağının sonsal dağılımın moduna dayanması ve bu sebeple BSD'ye kıyasla daha değişken sonuçlar vermesidir (Bock ve Mislevy, 1982).

BSD temelde sonsal (posterior) dağılımın ortalamasını ve varyansını bulmaya odaklanır ancak her aşamada normallik varsayımı gerektirmez (Wang ve Vispoel, 1998, s.113). Bu amaçla, sabit sayıdaki belirli yetenek düzeylerinde (Gauss-Hermite quadrature points) olasılık, yoğunluk veya ağırlıkların hesaplanması söz konusudur. BSD'nin önsel dağılıma dayanması sebebiyle önsel dağılımın tanımlanmasının doğruluğu önem arz etmektedir. Önsel dağılım ne kadar doğru tanımlanırsa BSD kestirimleri hatadan o kadar arınık olacaktır. Ayrıca madde sayısının az olması durumunda yöntem, yetenek düzeyleri kestirimlerinin ortalamasına doğru yanlılık göstermektedir. Buna göre madde sayısı arttıkça yanlılığın azalacağı söylenebilir ancak ne kadar sayıda maddenin yanlılığı ne düzeyde azaltacağı bilinmemektedir (Wainer ve Thissen, 1987; Akt: Embretson ve Reise, 2000, s.179).

Bu çalışmada kullanılan madde havuzlarının büyüklükleri düşünülerek BSD yetenek kestirim yöntemi hızlı, yansız ve etkili kestirim yapabilmesi sebebiyle tercih edilmiştir. Gu ve Reckase'e göre (2007) BSD, önsel dağılımın uygun

olmaması durumunda MOK'a kıyasla daha yanlı kestirim yapmaktadır (Gu ve Reckase, 2007; Akt: Gökçe, 2012, s.33). Ancak bu çalışma, simülatif veri üzerinden Monte Carlo ve gerçek veri üzerinden post-hoc simülasyonuna dayanmaktadır. Çalışmada önsel dağılımların bilinmesi sebebiyle BSD'nin MOK'a ve MSD'ye kıyasla daha uygun bir kestirim yöntemi olacağı düşünülmüştür.

1.5.1.5.2. Ağırlıklandırılmış Olabilirlik Kestirim Yöntemi

Lord'a göre (1983) MOK dışa doğru yanlı (outward) kestirim yapmakta ve bu yanlılık yeteneğin negatif değerlerinde pozitiflere kıyasla daha yüksek olmaktadır (Lord, 1983). AOK yöntemi MOK'un yanlılığını azaltmak amacıyla Warm (1989) tarafından geliştirilmiştir. MOK, yeteneğin olası değerlerinin ranjı üzerinden olabilirlik fonksiyonunu maksimize etmeye dayanmakta ve olabilirlik fonksiyonunun modunu temel alarak çalışmaktadır. Warm'a (1989) göre yetenek kestiriminde olabilirlik fonksiyonunun modunun değil ortalamasının dikkate alınması gerekmektedir. AOK'un odak noktası, yanlılığı azaltan ağırlıklandırma fonksiyonudur. Bu fonksiyon yetenek düzeyi ile madde parametrelerinin bir fonksiyonudur ve her test için özeldir (Warm, 1989).

Warm (1989) çalışmasında, MOK'un varyansının aksine, yöntemin normal dağılımın ortalama ve varyansının kestiriminin yansız olduğunu matematiksel olarak ispatlamıştır. AOK'un varyansı da MOK'un varyansına benzer şekilde asimptotik olarak normal dağılıma sahiptir. Warm'ın (1989) çalışmasında, sabit uzunluklu testlerde MOK ve MSD yöntemlerine kıyasla, yetenek ölçeğinin tüm ranjı boyunca kestirimlerin düşük ve sabit bir varyansa sahip olduğu görüldüğünden AOK'un daha yansız kestirimler verdiği söylenebilir. Ayrıca çalışmada AOK'un sabit uzunluklu testlerde MOK'a kıyasla daha az sayıda maddeyle testi sonlandırdığı görülmüştür ki bu da BBT veya BBST uygulamalarında istenen bir durumdur. Böylece test süresi kısaldığı gibi madde görünümü de azalmaktadır.

Nydick, Nozawa ve Zhu'ya (2012) göre birçok çalışmada yetenek kestirimi için MOK yöntemi kullanılsa da, AOK sınıflama literatüründe önemli bir yer edinmiştir (Wang, Hanson ve Lau, 1999; Eggen ve Straetmans, 2000; Wouda ve Eggen, 2009). Sabit uzunluklu testlerde AOK yönteminin diğer yetenek kestirim yöntemlerine kıyasla daha iyi performans sergilediği görülürken, değişken uzunluklu testlerde yanlılık değerinin diğer yöntemlere göre yüksek olması

sebebiyle AOK yöntemini kullanmanın sakıncalı olabileceği görülmüştür (Wang, Hanson ve Lau, 1999; Yi, Wang ve Ban, 2000). Ancak çalışmalar arasında tam bir tutarlılık olmaması, sabit uzunluklu testlerde AOK yansız bir kestirici iken değişken uzunluklu testlerde (özellikle sınıflama içeren BBST'lerde) yöntemin diğer yöntemlere kıyasla nasıl bir performans sergilediğinin alanyazında fazla incelenmemiş olması sebebiyle AOK bu araştırmanın değişkenlerinden biri olarak ele alınmıştır.

1.5.1.6. Sonlandırma Kriteri

BBST'nin beşinci bileşeninin odak noktasında üç farklı sınıflama kriteri yer almaktadır: MTK temelli Güven Aralığı (GA), Ardışık Olasılık Oran Testi (AOOT) ve Bayesci Karar Kuramı (BKK). Üç kriter de geleneksel kağıt-kalem testlerinden daha kısa madde kullanımı sağlamakta ve genellikle benzer sınıflama doğruluğu göstermektedir (Kingsbury ve Weiss, 1983; Rudner, 2002; Akt: Thompson, 2007b). Ancak bu kriterlerin uygunluğu da seçilen psikometrik modeller, yetenek kestirim yöntemi ve madde seçme yöntemine dayanmaktadır. Bu çalışmanın odak noktası GA, AOOT ve GOO sınıflama kriterleri olduğundan aşağıda bu üç kritere yer verilmiştir.

1.5.1.6.1. Ardışık Olasılık Oran Testi Sonlandırma Kriteri

AOOT, Wald (1947) tarafından askeri araçların kalite kontrolü için II. Dünya Savaşı zamanında geliştirilmiştir ve özünde iki basit hipotezden hangisinin daha doğru olduğuna karar vermeyi amaçlayan istatistiksel bir testtir. AOOT ilk kez Fergusson (1969) tarafından KTK kapsamında sınıflama testlerine uygulanmış; daha sonra Reckase (1983) tarafından ise MTK kapsamında kullanılmıştır. Ayrıca AOOT, Luecht (1996), Mead (2006), Zenisky, Hambleton ve Luecht (2010) tarafından BBT ve çok aşamalı testlere (multi-stage testing) uygulanmıştır (Akt: van Groen, Eggen ve Veldkamp, 2014).

AOOT'nin altındaki temel felsefe, iki alternatif hipotez altında gözlenen cevap dağılımının olabilirliğini belirleyerek iki hipotezden birinin doğruluğuna karar verilmesidir. Eğer hipotezlerden birinin olabilirliği diğerinden oldukça büyükse bu hipotez kabul edilirken; iki hipotezin olabilirlikleri benzerse öğrenci yeni bir madde alır ve süreç bu şekilde devam eder (Reckase, 1983). Bir testin öğrencileri bir kesme noktasıyla başarılı ve başarısız olmak üzere iki kategoriye ayırmayı

amaçladığını varsayarsak hipotezler, δ (indifference region: farksızlık bölgesi: FB) Ho'ın başarılı ve başarısız bölgesine koyulan sabit olmak üzere Eşitlik 7 ve 8'deki gibi kurulacaktır:

$H_0 : \theta i = \theta 0 - \delta$ (7)

$H_1 : \theta i = \theta 0 + \delta$ (8)

BBST'de herhangi bir sınıflama kriteri öğrencinin testi almaya devam edip etmeyeceğine ve etmeyecekse hangi sınıfa yerleştirileceğine karar verir. Bu noktada AOOT istatistiğini kritik değerlerle karşılaştırır. Belli cevap dağılımına sahip bir öğrenci (i) ve bir madde (j) için log-olabilirlik (Pij: 3 PLM olmak üzere):

$$\log [L(\theta|yi,J)] = \sum_{j=1}^{J} [\, y_{ij} \log [pj(\theta)] + (1 - y_{ij}) \log [1 - pj(\theta)] \quad (9)$$

θu = öğrencinin başarılı kategorisinde ve θ1 = öğrencinin başarısız kategorisinde bulunması olmak üzere, i öğrencisinin log-olabilirlik oranı:

$$C_{i,j} = \log [LR(\theta u,\theta 1|y_{i,j})] = \log \left[\frac{L(\theta u|yi,j)}{L(\theta 1|yi,j)}\right] = \log [L(\theta u|yi,j)] - \log [L(\theta 1|yi,j)] \quad (10)$$

Bu eşitlik büyük ve pozitif bir değerse öğrenci başarılı; negatifse öğrenci başarısız kategorisine atanacaktır. Olabilirlik yerine log-olabilirlik test istatistiğini, belirlenmiş I.-II. tür hata düzeylerinde kullanmada Wald (1947) aşağıdaki eşitliklerin kullanılmasını önermiştir:

$Cl = \log [A] = \log [\beta / (1 - \alpha)]$ (11)

$Cu = \log [B] = \log [(1 - \beta) / \alpha]$ (12)

Bu eşitliklerden ilki başarısız kategorisini; ikincisi ise başarılı kategorisini belirsizlikten ayıran kritik değerdir. Minimum ve maksimum madde sayılarının arasındaki her maddeden sonra Ci,j hesaplanır:

$C_{i,j} < Cl$ ise öğrenci başarısız olarak sınıflanır ve test sonlanır.

$C_{i,j} > Cu$ ise öğrenci başarılı olarak sınıflanır ve test sonlanır.

$C_l \leq C_{i,j} \leq C_u$ ise öğrenciye diğer madde uygulanır ve test devam eder.

j maksimum sayıya ulaştığında ise karar vermek için $[(C_l + C_u) / 2]$ kritik değeri hesaplanır (Finkelman, 2008). Araştırmacılar genellikle α = β belirler ve dolayısıyla $[(C_l + C_u) / 2] = 0$ hesaplanır; ancak uygulayıcılar bazen, yanlış sınıflamanın son hatalarına bağlı olarak bir tür hatadan kaçınmayı isteyebilmektedir.

Ne yazık ki AOOT'nin kolay bir hipotez testi olmasının yanında birçok dezavantajı vardır. Bunlardan biri, AOOT $\theta_i \neq \theta_l$ veya $\theta_i \neq \theta_u$ olduğunda diğer süreçlere göre daha etkisizdir (Finkelman, 2008). Bu noktada GOO, AOOT'nin modifikasyonu olması ve tüm hipotezi test edebilmesi açısından önemlidir (Bartroff, Finkelman ve Lai, 2008; Thompson, 2009). Bu çalışmada AOOT sınıflama kriteri için Nydick'in (2013) çalışması göz önünde bulundurularak tolere edilebilir hatalar için farksızlık bölgesi 0,05 ve 0,10 değerleri dikkate alınmıştır.

1.5.6.6.2. Genelleştirilmiş Olabilirlik Oranı Sonlandırma Kriteri

GOO, AOOT'nin modifiye edilmiş daha genel bir halidir. AOOT'de eşitlik olması durumuna dair hipotezler yer alırken GOO'da eşitlik olmaması durumu da dikkate alınmaktadır. Böylece AOOT'nin yukarıda bahsedilen dezavantajı ortadan kaldırılmış olmaktadır:

$$H_0 : \theta_i \leq \theta_l = \theta_0 - \delta \quad (13)$$

$$H_1 : \theta_i \geq \theta_u = \theta_0 + \delta \quad (14)$$

Eğer $\theta \leq \theta 0$ ise $\theta' = \theta 0 + \delta$ veya $\theta > \theta 0$ ya da $\theta jmax$ ise $\theta' = \theta jmax$ olmak üzere Bartroff ve diğerleri (2008) AOOT'den farklı olarak olabilirlik oranı tanımlanmaktadır:

$$G_{i,j} = \log\ [GLR(\theta_0\ |y_{i,j})] = \log\left[L(\hat{\theta}\ |y_{i,j})\right] - \log\ [L\left(\theta'|y_{i,j}\right)] \quad (15)$$

GOO'da da farklı hipotezlerle, farklı test istatistikleriyle ve farklı kritik değerlerle AOOT'ninkiyle aynı prosedür uygulanır. Test istatistiği ve kritik değerler için karmaşık bir yöntem öneren Bartroff ve diğerlerinin (2008) aksine Thompson (2009) GOO'nın AOOT ile aynı özelliklere sahip olması gerektiğini ancak sadece GOO'da θ_1 ve θ_2'nin değişebileceğini belirtmiştir:

$$\log\left[GLR\left(\theta_u, \theta_l|y_{i,j}\right)\right] = \underset{\theta 1 \geq \theta u}{sup}\left(\log\left[L(\theta 1|y_{i,j}\right]\right) - \underset{\theta 2 \leq \theta l}{sup}\left(\log\left[L(\theta 2|y_{i,j}\right]\right) \quad (16)$$

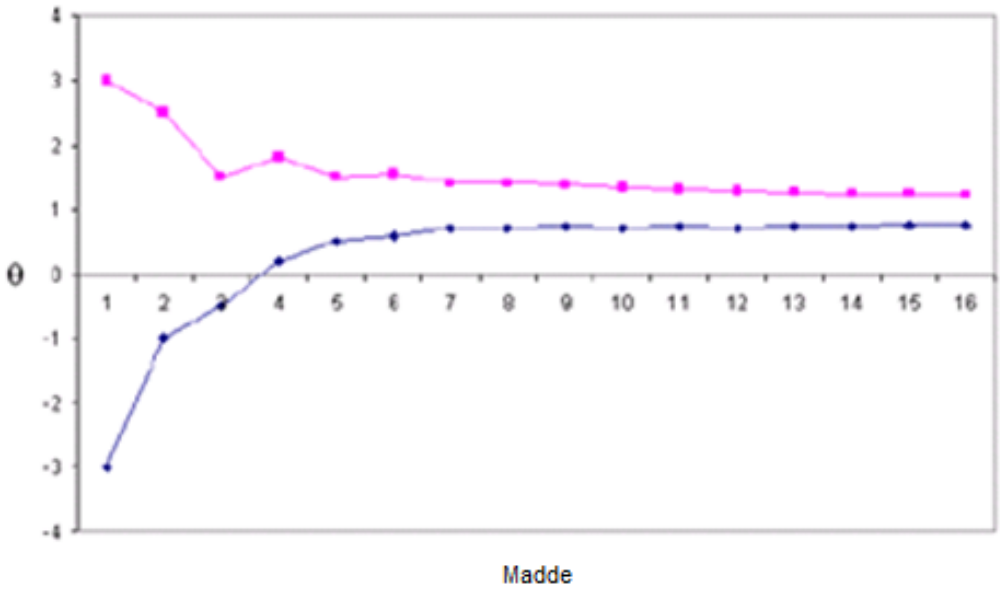
Bu eşitlikten elde edilen sonuç AOTT'deki C_l ve C_u ile karşılaştırılmaktadır. GOO ve AOOT'nin ikisi de olabilirlik oran testinin bazı versiyonlarını, sadece alınan maddelere dayalı olarak belirlenen kritik değerlerle karşılaştırmaktadır. Bu çalışmada GOO sonlandırma kriteri için AOOT'ninkilere benzer şekilde tolere edilebilir hatalar için farksızlık bölgesi 0,05 ve 0,10 değerleri dikkate alınmıştır.

1.5.1.6.3.Güven Aralığı Sonlandırma Kriteri

GA, sınıflama amacını istatistiksel bir kestirim problemi gibi formüle etmektedir (Eggen ve Straetmans, 2000). Test, j öğrencisi için kestirilen θ'yı elde edebilmek ve bu kestirimin belirlenen güven aralığına göre kesme puanının hangi tarafına düştüğünü belirleyebilmek amacıyla dizayn edilmektedir. Bu tanımlamanın hesaplanmasında da ölçmenin koşullu standart hatası (CSEM: Conditional Standard Error of Measurement) kullanılmaktadır (Thompson, 2007b):

$$\theta_j - z_{\alpha\,(CSEM)} \leq \theta_j \leq \theta_j + z_{\alpha\,(CSEM)}) \qquad (17)$$

Bu formülde zα, 1-α güven aralığına denk gelen normal sapmadır. Bu yaklaşımın basit bir örneği iki gruplu uzmanlık (mastery) testleridir. Bu testlerde her maddeden sonra güven aralığının kesme puanının altında mı üstünde mi olduğu değerlendirilmektedir. Eğer aralık tam olarak kesme puanının üstündeyse öğrenci geçti-başarılı şeklinde; aralık kesme puanının tam olarak altındaysa kaldı-başarısız olarak sınıflandırılabilmektedir. Eğer aralık kesme puanını içeriyorsa öğrenciye yeni bir madde sunulmaktadır. Aşağıdaki örnekte öğrenci yetenek düzeyi kesme puanı olan 0,75'in tamamen altına düşene kadar öğrenciye 16 madde uygulanmaktadır.



Şekil 1.1. Güven Aralığı Yöntemi Örneği (Thompson, 2007b)

Bu çalışmada GA sonlandırma kriteri için Eggen ve Straetmans'in (2000) çalışmalarında incelemiş olduğu %70 ve %90 güven aralığı değerleri ele alınmıştır.

2. İLGİLİ ARAŞTIRMALAR

Bu bölümde ilgili araştırmalar iki alt başlık halinde ele alınmıştır. İlk alt başlıkta, bu çalışmanın bağımsız değişkenlerinden olması sebebiyle, yetenek kestirim ve madde seçme yöntemleri ile ilgili yurt içi ve yurt dışı alanyazındaki araştırmalara yer verilmiştir. İkinci alt başlıkta ise, çalışmanın önemi bölümünde de belirtildiği üzere, bu araştırmanın Bireyselleştirilmiş Bilgisayarlı Sınıflama Testleri (BBST) hakkında ülkemizde yapılan ilk çalışma olması sebebiyle, sadece yurt dışı alanyazından çalışmalar yer almaktadır.

2.1. Yetenek Kestirimi ve Madde Seçme Yöntemleri ile İlgili Çalışmalar

Warm (1989) çalışmasında Maksimum Olabilirlik Kestirim yöntemi (MOK) ile benzer asimptotik varyansa ve normal dağılıma sahip ancak MOK'tan daha yansız olan Ağırlıklandırılmış Olabilirlik Kestirim yöntemini (AOK) önermiştir. Çalışmada geleneksel ve bireyselleştirilmiş testlerde AOK yöntemi, MOK ve Maksimum Sonsal Dağılım (MSD) ile karşılaştırılmıştır. Geleneksel kağıt-kalem testleri için 10, 20, 30, 40, 50, 60 madde olmak üzere 6 farklı test uzunluğu ve 2 farklı ayırt edicilik değeri (1,0 ve 2,0) çaprazlanmıştır. Bu noktada simülatif verinin geleneksel testleri yansıtması adına b parametreleri normal dağılımdan türetilmiş ve c parametreleri 0,20'ye sabitlenmiştir. Bireyselleştirilmiş testler için ise yine 6 farklı test uzunluğu ve 2 farklı ayırt edicilik değeri çaprazlanarak koşullar oluşturulmuştur. Yetenek düzeyleri, 17 yetenek düzeyinde 1000'er bireyden olacak şekilde 17.000 birey üzerinden türetilmiştir. Sonlandırma kuralı test uzunluğu olarak belirlenmiştir. Karşılaştırmalarda ortalama, standart sapma, ortalama hataların karekökü (OHK) hesaplanmıştır. Araştırmada tüm koşullarda benzer sonuçlar ortaya çıkmıştır. Ortalama hata bakımından en iyi yöntemin AOK olduğu; bunu MOK'un takip ettiği ve en yüksek ortalama hata veren yöntemin MSD olduğu sonucuna ulaşılmıştır. Buna karşılık standart hatalar bakımından MSD'nin daha iyi performans gösterdiği; OHK bakımından ise AOK'un diğer iki yöntemden daha düşük değerlere sahip olduğu görülmüştür. Ortalama kestirim hatası bakımından AOK ve MOK'un yansız kestirimler yaptığı ancak MSD'nin kestirimlerinin yanlı olduğu söylenebilir. Bireyselleştirilmiş testlerde test etkililiği (testin sonlanması için gereken madde sayısı) bakımından AOK ve MSD'nin benzer olduğu ve bunların MOK'un madde sayısının yarısı kadar madde gerektirdiği görülmüştür. Hesaplama zamanı

bakımından ise AOK'un diğer iki yönteme kıyasla uzun zaman gerektirdiği dikkat çekmiştir. Araştırma sonuçları AOK'un yetenek ölçeğinin geniş bir ranjındaki birçok koşulda MOK ve MSD'ye kıyasla daha iyi performans gösterdiğini göstermiştir.

Breslow ve Holubkov (1997) çalışmalarında AOK, MOK ve Yalancı Olabilirlik Kestirimi (YOK) yöntemlerinin iki kategorili veri setine uyumunu incelemişlerdir. Simülasyon çalışması sonucunda kestirimin standart hatası bakımından en düşük değeri MOK vermiş; bunu takiben AOK yöntemi gelmiş ve en yüksek standart hatayı ise YOK yöntemi göstermiştir. Çalışma sonucunda MOK'un diğer iki yönteme kıyasla, hesaplanan regresyon katsayısına ve ortalama standart hata değerine göre daha iyi performans gösterdiği sonucuna ulaşılmıştır.

Wang (1997) çalışmasında beta önsel dağılımını kullanan BSD (BÖ-BSD) kestirim yöntemi ile BSD ve MOK kestirim yöntemlerini karşılaştırmıştır. Araştırma sonuçlarına göre, MOK yansız kestirimlerinin yanında yüksek standart hata (SH) değerleri vermiştir. BÖ-BSD ise diğer iki yöntemden düşük yanlılık değeri vermiş ancak SH ve RMSE bakımından BSD'den yüksek değerler üretmiştir.

Wang ve Vispoel (1998) çalışmalarında, sabit ve değişken uzunluklu testlerde ve farklı madde havuzu özelliklerinde Owen'ın Bayesci Kestirim yöntemi (OBK), MOK, BSD ve Maksimum Sonsal Dağılım (MSD) yetenek kestirim yöntemlerini karşılaştırmışlardır. Çalışmanın bağımlı değişkenleri yanlılık, SH, RMSE, uyum ve test etkililiği olarak belirlenmiştir. Bu amaçla 17 yetenek düzeyinde, her bir yetenek düzeyinde 1000'er birey olacak şekilde toplam 17.000 yetenek parametresi türetilmiş; 300 ve 50 maddelik madde havuzlarıyla havuzların özelliklerini yansıtan maddelerin ayırt edicilik ve güçlük düzeyleri çaprazlanmıştır. Çalışma sonunda tüm madde havuzlarında RMSE'ler bakımından BSD yönteminin en düşük değere sahip olduğu ve en iyi performansı gösterdiği; MOK'un ise daha az yanlılık gösterse de daha yüksek SH ve RMSE, düşük uyum ve düşük uygulama etkililiği verdiği sonucuna ulaşılmıştır. Çalışmada Bayesci kestirim yöntemlerinden en kötü performans gösterenin OBK yöntemi; en iyi performans gösterenin ise BSD yöntemi olduğu görülmüştür.

Wang, Hanson ve Lau (1999) araştırmalarında 4 parametreli beta önsel dağılımı kullanan MSD yöntemi (BÖ-MSD) ile MSD, BSD, BÖ-BSD, AOK ve MOK kestirim yöntemlerini (i) 30 maddeli sabit uzunluklu testlerde (içerik dengeleme-madde

görünüm sıklığı ele alınmaksızın), (ii) sonsal varyansın sabitlendiği değişken uzunluklu testlerde ve (iii) içerik dengeleme ve madde kullanım sıklığının kontrol edildiği BBT'de, önsel dağılımın şekli, madde havuzu özellikleri, yanlılık, SH, RMSE bakımından karşılaştırmışlardır. Simülasyon çalışmasında 17 yetenek düzeyinde toplam 800 bireye ait yetenek parametresi türetilmiş; biri 420 maddeden oluşan, diğeri ise 420 maddelik havuzdan alınan 120 maddeden oluşan iki gerçek madde havuzu; biri ideal ayırt ediciliğe [N(1,1, 0,2)] diğeri ise ideal güçlüğe [N(0,1)] sahip iki simülatif madde havuzu ele alınmıştır. Simülasyonda tüm madde seçimleri Kingsbury ve Zara (1989) tarafından önerilen maksimum bilgi yöntemiyle yapılmıştır. Sonuçları değerlendirmede yanlılık, SH ve RMSE'ler dikkate alınmıştır. Sabit uzunluklu BBT uygulamasında RMSE'ler bakımından en düşük değere sahip olan yöntemlerin MSD ve AOK olduğu; değişken uzunluklu BBT uygulamasında ise RMSE'ler bakımından en düşük değere sahip olan yöntemlerin BÖ-MSD ve BÖ-BSD olduğu ortaya çıkmıştır. Bu çalışma sonucunda özellikle değişken uzunluklu testlerde AOK kullanılmasının sakıncalı olabileceği yorumu yapılmıştır.

Yi, Wang ve Ban (2000) çalışmalarında sabit uzunluklu ve değişken uzunluklu testlerde MOK, AOK, BSD ve MSD'yi yanlılık, SH, RMSE ve test etkililiği bakımından karşılaştırmıştır. Çalışmada madde seçme yöntemi olarak MFB kullanılmış; başlama noktası 0 olarak belirlenmiş; madde havuzu a, b, c parametrelerinin ortalamaları sırasıyla 0,97; 0,18 ve 0,15; standart sapmaları da 0,29; 0,97 ve 0,05 şeklinde belirlenmiştir. Bireylerin yetenek parametreleri -4,4 aralığı arasında 1000'er bireyden türetilmiştir. Sabit hata kuralında yanlılık bakımından AOK'un MOK'tan daha yüksek değerler verdiği ancak SE'ler bakımından daha iyi performans gösterdiği görülmüştür. En az sayıda maddeyle testi sonlandıran yöntemin MSD olduğu; bunu takiben sırasıyla BSD'nin ve AOK'un geldiği; en yüksek sayıda madde gerektiren yöntemin ise MOK olduğu sonucuna ulaşılmıştır.

Cheng ve Liou (2000) araştırmalarında iki farklı yetenek kestirimini (AOK ve MOK) dört farklı madde seçme yöntemiyle (Maksimum Fisher Bilgisi (MFB)-uygun madde güçlüğü; MFB-en yüksek bilgiyi veren madde; Kullback Leibler Bilgisi (KLB)-uygun madde güçlüğü; KLB-en yüksek bilgiyi veren madde) çaprazlayarak incelemişlerdir. Çalışmada gerçek veri seti kullanılmış ve koşullar 1000 tekrar sonucu ortalama yanlılık ve karekök farklılıkları bakımından karşılaştırılmıştır.

Araştırma sonunda ortalama yanlılık bakımından AOK'un en yansız kestirici olduğu; AOK ile KLB'nin birlikte ele alındığı tüm koşullarda gerçek yetenek düzeyinin daha erken (hızlı) kestirildiği; MOK ile en yüksek bilgi veren maddenin seçimine dayanan KLB ve MFB yöntemlerinin daha yansız kestirim yaptığı görülmüştür. Karekök farkları bakımından ise en yüksek bilgiyi veren madde seçiminin yöntem fark etmeksizin en düşük hatayı verdiği; AOK ile KLB'nin bir arada ele alındığı koşulların en az hataya sahip olduğu ancak en uzun hesaplama süresini gerektirdiği sonucuna ulaşılmıştır. Ayrıca AOK ile KLB ve MFB yöntemlerinin 10 maddeden sonra benzer çalıştığı görülmüştür.

Wang ve Wang (2001) çalışmalarında MOK, AOK, BSD ve MSD'yi çok kategorili madde havuzunda Ardışık Tepki Modeli (ATM; Graded Response Model: GRM) temelinde incelemiştir. Araştırmada sabit test uzunluğu ve sabit güvenirlik düzeyi (değişken test uzunluğu) sonlandırma kuralları ile üç farklı madde havuzu büyüklüğünü çaprazlayarak yetenek kestirim yöntemlerinin hangi koşulda daha iyi performans gösterdiği belirlemeye çalışılmıştır. Bağımlı değişken olarak yanlılık, SH ve RMSE'ler dikkate alınmış; bu değişkenlere varyans analizi (ANOVA) uygulanmış ve değişkenlerin etki büyüklükleri hesaplanmıştır. Çalışma sonucunda AOK'un test uzunluğu ve test güvenirliği bakımından BSD'ye, MSD'ye ve MOK'a kıyasla daha üstün olduğu; BSD'nin ise MSD'ye kıyasla daha iyi performans gösterdiği görülmüştür. Değişken uzunluklu testlerde özellikle beklenenin aksine AOK'un MOK'taki hatayı azaltmadığı bulgusu dikkat çekmiştir. AOK ve MOK'un Bayesci yöntemlere kıyasla yanlılık değerleri daha az olsa da SH'leri daha yüksek bulunmuştur. Bu iki yöntemden AOK'un SH'leri MOK'unkilerden daha düşüktür. Ayrıca bu iki yöntem Bayesci yöntemlere göre yeteneğin uç değerlerinde RMSE bakımından daha düşük değerler vermiş olsa da test etkililiği bakımından Bayesci yöntemlerin iki katı kadar maddede testi sonlandırabilmiştir. ANOVA sonuçlarında ise değişken uzunluklu testlerde yanlılık değişkenindeki toplam varyansın %74'ünün yetenek kestirim yöntemlerinden geldiği; sabit uzunluklu testlerde de toplam varyansın büyük bir kısmının yetenek kestirim yöntemlerinden geldiği; RMSE'lerdeki toplam varyansın küçük bir bölümünün ise havuz özelliklerinden geldiği ancak bu değişkenliğin manidar olmadığı bulunmuştur.

Penfield ve Bergeron (2005) çalışmalarında çoklu puanlanan maddelerden oluşan madde havuzunda ve sabit uzunluklu testlerde yetenek kestirim yöntemlerini

incelemişlerdir. Bu amaçla öncelikle AOK, genelleştirilmiş kısmi kredi modeline (GKKM; Generalized Partial Credit Model: GPCM) uyarlanmış; ardından AOK ile MOK ve BSD yöntemleri karşılaştırılmıştır. Çalışmada (-4,4) yetenek düzeyleri arasındaki 15 noktada 1000'er bireye ait yetenek parametresi türetilmiş ve madde sayısı 6, 12, 24 olan 3 farklı 5 kategorili madde havuzu oluşturulmuştur. Araştırmada yer parametreleri b1 = −1,5, b2 = −0,5, b3 = 0,5 ve b4 = 1,5 şeklinde; madde ayırt edicilikleri ise a = 0,4 (düşük), 1,0 (orta), 1,6 (yüksek) şeklinde 3 farklı koşulla yer almış; test uzunlukları ve ayırt edicilikler çaprazlanarak toplam 9 koşul elde edilmiştir. Karşılaştırmada yanlılık, ortalama hata ve RMSE'ler dikkate alınmıştır. Simülasyon sonuçlarına göre AOK yönteminin yanlılık ve ortalama hata bakımından MOK ve BSD yöntemlerinden daha iyi performans gösterdiği bulunmuştur. RMSE'ler bakımından ise en düşük hatanın BSD'ye ait olduğu; bunu takiben AOK yönteminin geldiği ve en yüksek hatanın MOK tarafından verildiği görülmüştür. Hatalar bakımından yöntemler arasındaki en güçlü farklılık test uzunluğunun 6 olması ve ayırt eidicliğin düşük olması koşullarında ortaya çıkmıştır.

Diao ve Reckase (2009) çalışmalarında çok boyutlu BBT (ÇB-BBT) uygulamalarında sabit uzunluklu testlerde madde seçimi ve yetenek kestirimi yöntemlerinin karşılaştırılmasını amaçlamıştır. Bu amaçla yetenek kestirim yöntemlerinden MOK, ara yetenek kestirimlerinde MSD ve son yetenek kestiriminde BSD olmak üzere Bayesci kestirim yöntemleri; madde seçim yöntemlerinden ise MFB ve KLB seçilmiştir. Karşılaştırmalar test uzunluğu, önsellerin kullanımı, ortalama yanlılık ve RMSE'ler üzerinden yapılmıştır. Araştırmada gerçek veri seti kullanılmıştır. Çalışma sonucunda madde seçme yöntemlerinin MOK altında benzer hataları verdiği; özellikle 20 maddeli kısa test koşullarında Bayesci yöntemlerin MOK'a kıyasla daha iyi performans gösterdiği; önsellerin üç koşulunda da tüm test uzunluklarında Bayesci kestirim yöntemlerin iyi çalıştığı; yetenek kestiriminin sabit ve uygun olduğu görülmüştür.

Kalender (2011) araştırmasında, farklı yetenek kestirim ve sonlandırma kurallarını dikkate alarak uyguladığı simülasyon çalışması sonuçlarını, ÖSS - Fen Bilgisi alt testinin kağıt-kalem formatı sonuçları ile karşılaştırmıştır. Çalışma iki aşamadan oluşmaktadır: İlk aşamada, ÖSS'nin BBT ve kağıt kalem formatlarından elde edilen yetenek kestirimlerini karşılaştırmak amacı ile post-hoc simülasyonu

uygulanmıştır. Bu noktada yetenek kestirim yöntemlerinden MOK ve BSD; sonlandırma kurallarından ise sabit test uzunluğu ve değişken test uzunluğunu (sabit hata) ele almıştır. Çalışmanın ikinci aşamasında BBT performansını simülasyon dışında bir ortamda gözlemlemek amacı ile test bir grup öğrenciye uygulanmıştır. Post-hoc simülasyon bulguları BBT uygulamasının ÖSS için BSD yöntemi ile 0,30 ya da daha yüksek standart hata eşik değeri ile uygulanabileceğini göstermiştir. İki formattan elde edilen yetenek kestirimleri arasındaki korelasyon 0,95 olarak bulunmuştur. BBT ile kullanılan soru sayısı ortalaması ise 18,4'tür. Gerçek bireylere uygulanan BBT ile kağıt kalem formatındaki ÖSS yetenek kestirimleri arasındaki korelasyon 0,74 olarak hesaplanmıştır. Gerçek bireylere uygulanan BBT ile bireylere sorulan soru sayısında yaklaşık %50 oranında düşüş sağlanmıştır. Sonuçlar BBT formatının ÖSS Fen Bilgisi alt testi için kağıt kalem testi ile karşılaştırıldığında, daha yüksek güvenilirliğe sahip yetenek kestirimlerini daha az sayıda soru ile sağlandığını göstermiştir.

Tao, Shi ve Chang (2012) çalışmalarında ikili ve çoklu puanlanan maddelerden oluşan karma testler için yeni bir yetenek kestirim yöntemi olarak madde ağırlıklı olabilirlik kestirim (MAOK) yöntemini sunmuşlardır. Ayrıca araştırmada MAOK'un yanlılık bakımından, MOK ve AOK yöntemleriyle karşılaştırılması da amaçlanmıştır. Çalışmada hem gerçek veri seti hem de simülatif veri kullanılmıştır. Analiz sonuçlarına göre MAOK'un diğer iki yönteme kıyasla daha yansız ve daha tutarlı bir kestirici olduğu sonucuna ulaşılmıştır. MAOK'un özellikle yeteneğin uç değerlerinde MOK'a kıyasla oldukça iyi performans gösterdiği görülmüştür. Yanlılık, RMSE ve mutlak hata değerleri bakımından ise en düşük değerlere MAOK yönteminin sahip olduğu; bunu takiben MOK'un geldiği ve en yüksek yanlılık, RMSE ve mutlak hata değerlerinin ise AOK yöntemi ile elde edildiği görülmüştür.

Bulut ve Kan (2012) çalışmalarında BBT'nin ALES'e uygunluğunu incelemişlerdir. Bu amaçla 2008 yılında uygulanmış sınavdan 10.000 kişi rasgele seçilmiş ve bu 10.000 kişinin yanıtlarından 3 PLM'e göre madde ve yetenek parametreleri kestirilmiştir. Mevcut sorulardan bir havuz oluşturularak BBT uygulaması simulatif olarak yapılmıştır. Yapılan post-hoc simülasyon sonrasında BBT uygulamasının BSD yöntemi ile 0,25, 0,30 ve 0,40 hata eşik değerleriyle yapılabileceği

saptanmıştır. Araştırmada BBT uygulamaları ile soru sayılarında %70'e varan azalma sağlandığı; geleneksel kâğıt kalem testi ile BBT uygulamaları arasında 0,93'ün üzerinde korelasyonlar elde edildiği; buna dayanarak uygun koşullar sağlandığında ALES'in BBT ortamında uygulanabileceği ortaya koyulmuştur.

Gökçe (2012) çalışmasında Monte Carlo simülasyonunda AOK ve BSD'yi; post-hoc simülasyonunda da AOK ve MOK'u farklı başlama ve sonlandırma kurallarıyla bir arada ele almıştır. Araştırmada ayrıca içerik dengeleme (content balancing) ve madde kullanım sıklığı (item exposure) da incelenmiştir. Çalışmada başlama kuralları madde güçlük düzeylerinden kolay ($-1,5 < b < -0,5$), orta ($-0,5 < b < 0,5$) ve zor ($0,5 < b < 1,5$) olacak şekilde ayarlanmış; sonlandırma kuralları ise sabit madde ve sabit güvenirlik düzeyi şeklinde belirlenmiştir. Araştırma sonuçlarına göre sonlandırma kuralı sabit madde sayısı ve başlama kuralı orta güçlükteki madde olduğunda AOK, BSD'ye kıyasla daha yansız kestirim yapmış ve gerçek ile kestirilen yetenekler arasında daha yüksek korelasyon vermiştir. Ayrıca sonlandırma kuralı sabit güvenirlik olduğunda da AOK BSD'den daha az sayıda maddeyle belirlenen güvenirlikte kestirim yapabilmiş ve yine gerçek ile kestirilen yetenekler arasında daha yüksek korelasyon vermiştir. Bununla birlikte AOK ile MOK'un karşılaştırıldığı post-hoc simülasyon çalışmasında da benzer şekilde AOK'un MOK'a kıyasla daha az sayıda maddeyle daha yansız kestirim yaptığı ve gerçek ile kestirilen yetenekler arasında daha yüksek korelasyon verdiği görülmüştür.

Sulak (2013) çalışmasında farklı madde seçme yöntemleri, yetenek kestirim yöntemleri ve test sonlandırma kurallarını karşılaştırılmıştır. Bu amaçla 250 maddelik bir madde havuzu ve 2000 bireye ait yetenek düzeyleri [N(0,1)] simüle edilmiştir. Araştırmanın koşulları, madde seçme yöntemlerinden MFB, KLB, a-tabakalama, olabilirlik ağırlıklı bilgi ölçütü (OABÖ), Ardışık maksimum bilgi oranı (AMBO); yetenek kestirim yöntemlerinden MOK ve BSD yöntemleri; testi sonlandırma kurallarından ise sabit 40 madde ile 0,2 ve 0,4'ten küçük standart hata ele alınarak oluşturulmuştur. Buna göre 5 madde seçme yöntemi, 2 yetenek kestirim yöntemi ve 3 sonlandırma kuralıyla toplam 30 koşul oluşturulmuştur. Elde edilen bulguların analizinde; sabit test uzunluğuna dayalı sonlandırma kuralında tahminin standart hatası (SH); sabit standart hataya dayalı sonlandırma kuralında ise ortalama madde sayısı kullanılmıştır. Ayrıca çalışmada madde seçme

yöntemlerinin madde kullanım sıklıkları da incelenmiştir. Sabit test uzunluğu sonlandırma kuralına göre yapılan karşılaştırmalarda, MOK yetenek kestirimi kullanıldığında elde edilen SH değerleri, BSD yetenek kestirimi kullanıldığında elde edilen SH değerlerinden daha yüksek bulunmuştur. Madde havuzu kullanımında ise en iyi sonuç a-tabakalama madde seçme yönteminden elde edilmiştir. Sonlandırma kuralının SH<0.2 olduğu koşullarda, MOK yetenek kestirimi kullanıldığında en düşük ve en yüksek madde sayısı ortalaması sırasıyla AMBO ve MFB madde seçme yönteminden; BSD yetenek kestirimi kullanıldığında ise, KLB ve OABÖ madde seçme yönteminden elde edilmiştir. Sonlandırma kuralının SH<0.4 olduğu koşullarda ise, MOK yetenek kestirimi kullanıldığında en düşük ve en yüksek madde sayısı ortalaması sırasıyla MFB ve KLB'den; BSD yetenek kestirimi kullanıldığında ise MFB ve a-tabakalama yönteminden elde edilmiştir. Sonlandırma kuralının SH<0.2 ve SH<0.4 olduğu koşullarda, bütün madde seçme yöntemlerinde, MOK yetenek kestirimi kullanıldığında elde edilen ortalama madde sayısı, BSD yetenek kestirimi kullanıldığında elde edilen ortalama madde sayısından daha yüksek bulunmuştur. Araştırma sonucunda BSD yetenek kestiriminin test uzunluğunu kısalttığı; bütün madde seçme yöntemlerinin madde havuzu kullanımına ilişkin iyi bir denge göstermediği ve yüksek a-parametresine sahip maddeleri daha çok kullandıkları sonucuna varılmıştır.

Eroğlu (2013) çalışmasında BBT uygulamasındaki farklı sonlandırma kurallarını ölçme kesinliği ve test uzunluğu bakımından karşılaştırmıştır. Çalışmada 250 ve 500 maddeden oluşan madde havuzlarında sonlandırma kurallarından sabit test uzunluğu, SH, SH ve en az madde sayısı, yetenekteki değişim, yetenekteki değişim ve en az madde sayısı; başlama kurallarından 0 ve -1,1 yetenek düzeyi aralığı; yetenek kestirim yöntemlerinden MOK ve BSD incelenmiştir. Birey parametreleri -3,3 aralığından 1000 birey üzerinden türetilmiş; simülasyonda 25 tekrar yeterli görülmüştür. Bağımlı değişken olarak RMSE, yanlılık, uyum ve test uzunlukları karşılaştırılmıştır. Sonuç olarak 20 madde sabit uzunluk ve 0,22 SH koşullarında RMSE ve yanlılığın düşük elde edildiği; uyumun fazla değişmediği; madde havuzunun artmasıyla ve BSD kullanılmasıyla RMSE ve yanlılık değerlerinin düştüğü; teste başlama kurallarının ise farklılık çıkarmadığı görülmüştür.

Kezer'in (2013) araştırmasının amacı, gerçek BBT uygulaması ile kağıt-kalem testinin yetenek kestirimi ve kullanılan madde sayısı bakımından; simülasyon çalışması ile de farklı başlama-sonlandırma kurallarının ve yetenek kestirim yöntemlerinin karşılaştırılmasıdır. Bu amaçla Ankara Üniversitesi Yabancı Diller Yüksek Okulu İngilizce Hazırlık Programı'ndaki öğrencilere 100 maddelik İngilizce Kelime Testi hem BBT hem de kağıt-kalem testi olarak uygulanmıştır. Gerçek BBT uygulaması için araştırmacı tarafından çevrimiçi ortam oluşturulmuştur. Gerçek BBT uygulaması ile kağıt-kalem testi karşılaştırmasında 67 öğrenciye ait veriden; simulatif veride ise gerçek BBT uygulamasını alan 994 bireyin verisinden yararlanılmıştır. Gerçek BBT uygulamasında madde kullanım sıklığı kontrol edilmiştir. Simülasyon çalışmasında ise BBT uygulamalarında kullanılan farklı başlangıç kuralları, yetenek kestirim yöntemleri ve sonlandırma kuralları çerçevesinde araştırma için 18 farklı koşul oluşturulmuştur. Testin başlangıcında bireylerin başlangıç yetenek puanları, bir durum için 0 (sıfır), diğer bir durum için ise daha önceden kestirilmiş yetenekler olarak alınmıştır. Yetenek kestirimi için MTK'da mevcut olan üç kestirim yöntemi de MOK, MSD ve BSD üç farklı durum olarak alınmıştır. BBT uygulamalarında sıklıkla kullanılan sonlandırma kurallarından standart hatanın 0,50'nin altında olması durumu, standart hatanın 0,30'un altında olması durumu ve sabit uzunluk durumu üç sonlandırma kuralı olarak araştırmada incelenmiştir. Araştırma sonuçlarına göre BBT uygulamasıyla kağıt-kalem testine göre uygulanan maddede ve zamanda tasarruf sağlanmış; iki uygulama arasında yetenek parametreleri bakımından manidar ve yüksek bir korelasyon olduğu görülmüş; simülasyon çalışmasına göre de farklı koşullar arasında manidar ve yüksek bir ilişki olduğu saptanmış; 18 koşuldan kestirilen yetenek parametreleriyle kağıt-kalem uygulaması arasında da manidar ve yüksek bir ilişki görülmüştür. Sonlandırma kurallarından standart hatanın 0,50'nin altında olmasıyla kestirilen yetenek parametreleri diğer durumlardan elde edilen parametrelerle en düşük ilişkiyi vermiştir. Farklı yetenek kestirim yöntemleri arasında da farklılık olmadığı ancak MOK'un beklenen bir şekilde daha fazla sayıda madde gerektirdiği sonucuna ulaşılmıştır.

2.2. Bireyselleştirilmiş Bilgisayarlı Sınıflama Testleri (BBST) ile İlgili Çalışmalar

Kingsburry ve Weiss (1980) çalışmalarında BBST uygulamasını Monte Carlo simülasyonuyla AOOT ve Bireyselleştirilmiş Uzmanlık Testi (BUT) sonlandırma kriterleri ile sabit uzunluklu geleneksel kağıt-kalem testini ortalama test uzunluğu, sınıflama doğruluğu ve hata türleri bakımından karşılaştırmışlardır. Bu amaçla madde parametrelerinin birbirlerinden farklılaştığı 100'er maddelik 4 madde havuzu oluşturulmuştur. Bunlar: i) a parametresi 1,0; b parametresi 0,0 ve c parametresi 0,2'ye sabitlenmiş şekilde tekbiçimli (uniform) dağılımdan oluşturulan madde havuzu; ii) b parametresi -2,5 ile 2,5 arasında 0,5 aralıklarla değişen madde havuzu; iii) b parametresinin değişkenliğinin yanında a parametresi 0,5 ile 2,0 arasında 0,5 aralıklarla değişen madde havuzu; iv) a ve b parametrelerinin yanında c parametresi 0,1, 0,2 ve 0,3 olacak şekilde değişen madde havuzudur. Her havuz için N(0,1) dağılımından 500 öğrenciye ait yetenek parametresi simüle edilmiştir. Geleneksel testin üç farklı sabit uzunluğu (10 madde, 25 madde ve 50 madde) madde havuzundan rasgele alınmıştır. AOOT ve BUT, üç farklı test uzunluğu ve dört farklı madde havuzunda simüle edilmiştir. Karşılaştırmalar öncelikle geleneksel test, AOOT ve BUT üzerinden yapılmış; bunun yanında karar vermede ortalama test uzunluğu ve ortalama doğru sınıflama oranları da incelenmiştir. Araştırma sonuçlarına göre tekbiçimli madde havuzunda en az sayıda madde ile sonlanan testin, üç test uzunluğu için de AOOT olduğu; onu takiben BUT'un geldiği ve en fazla sayıda maddeyle sonlanan testin ise geleneksel kağıt-kalem testinin olduğu söylenebilir. Diğer madde havuzlarında ise AOOT ile BUT'un birbirine benzer sonuçlar verdiği ve geleneksel testin diğer sonlandırma kriterlerine kıyasla daha fazla sayıda madde gerektirdiği sonucuna ulaşılmıştır. Araştırma sonuçları sınıflama kriterleri bakımından birbirine benzerlik gösterse de en az sayıda maddeyle, en geçerli sınıflama kararları ve en dengeli hata oranlarının AOOT sınıflama kriteriyle elde edildiği görülmüştür.

Reckase (1983) çalışmasında, AOOT'nin tek kesme puanı olması durumunda iki farklı MTK modeliyle gerçekleştirilen bireyselleştirilmiş testlere uygulanabilirliğini incelemiştir. Çalışmanın amaçları (i) Maddeler, havuzdan rasgele seçilmediğinde AOOT'nin nasıl çalıştığının belirlenmesi; (ii) Farklı değerlere sahip farksızlık bölgesi durumlarında sınıflama doğruluğu ve ortalama test uzunluğunun

değişiminin incelenmesi ve (iii) 3 PLM ile kalibre edilen havuzdan 1 PLM temel alınarak gerçekleştirilen BBST simülasyon sonuçlarına göre tahmin parametresinin sınıflama doğruluğu ve ortalama test uzunluğu üzerindeki etkisinin incelenmesidir. Çalışmada iki farklı madde havuzu kullanılmıştır: 72 maddelik kelime testi 1 PLM ve 3 PLM ile kalibre edilmiştir. Öğrenci yetenekleri ise -3,3 aralığında 0,25 aralıkla türetilmiştir. BBST, her yetenek düzeyinde random sayı jeneratörü için farklı seed numarası kullanılarak 25 tekrarla simüle edilmiştir. Çalışmada tüm yetenek kestirimlerinde MOK kullanılmış; madde seçiminde ise iki modele uygun olarak kestirilen yetenekte en yüksek bilgiyi veren madde seçilmiştir. Ayrıca c parametresinin sınıflama doğruluğu üzerindeki etkisinin belirlenebilmesi amacıyla, madde cevapları 3 PLM ile elde edilmiş ancak bireyselleştirilmiş test süreci ve olabilirlik oranı 1 PLM ile temel alınarak hesaplanmıştır. Çalışmada olabilirlik oranı hesaplamada üç farklı farksızlık bölgesi (FB)ele alınmıştır. Bunlar ±0,3; ±0,8 ve ±1,0'dir. Tüm durumlar için kesme noktası 0 alınmıştır. Çalışma sonuçlarına göre 1 PLM'de farksızlık bölgelerine göre sınıflama doğruluğu benzerdir ancak karar vermede gerekli ortalama madde sayısı farksızlık bölgesi daraldıkça artmaktadır. Testin sonlanabilmesi için FB: ±1,0 arasında iken ortalama 3-5 madde; FB: ±0,3 arasında iken ise ortalama 6-11 madde gerekmektedir. 1 PLM'de öğrencinin gerçek yetenek düzeyi kesme noktasında fazla yakın değilse çok az sayıda maddeyle sınıflama yapılabilmektedir ve farksızlık bölgesinin genişliği sınıflama doğruluğunu fazla etkilememektedir. 3 PLM'de ise 1 PLM'ye benzer şekilde farksızlık bölgesi fark etmeksizin sınıflama doğruluğu benzerdir ancak ±0,8 ve ±1,0 farksızlık bölgesi durumunda daha iyi hata olasılıkları elde edilmektedir. Ayrıca yine 1 PLM'ye benzer şekilde ±0,3 farksızlık bölgesi daha fazla sayıda madde gerektirirken, ±0,8 ve ±1,0 benzer sayıda ve diğerine kıyasla daha az madde gerektirmiştir. 1 PLM ile karşılaştırıldığında ise 3 PLM'den hesaplanan ortalama madde sayısının daha az olduğu görülmüştür. Analiz sonuçlarına göre c parametresinin sınıflama doğruluğu ve ortalama madde sayısı üzerinde önemli bir etkisi olduğuna işaret edilmiştir.

Spray ve Reckase (1994) çalışmalarında (i) kesme noktasında en yüksek bilgiyi veren; (ii) gerçek yetenek düzeyinde en yüksek bilgiyi veren ve (iii) kestirilen geçici yetenek düzeyinde en yüksek bilgiyi veren madde seçme yöntemlerini ortalama madde sayısı bakımından karşılaştırmışlardır. Bu amaçla gerçek veri setinden

madde havuzu oluşturulmuş; öğrenci yetenek düzeyleri (-3,3) ranjında 0,25 aralıkla türetilmiş ve her yetenek düzeyinde 1000 tekrarla simülasyon tamamlanmıştır. Çalışmada, sonuçları genelleyebilmek amacıyla, -0,5; 0,0; 0,1 olmak üzere üç farklı kesme noktası kullanılmıştır. Çalışmanın gerçek uygulamaya benzemesi amacıyla BBST uygulamasında maksimum test uzunluğu 50 maddeye ayarlanmıştır. Farksızlık bölgesi olarak 1,0 tanımlanmıştır. Araştırma sonuçlarına göre, kesme noktasında en yüksek bilgiyi veren maddenin seçilmesi durumunda karar vermede kullanılan madde sayısı, öğrenci yetenek düzeyinde maksimum bilgi veren madde seçme yöntemine göre daha azdır. Çalışma sonucuna göre kesme noktasında en yüksek bilgiyi veren maddenin seçilmesi durumunda, bireyin geçici veya gerçek yetenek düzeyinde en yüksek bilgi veren madde seçimine kıyasla BBST uygulamalarının kısalacağı söylenebilir.

Spray ve Reckase (1996) çalışmalarında AOOT ve BUT yöntemlerini, bireyleri iki kategoriye sınıflamada, madde sayısı ve sınıflama hataları bakımından karşılaştırmışlardır. Bu amaçla 200 maddelik madde havuzu ve -3,3 aralığında 0,25 aralıklarla 1000'er bireye ait yetenek parametresi türetilmiş; simülasyondaki tüm koşullarda MFB-Kesme Noktası (MFB-KN) kullanılmış; bireylerin alabileceği maksimum madde sayısı 50'ye ayarlanmıştır. Araştırma sonuçlarına göre genel olarak AOOT'nin BUT'a kıyasla daha az sayıda madde ve benzer düzeyde sınıflama hatası verdiği; AOOT ve BUT'un benzer sayıda maddeyle sınıflama yapabildiği koşulda AOOT ile daha düşük sınıflama hatası hesaplandığı görülmüştür.

Lau (1996) çalışmasında tek boyutluluk varsayımının sağlanamadığı veya ihlal edildiği durumlarda AOOT sınıflama kriteriyle hesaplanan geçti-kaldı kararlarının uygunluğunu incelemiştir. Bu amaçla araştırmada Monte Carlo çalışmasıyla türetilen iki boyutlu madde havuzu tek boyutlu MTK temelinde simüle edilmiştir. Araştırmanın bağımsız değişkenlerini tek boyutlu MTK modellerinden 1 PLM ve 3 PLM, yetenek kestirimleri arasındaki korelasyonlar, test uzunluğunun en fazla 50 madde ile sabitlenmesi ve kesme puanları; bağımlı değişkenlerini ise ortalama test uzunluğu ve ortalama sınıflama doğruluğu oluşturmaktadır. Çalışma sonuçlarına göre tek boyutluluk varsayımının ihlal edilmesi durumunda AOOT'nin 1 PLM ve 3 PLM ile birlikte kullanışlı sonuçlar verdiği ve modellerden 3 PLM'in 1 PLM'e kıyasla test etkililiği (ortalama test uzunluğu) bakımından daha iyi performans gösterdiği

görülmüştür. Bunların yanında test uzunluğunun en fazla 50 maddeye sabitlenmesinin sınıflama doğruluğu ve test uzunluğu üzerindeki etkisinin kullanılan MTK modeline bağlı olduğu söylenebilir. Son olarak iki boyutlu madde havuzu üzerinden tek boyutlu MTK modelleriyle uygulanan kesme puanı kestiriminde kesme puanlarının yetenek ölçeği üzerinde kullanılan modele göre farklılık gösterdiği; bir başka deyişle ölçeğin bazı kısımlarında yüksek kestirimler yapılırken bazı kısımlarında düşük kestirimler yapıldığı; bunların da sınıflama doğruluğunu düşürdüğü sonucuna ulaşılmıştır. Buna göre çalışmanın en önemli sonucu, AOOT'nin tek boyutluluk varsayımının ihlal edildiği koşullarda bile başarılı performans göstermesi ve 3 PLM'in 1 PLM'den daha başarılı sınıflama doğruluğu ve test uzunluğu vermesidir.

Spray, Abdel-Fattah, Huang ve Lau (1997) araştırmalarında, Lau'nun (1996) çalışmasına benzer olarak, AOOT'nin çok boyutlu madde havuzlarındaki ve çok boyutlu örtük özelliğe sahip ölçmelerdeki performansını incelemişlerdir. Bu amaçla her biri 30 maddeden oluşan 6 paralel matematik testinden oluşan iki boyutlu madde havuzu üzerinden tek boyutlu MTK temelinde post-hoc simülasyonlar gerçekleştirmişlerdir. Araştırma koşullarında bağımsız değişkenler, (i) herhangi bir sabitlemenin olmadığı durum, (ii) en fazla 60 ve 120 madde sabitlemesi ve (iii) madde sayısı sabitlemesi ile madde görünüm sıklığı kontrolü olarak ele alınmıştır. Çalışmanın bağımlı değişkenleri ise test uzunluğu ve sınıflama doğruluğudur. Çalışma sonucunda test uzunluğu arttıkça sınıflama doğruluğunun da beklenen şekilde arttığı; madde görünüm sıklığı kontrolünün düşük yetenek düzeyinde daha fazla etkiye sahip olduğu; madde havuzunun çok boyutlu olması durumunda bile AOOT sınıflama kriterinin sınıflama doğruluğu bakımından iyi performans göstermiş olduğu görülmüştür. Buna göre çalışmanın sonuçları Lau'nun (1996) çalışmasının sonucuyla örtüşmektedir. Kısaca tekrarlamak gerekirse, AOOT sınıflama kriterinin tek boyutluluk varsayımının ihlal edildiği durumlarda da başarılı performans gösterdiği söylenebilir.

Lau ve Wang (1998) çalışmalarında AOOT'nin çok kategorili maddelere uygulanmasını; iki kategorili ve çok kategorili maddeler ile karma testlerde AOOT'nin uygunluk ve etkililik bakımından performansının belirlenmesini amaçlamışlardır. Bu amaçla 247 iki kategorili ve 266 çok kategorili maddelerden oluşan gerçek madde havuzları üzerinden post-hoc simülasyon

gerçekleştirmişlerdir. Araştırmada ayrıca üç farklı madde seçme yöntemi (tesadüfi madde seçimi, kesme noktasında madde bilgisi ve tesadüfi madde seçiminin kombinasyonu ile kesme noktasında madde bilgisi) ele alınmıştır. Simülasyonda bireylerin en az 10, en çok 50 maddeyi cevaplaması sağlanmıştır. Araştırma sonucunda hem iki kategorili hem çok kategorili hem de karma maddelerden oluşan madde havuzlarında en az sayıda madde ve en düşük sınıflama hatasının kesme noktasında bilgi veren madde seçimiyle elde edildiği görülmüştür. Buna göre tüm madde havuzlarının birbiriyle tutarlı sonuçlar verdiğine ve AOOT'nin çok kategorili maddelerle ve karma testlerle de uyum gösterdiğine işaret edilmiştir.

Eggen (1999) çalışmasında MFB ve KLB'nin etkililiğini iki ve üç kategorili sınıflamalar üzerinden ortalama madde sayısı ve sınıflama doğruluğu bakımından karşılaştırmıştır. Bu amaçla gerçek veri setinden 250 maddelik madde havuzu oluşturulmuş ve yine gerçek veriden [N(0,29, 0,52)] bir adet tesadüfi yetenek düzeyi türetilmiştir. Bu yetenek düzeyinde simülasyonlar 5000 kere tekrarlanmıştır. Tüm koşullarda MFB-Kestirilen Yetenek (MFB-KY), MFB-Kesme Noktası (KN), KLB madde seçme yöntemleri incelenmiştir. Çalışmada ikili ve üçlü sınıflamada tüm yöntemlerin madde sayısı bakımından benzer; sınıflama doğrulu bakımından yüksek (ikili ve üçlü sınıflama için sırasıyla %95 ve %89 civarında) değerler verdiği görülmüştür. Çalışmada MFB-KN'nin MFB-KY'ye kıyasla daha iyi performans gösterdiği dikkat çeken bir bulgu olmuştur. Çalışma sonucuna göre KLB'nin MFB'ye benzer ve bazen MFB'den daha iyi performans gösterdiği görülmüştür.

Lau ve Wang (1999) çalışmalarında çok kategorili maddeler üzerinden AOOT'nin performansını KLB ve MFB olmak üzere iki madde seçme yöntemine, madde görünüm sıklık kontrol yöntemlerine, kesme noktalarına, madde havuzlarına ve farksızlık bölgesi değerlerine göre incelemişlerdir. Bu amaçla iki farklı gerçek veri seti üzerinden post-hoc simülasyon ile koşullar tekrarlanmış ve sonuçlar hesaplanmıştır. Çalışmada bireylerin en az 3 en fazla 30 madde alması sağlanmış; değerlendirme kriterleri ise ortalama test uzunluğu, sınıflama hatası ve madde görünüm sıklığı olarak belirlenmiştir. Araştırma sonucunda hatalar ve madde sayısı bakımından MFB ve KLB arasında farklılık olmadığı; madde kullanım sıklık kontrol yönteminin olmadığı koşulda daha az sayıda maddeyle daha yüksek doğrulukla sınıflama yapılabildiği; kesme noktalarına göre farklılıklar olduğu; daha fazla sayıda madde içeren havuz üzerinden yapılan çalışmada daha

az sayıda maddeyle daha yüksek doğrulukla sınıflama yapılabildiği; farksızlık bölgesi büyüdükçe hatanın arttığı ancak madde sayısının azaldığı görülmüştür.

Eggen ve Straetmans (2000) çalışmalarında GA ve AOOT sınıflama kriterleriyle öğrencileri gerçek veri seti üzerinden üç kategoriye sınıflamayı amaçlamışlardır. Bu amaçla tesadüfi madde seçme yöntemi; MFB-KY; MFB-KN; MFB-KY ve içerik dengeleme; MFB-KN ve içerik dengeleme; MFB-KY ve madde kullanım sıklık kontrolü; MFB-KN ve madde kullanım sıklık kontrolü; MFB-KY, içerik dengeleme ve madde kullanım sıklık kontrolü; MFB-KN, içerik dengeleme ve madde kullanım sıklık kontrolünün bir arada ele alındığı koşullar incelenmiştir. Kalibrasyonda 1198 öğrenciye uygulanan 250 madde için 2 PLM kullanılmıştır. Madde havuzundaki maddeler, uzman görüşleri doğrultusunda düzey 1 (kolay), düzey 2 (orta), düzey 3 (zor) şeklinde kodlanmış ve iki kesme noktası belirlenmeye çalışılmıştır. Düşük kesme puanı düzey 1'de kestirilen en yüksek yetenek değerinin %70'i olan -0,13 ve yüksek kesme puanı da düzey 2'de kestirilen en yüksek yetenek değerinin %70'i olan 0,33 olarak hesaplanmıştır. Ardından veri setine post-hoc simülasyon uygulanmıştır. Simülasyon çalışmasında teste başlama kuralı olarak kolay madde verilmesi seçilmiş; yetenek kestiriminde AOK kullanılmış ve maksimum madde sayısı olarak 25 madde belirlenmiştir. Simülasyon koşulları ortalama madde sayısı ve doğru sınıflama oranıyla değerlendirilmiş; ayrıca madde kullanım sıklık oranları da dikkate alınmıştır. Araştırma sonuçlarına göre üç kategorili sınıflamada madde seçme yöntemleri benzer sonuçlar vermiş olsa da MFB-KY'nin MFB-KN'den daha az sayıda maddeyle testi sonlandırdığı görülmüş; MFB-KN'nin daha az sınıflama hatası içerdiği ve içerik dengelemenin madde kullanım sıklık kontrolüne kıyasla sınıflama hatası bakımından daha iyi performans gösterdiği görülmüştür. Karşılaştırma yapabilmek adında kağıt-kalem testi sabit uzunluklu 25 madde ile de simüle edilmiş ve bu durumda %87 doğru sınıflama yapıldığı görülmüştür. GA'da bazı koşullarda %87'nin geçilemediği görülürken AOOT'de tüm koşullarda %87'den yüksek doğru sınıflama yapılmıştır. Kağıt-kalem testinden %22 ile %44 oranları arasında daha az sayıda maddeyle test sonlanmıştır. Madde sayısı ve doğru sınıflama oranına göre AOOT, GA'dan daha iyi performans göstermiştir. Ancak çalışmada iki yöntemin tamamen benzer alt yapısı olmamasından dolayı sonuçların genellenebilirliğinin düşük olabileceği ifade edilmiştir.

Lin ve Spray (2000) çalışmalarında AOOT altında MFB, KLB ve Ağırlıklı Log-Odds Oranı (ALOO; Weighted Log-Odds Ratio: WLOR) madde seçme yöntemlerini ortalama test uzunluğu ve sınıflama doğruluğu bakımından karşılaştırmışlardır. Çalışmada ayrıca 11 farksızlık bölgesi, tam (360 maddelik) ve yarım (180 maddelik) olmak üzere iki gerçek madde havuzu ve üç madde kullanım sıklığı kontrol yöntemi de ele alınmıştır. Normal dağılımdan türetilen 10.000 yetenek düzeyi üzerinden gerçekleştirilen simülasyon çalışması sonuçlarına göre madde seçme yöntemlerinin, farksızlık bölgesi, madde havuzu ve kullanım sıklığı yöntemi fark etmeksizin benzer sonuçlar verdiği görülmüştür.

Jiao ve Lau (2003) çalışmalarında AOOT altında model uyumsuzluklarını incelemişlerdir. Bu amaçla 1 PLM, 2 PLM ve 3 PLM ele alınmıştır. Modellerden biri gerçek model varsayılırken diğer modellerin kullanılması durumundaki uyumsuzluklar sınıflama hataları bakımından incelenmiştir. Araştırma sonuçlarına göre 1 PLM ve 2 PLM gerçek model iken diğer modellerin kullanılması sınıflama hatalarını etkilememiştir. 3 PLM'nin gerçek model olması durumunda ise 1 PLM'nin yanlış ve pozitif sınıflama hatalarını; 2 PLM'nin ise yanlış ve negatif sınıflama hatalarını artırdığı görülmüştür. Bu bulguya dayanarak MTK model seçiminin BBST uygulamaları için önemli olduğu sonucuna ulaşılmıştır.

Thompson (2007a) çalışmasında, çok kategorili maddelerde çoklu kesme puanlarının kullanıldığı BBST koşullarını simüle ederek sonuçları ortalama test uzunluğu ve doğru sınıflama oranı bakımından karşılaştırmıştır. Bu amaçla beş bağımsız değişken ele alınmış ve toplam 32 koşul incelenmiştir. Bunlar (i) AOOT ve BUT olmak üzere iki farklı sınıflama kriteri; (ii) MFB-KY ve MFB-KN olmak üzere iki farklı madde seçme yöntemi; (iii) 3 PLM ve GKKM olmak üzere iki psikometrik model; (iv) uniform ve normal dağılımdan türetilen iki madde havuzu; (v) bir ve iki olmak üzere iki farklı kesme puanıdır. Bu amaçla kesme puanları keyfi olarak ikili sınıflamada 0,675 ve üçlü sınıflamada ise -0,675 ve 0,675 olarak belirlenmiştir. Madde havuzları uniform ve normal dağılımdan üretilmiş; kesme puanı sayısına göre toplam 6 madde havuzu oluşturulmuştur. Farksızlık bölgesinin doğru sınıflama ve ortalama test uzunluğunda etkisi olduğundan çalışmada 16 farksızlık bölgesi değeri için simülasyon tekrarlanmış; BUT ile hata oranı en yakın olan farksızlık bölgesi temel alınarak kriterler ortalama test uzunluğu bakımından karşılaştırılmıştır. Simülasyonda her 32 koşul için 25 tekrar yapılmış ve sonuçlar

ANOVA ile özetlenmiştir. Öğrenci yetenekleri 13 düzeye ayrılmış ve yetenek parametreleri her yetenek düzeyinde 1000 öğrenci bulunacak şekilde türetilmiştir. Her düzey için ortalama test uzunluğu ve sınıflama doğruluğu ayrı hesaplanmıştır. Çalışma sonunda en az sayıda madde GKKM'de tek bir kesme puanı olması durumunda AOOT ve BUT kriterlerinin her ikisi için de MFB-KN yönteminden elde edilmiştir. En uzun test ise 3 PLM'de iki kesme puanı kullanılan BUT sonlandırma kriteri ve MFB-KN madde seçme yöntemi ile elde edilmiştir. ANOVA sonuçlarına göre ortalama test uzunluğu ve doğru sınıflama oranına ait hata varyansı oldukça düşüktür. Buna göre örnekleme hatasının düşük ve simülasyonların tekrarlar arasında durağan olduğu söylenebilir. Ortalama test uzunluğu için varyansın %54'ü kesme puanı sayısından geldiği; doğru sınıflama oranı için ise varyansın %79,3'ü MTK modelinden kaynaklandığı görülmüştür.

Thompson ve Ro (2007) çalışmalarında AOOT, GA ve Bileşik Olabilirlik Oranı (BOO; Combined Likelihood Ratio: CLR) olmak üzere üç farklı sınıflama kriterinin farklı düzeydeki güvenirliklerini ve farksızlık bölgesi değerlerini, ortalama test uzunluğu ve sınıflama doğruluğu bakımından incelemişlerdir. Bu amaçla 750 maddelik madde havuzunun parametreleri ve 10.000 bireye ait yetenek düzeyi normal dağılımdan türetilmiş; kesme noktası 0,5 ve maksimum madde sayısı 200 olarak belirlenmiştir. Simülasyon AOOT için beş farksızlık bölgesi değeri, BOO ve GA sınıflama kriterleri için ise iki farklı güvenirlik düzeyi ele alınmıştır. Ayrıca BOO için MFB-KY ve MFB-KN madde seçme yöntemleri çaprazlanmıştır. Araştırma sonunda AOOT için en az sayıda madde (20 civarı) 0,5 farksızlık düzeyi ile ve en fazla sayıda madde (80 civarı) 0,2 farksızlık düzeyiyle elde edilmiştir. En az sayıda maddeyle testin sonlanması BOO için MFB-KN ile; GA için ise %95 güven aralığında hesaplanmıştır. Tüm koşullar için hesaplanan sınıflama doğruluğunun %90'nın üstünde olduğu sonucuna ulaşılmıştır.

Thompson (2009) çalışmasında AOOT ve GA sınıflama kriterlerini madde seçme yöntemlerine ve madde havuzlarına göre test uzunluğu ve sınıflama doğruluğu bakımından karşılaştırmıştır. Bu amaçla çalışmada MFB-KY ve MFB-KN madde seçme yöntemleri; sivri ve basık dağılıma sahip 350 ve 700 maddelik madde havuzları çaprazlanmıştır. Yetenek düzeylerinin 10.000 birey üzerinden ve normal dağılımdan türetilmiş olduğu bu çalışma sonuçlarına göre tüm madde havuzlarında AOOT sınıflama kriteri ve MFB-KN madde seçme yönteminin en az

sayıda maddeyle ve en yüksek sınıflama doğruluğuyla testi sonlandırdığı, bir başka deyişle en iyi performansı gösterdiği bulunmuştur. Bunun yanında 750 maddelik sivri dağılımlı havuzdan daha az sayıda maddeyle daha yüksek güvenirlikte testin sonlandığı; sınıflama kriterlerinin benzer performans gösterdiği; kesme noktasında bilgi veren madde seçiminin ise az sayıda maddeyle daha yüksek bir sınıflama doğruluğu sağladığı görülmüştür.

Wouda ve Eggen (2009) çalışmalarında Stokastik Azaltmalı Ardışık Olasılık Oran Testini (SA-AOOT; Sequential Probability Ratio Test with Stochastic Curtailment: SC-SPRT) üç kategorili sınıflama durumuna uygun şekilde modifiye etmiş; SA-AOOT ile Ucu Kesik Ardışık Olabilirlik Oran Testi (UK-AOOT; The Truncated Sequential Probability Ratio Test: T-SPRT) sınıflama kriterlerini karşılaştırmıştır. Ayrıca Finkelaman'ın (2003) çalışmasını yetenek kestirim yöntemini AOK şeklinde değiştirerek tekrarlamış ve sonuçları karşılaşmıştır. Sonuçta Finkelman'ın çalışmasına kıyasla, SA-AOOT ile UK-AOOT arasında sınıflama doğruluğu bakımından değişim olmadığı ancak ortalama madde sayısında farklılıklar çıktığı görülmüştür. Gerçek veri seti üzerinden yapılan post-hoc simülasyon çalışma sonuçlarına göre ise iki yöntem arasında sınıflama doğruluğu ve madde sayısı bakımından manidar farklılık olmadığı görülmüştür.

Thompson (2011) çalışmasında AOOT, GOO ve GA sınıflama kriterlerini ortalama test uzunluğu ve sınıflama doğruluğu bakımından karşılaşmıştır. Bu amaçla normal dağılımdan 10.000 bireye ait yetenek parametreleri türetilmiş ve 50, 100 ve 200 maddelik madde havuzları oluşturulmuştur. AOOT ve GOO için farksızlık bölgesi değerleri 0,3 ve 0,2 şeklinde ayarlanmıştır. Araştırma sonuçlarına göre GOO için 0,3 farksızlık bölgesi değeri ile madde sayısı ve sınıflama doğruluğu bakımından en uygun testin oluşturulduğu görülmüştür.

Nydick, Nozawa ve Zhu (2012) çalışmalarında üç ve beş kategorili sınıflamalarda AOOT, GOO, SA-AOOT ve GA'nın ortalama test uzunluğu ve doğru sınıflama oranları bakımından incelenmesini amaçlamıştır. Bu amaçla 600 maddelik büyük ölçekli bir sınava ait gerçek veri seti 3 PLM ile kalibre edilmiştir. 3 kategorili sınıflama için -0,47 ve 1,18 noktaları; 5 kategorili sınıflama için ise -1,39; -0,47; 0,28 ve 1,18 kesme puanları keyfi olarak belirlenmiştir. Ayrıca 8 alt boyutun her birinden test başlangıcında birer madde alınmasına karar verilmiştir. Simülasyon, herhangi bir öğrenci için minimum 8 ve maksimum 21 maddeyle manipüle

edilmiştir. İlk 4 madde -0,5 ve 0,5 aralığından rasgele seçilmiş ve MOK ile kestirim yapılmıştır. Yetenek kestiriminde, ilk maddeler haricinde, AOK kullanılmıştır. Analiz sonuçlarına göre SA-AOOT ve GOO sınıflama kriterleri test uzunluğu ve doğru sınıflama oranı bakımından diğer yöntemlere kıyasla iyi performans göstermiş; GOO, SA-AOOT ile benzer sonuçlar vermiş; SA-AOOT'nin yoğun hesaplama gerektirmesi nedeniyle GOO'nun daha tercih edilebilir olduğu belirtilmiştir.

Huebner (2012) çalışmasında AOOT sınıflama kriteri altında madde kullanım sıklığı kontrol yöntemlerini ortalama test uzunluğu, sınıflama doğruluğu, maksimum madde kullanım sıklığı oranı, havuzdaki fazla kullanılan maddelerin oranı ve fazla kullanılan maddelerin ortalama kullanım sıklığı bakımından karşılaştırmıştır. Bu amaçla Thompson'ın (2009) çalışmasındaki madde havuzuna benzer şekilde iki farklı güçlükte havuz simüle etmiş; yetenek düzeyini normal dağılımdan 1000 birey olacak şekilde düzenlemiştir. Çalışma sonucunda madde kullanım sıklığı yöntemlerinin ortalama test uzunluğu ve sınıflama doğruluğu bakımından birbirine benzer sonuçlar verdiği; testin güçlüğü arttıkça testin sonlanması için gereken madde sayısının arttığı ve sınıflama doğruluğunun yükseldiği görülmüştür. Ayrıca madde kullanım sıklığı bakımından yöntemler arasında diğer bağımlı değişkenler bakımından farklılıklar olduğu görülmüştür.

2.3. İlgili Araştırmalar Özet

Yetenek kestirim yöntemleriyle ilgili çalışmalar özetlenecek olursa, BBT uygulamalarında yetenek kestirim yöntemlerinin yanlılık, RMSE, SH ve madde sayısı değişkenlerinin üzerinde önemli etkisi olduğu görülmektedir (Warm, 1989; Wang, 1997; Wang ve Vispoel, 1998; Wang, Hanson ve Lau, 1999; Yi, Wang ve Ban, 2000; Wang ve Wang, 2001; Tao, Shi ve Chang, 2012; Gökçe, 2012; Eroğlu, 2013; Kezer, 2013). Çalışmalara göre yanlılık bakımından AOK yönteminin (Warm, 1989; Cheng ve Liou, 2000; Gökçe, 2012); RMSE ve SH'ler bakımından BSD yönteminin (Wang, 1997; Wang ve Vispoel, 1998; Wang ve Wang, 2001); madde sayısı bakımından ise BSD ve MSD gibi Bayesci yöntemlerin (Yi, Wang ve Ban, 2000; Wang ve Wang, 2001; Kezer, 2013; Sulak, 2013) daha iyi performans gösterdiği görülmektedir Bununla birlikte kimi çalışmalarda AOK yönteminin diğer yetenek kestirim yöntemlerine kıyasla iyi performans göstermiş olduğu görülürken (Wang ve Wang, 2001; Gökçe, 2012; Cheng ve Liou, 2000); özellikle değişken uzunluklu testleri içeren kimi çalışmalarda AOK yöntemini kullanmanın sakıncalı

olabileceği göze çarpmaktadır (Wang, Hanson ve Lau, 1999; Yi, Wang ve Ban, 2000).

Bu çalışmada yetenek kestirim yöntemlerinden BSD ve AOK ele alınmış ve karşılaştırılmıştır. Çalışmada AOK'un incelenmesinin nedeni, ilgili çalışmalara göre sabit uzunluklu testlerde AOK'un diğer yöntemlere kıyasla daha yansız bir kestirici olması ancak BBST'nin de dâhil olduğu değişken uzunluklu testlerde daha kötü bir performans sergilemesidir. Buna dayanarak AOK yönteminin BBST uygulamalarında nasıl bir kestirici olduğunun incelenmesi önemlidir. BSD yetenek kesirim yönteminin seçilme nedeni ise hem RMSE ve SH'ler hem de test etkililiği bakımlarından diğer yöntemlere kıyasla daha iyi performans göstermesidir.

Madde seçme yöntemleriyle ilgili çalışmalar incelendiğinde, madde seçme yöntemlerinin özellikle test etkililiği bakımından önemli olduğu görülmektedir (Cheng ve Liou, 2000; Diao ve Reckase, 2009; Sulak, 2013). Yukarıdaki araştırmalarda, BBT uygulamalarında MFB ve KLB madde seçme yöntemlerinin, Bayesci yetenek kestirim yöntemleri olan BSD veya MSD ile bir arada ele alınması durumunda en az sayıda maddeyle testi sonlandırdığı görülmektedir (Diao ve Reckase, 2009; Sulak, 2013). Ayrıca AOK yetenek kestirim yöntemiyle birlikte KLB'nin madde sayısı bakımından oldukça iyi performans gösterdiği; yetenek kestirim yöntemi fark etmeksizin kestirilen yetenekte yüksek bilgi veren maddelerin daha düşük RMSE değeri verdiği; AOK ile birlikte KLB veya MFB kullanımının 10 maddeden sonra benzer çalıştığı görülmektedir (Cheng ve Liou, 2000).

Bu çalışmada madde seçme yöntemlerinden MFB ve KLB ele alınmıştır. Bunun sebebi her iki madde seçme yönteminden ikişer farklı yöntem türetilmiş olmasıdır. Bu iki madde seçme yöntemiyle BBST'de hem kestirilen yetenekte en yüksek bilgiyi veren maddenin seçimi (KY: Kestirilen Yetenek temelli) hem de kesme noktasında en yüksek bilgiyi veren maddenin seçimi (KN: Kesme Noktası temelli) mümkün olabilmektedir. Alayazın incelendiğinde MFB ve KLB madde seçme yöntemleri BSD veya AOK ile birlikte çalıştırıldığında oldukça az sayıda maddeye ihtiyaç duyulması çalışmada bu iki madde seçme yöntemin seçilmesine ayrıca neden olmuştur. KY ve KN arasındaki farklılık ise oldukça az sayıda çalışmada değinilmiş bir konudur. Örneğin Spray ve Reckase (1994) çalışmasında kesme noktasında en yüksek bilgiyi veren madde seçme yöntemiyle daha kısa test oluştuğunu gösterirken; Eggen ve Straetmans (2000) araştırmalarında tam aksini,

geçici yetenek düzeyinde en yüksek bilgi veren maddenin seçilmesi durumunda testin kısaldığını ortaya koymuştur. Bu bulgulara dayanarak, bu dört madde seçme yönteminden hangisinin BBST uygulamalarında daha iyi performans gösterdiğini belirleyebilmenin önemli olduğu düşünülmüştür.

BBST ile ilgili araştırmalar incelendiğinde, BBST uygulamalarında sonlandırma kriterlerinin tek başına veya birlikte çalışıldığı; madde seçme yöntemleriyle ve madde havuzlarıyla çaprazlandığı; kriterlerin çoklu kategoriye sınıflama ve çoklu puanlanan maddelerin kullanılması durumunda performanslarının nasıl olduğunun incelendiği görülmektedir. Araştırmalarda AOOT'nin sıklıkla diğer sonlandırma kriteriyle karşılaştırılması söz konusudur. İncelenen çalışmalara göre, AOOT'nin Bayesci sonlandırma yaklaşımlarına kıyasla daha az sayıda madde gerektirdiği (Spray ve Reckase, 1994); BUT ve geleneksel testlere kıyasla ise test uzunluğu ve sınıflama doğruluğu bakımından daha iyi sonuçlar verdiği görülmektedir (Kingsburry ve Weiss, 1980). Alanyazın incelendiğinde AOOT'nin, bu çalışmada incelenmiş olan GOO ve GA sonlandırma kriterleriyle bir arada ele alındığı herhangi bir çalışmaya rastlanmadığı görülmektedir. Bu nedenle bu çalışmada AOOT, GOO ve GA sınıflama kriterlerinin karşılaştırmasına yer verilmiştir.

Reckase'in (1983) çalışması incelendiğinde AOOT'nin, 3 PLM'de ±0,8 ve ±1,0 farksızlık bölgesi değerlerinde doğru sınıflama oranının yüksek ve ortalama madde sayısının düşük olduğu; farksızlık bölgesi değeri düştükçe doğru sınıflama oranının düştüğü ve madde sayısının arttığı; MTK modelleri bakımından 3 PLM ile hesaplanan doğru sınıflama oranının 1 PLM'e göre daha yüksek olduğu; 3 PLM'den elde edilen ortalama madde sayısının da 1 PLM'e kıyasla daha az olduğu görülmektedir. Ayrıca Lau'nun (1996) ile Jiao ve Lau'nun (2003) çalışmalarının sonuçları da BBST'de 3 PLM'in göreli olarak daha iyi sonuçlar verdiğini göstermektedir. Bu nedenle çalışmada 3 PLM ele alınmış; farksızlık bölgesi ve güven düzeyi değerlerinin değişimi de incelenmiştir.

Alanyazın incelendiğinde sonlandırma kriterlerinin farklı madde seçme yöntemleriyle çaprazlandığı çalışmalara rastlanmaktadır. Spray ve Reckase'in (1994) araştırmalarında AOOT'nin MFB-KN madde seçme yöntemi ele alındığında ortalama test uzunluğunun daha kısa olduğu belirlenmiş; ancak bu çalışmada MFB dışında herhangi bir madde seçme yönteminin yer almadığı görülmüştür. Bu çalışmada MFB'nin yanında KLB'ye de yer verilmesinin temel nedeni bu

durumdur. Benzer bir bulgu Lau ve Wang (1998), Eggen (1999) ile Thompson'ın (2009) araştırmasında da ortaya çıkmıştır. Eggen'in (1999) çalışmasında AOOT'nin hem MFB hem de KLB açısından incelenmiş olduğu ancak bu sefer de AOOT'nin yanında başka bir sonlandırma kriterinin yer almadığı görülmektedir. Alanyazındaki bu eksiklik, tek bir çalışmada birden fazla sonlandırma kriterinin, madde seçme yönteminin ve yetenek kestirim yönteminin karşılaştırılması gerekliliğini ortaya çıkarmaktadır.

Spray ve Reckase'in (1994) sonuçlarının aksine, Eggen ve Straetmans'ın (2000) çalışmasında, kestirilen yetenekte en yüksek bilgi veren madde seçimi kullanıldığında ortalama test uzunluğunun kısaldığı ve AOOT'nin GA'dan daha az maddeyle daha doğru sınıflama yapabildiği görülmüştür. Özetlemek gerekirse AOOT'nin diğer sınıflama kriterlerine üstünlük sağladığı ancak madde seçme yöntemleri bakımından farklı çalışmalarda çelişkili sonuçlar elde edildiği söylenebilir. Bu nedenle bu çalışmada madde seçme yöntemleriyle sonlandırma kriterleri birçok koşul altında incelenmiştir.

İlgili çalışmalar incelendiğinde, bu çalışmanın amacı olan, her biri iki farklı farksızlık bölgesi veya güven düzeyi değeriyle ele alınan üç sonlandırma kriterinin, iki yetenek kestirim yöntemi ve dört farklı madde seçme yöntemiyle çaprazlandığı herhangi bir çalışmaya rastlanmadığı görülmektedir. Bu nedenle çalışmada AOOT, GA ve GOO sonlandırma kriterleri; BSD ve AOK yetenek kestirim yöntemleriyle; madde seçme yöntemlerinden MFB-KY, KLB-KY, MFB-KN ve KLB-KN ile çaprazlanarak iki ayrı simülasyon üzerinden toplam 96 adet koşul oluşturulmuştur. Oluşturulan koşullardan hangisinin veya hangilerinin ortalama test uzunluğu, ortalama sınıflama doğruluğu ve ölçme kesinliği bakımından daha iyi performans gösterdiği belirlenmeye çalışılmıştır.

3. YÖNTEM

Bu bölümde araştırmanın türüne, simülasyonlarda kullanılacak verinin türetilmesine ve elde edilmesine, verinin istenen koşullara uygunluğuna, araştırma koşullarına ve veri analizine yer verilmiştir.

3.1. Araştırmanın Yöntemi

Bu araştırmada Bireyselleştirilmiş Bilgisayarlı Sınıflama Testi (BBST) uygulamalarında ele alınan farklı sınıflama kriterlerinin, yetenek kestirim yöntemlerinin ve madde seçme yöntemlerinin, sınıflama doğruluğu, test uzunluğu, yanlılık, korelasyon, RMSE ve mutlak hata bakımından incelenmesi amaçlandığından, araştırmanın betimsel olduğu söylenebilir. Betimsel araştırmalar, olayların, objelerin, varlıkların, kurumların, grupların ve çeşitli alanların "ne" olduğunu açıklamaya çalışır (Kaptan, 1977). Diğer taraftan bu araştırma bir simülasyon çalışmasıdır. Diğer araştırma yöntemleri, geçmişe dayalı olmak üzere "Ne, Nasıl, Neden oldu?" ile ilgilenirken, simülasyon çalışmaları "… olsa ne olurdu?" sorusuna cevap aramaktadır (Dooley, 2002).

3.2. Verinin Türetilmesi ve Elde Edilmesi

Bu çalışmada Monte Carlo (MC) ve Post-Hoc (PH) simülasyon çalışmalarında ayrı ayrı olmak üzere iki farklı veri seti kullanılmıştır. Araştırmanın MC simülasyonu bölümündeki veri seti R ortamında türetilmiştir (R Core Team, 2013). Araştırmanın PH simülasyonu bölümünde ise Kezer'in (2013) doktora çalışmasında elde ettiği gerçek madde cevap örüntüsü kullanılmıştır. Türetilen MC veri setinin araştırmada incelenecek koşullara uygun özellikleri göstermesi sağlanırken; gerçek bir madde cevap örüntüsü üzerinden gerçekleştirilen PH simülasyonunda veri setiyle ilgili herhangi bir manüpülasyon söz konusu olmamıştır. Böylece hem simülatif veri seti hem de gerçek veri seti üzerinden gerçekleştirilen BBST simülasyon sonuçlarının incelenmesi sağlanabilmiştir. MC ve PH simülasyonlarında kullanılan veri setleri ve araştırma amacına uygunlukları aşağıda detaylıca ele alınmıştır.

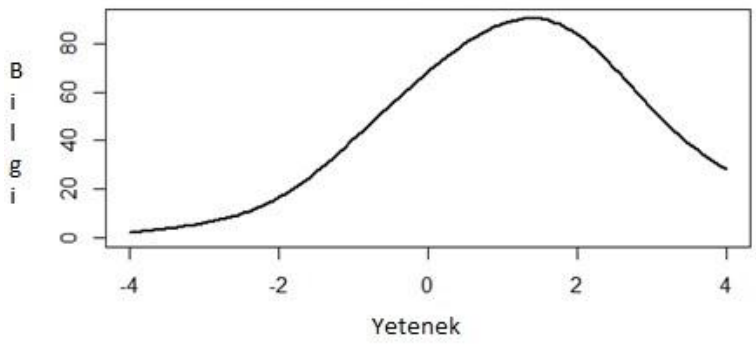
3.2.1. Monte Carlo (MC) Simülasyonu İçin Madde ve Yetenek Parametrelerinin Türetilmesi

Çalışmanın MC simülasyonu kısmında Thompson'ın (2011) araştırması dikkate alınmış ve madde havuzunun 3 PLM temel alınarak 500 maddeden oluşması sağlanmıştır. Araştırmada hem kestirilen yetenek (KY) hem de kesme noktası

(KN) temelli madde seçme yöntemleri karşılaştırılacağından madde havuzunun belirlenen kesme noktası olan 1,0 ve etrafında yüksek bilgi verecek; -3,+3 yetenek düzeyleri aralığını kapsayacak şekilde oluşturulmasına dikkat edilmiştir. Bu sebeple havuzdaki maddeler:

- a parametresinin orta ve yüksek değerlerde olabilmesi adına uniform dağılımdan [0,5; 2,0] aralığından;
- b parametresinin, Warm'ın da (1989) çalışmasında belirttiği gibi, gerçek uygulamadaki değerlere yakın olabilmesi adına normal dağılımdan ortalaması 1,0 ve standart sapması 1,5 olmak üzere;
- c parametresi ise yine gerçek bir uygulama düşünülerek normal dağılımdan ortalaması 0,15 ve standart sapması 0,05 olacak şekilde türetilmiştir. MC simülasyonu için türetilen madde parametrelerinin betimsel özellikleri Ek 1'de yer almaktadır.

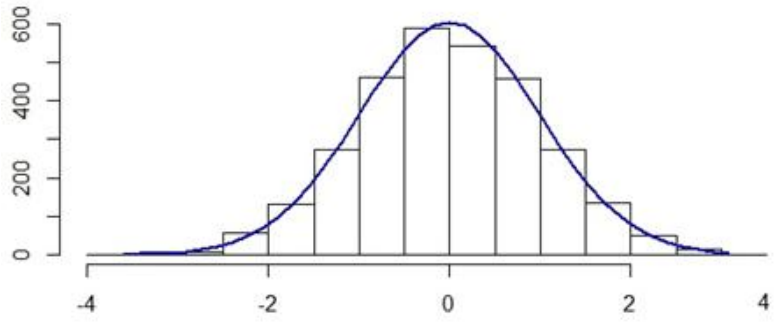
Şekil 2'de, oluşturulan madde havuzunun test bilgi fonksiyonu yer almaktadır. Şekil 2'de görüldüğü üzere madde havuzu amaçlandığı şekilde kesme noktası olan 1,0 etrafında yüksek bilgi veren ve -3,+3 yetenek düzeyi aralığını kapsayan maddelerden oluşmaktadır.



Şekil 3.1. Test Bilgi Fonksiyonu

Araştırmanın MC simülasyonu çalışmasında bireylerin yetenek parametreleri (-3,+3) yetenek düzeyi aralığında, ortalaması 0 ve standart sapması 1 olan normal dağılım yardımıyla (N(0,1)) toplam 3000 kişi üzerinden R yazılımında rasgele türetilmiştir (R Core Team, 2013). Türetilen yetenek parametrelerinin betimsel istatistikleri Ek 2'de gösterilmektedir. Aşağıda Şekil 3'te bireylere ait yetenek

dağılımı yer almaktadır. Şekil 3'te görüldüğü gibi bireylerin yeteneklerinin, amaçlanan şekilde -3,+3 yetenek düzeyleri aralığında normal dağıldığı söylenebilir.



Şekil 3.2. Yetenek Parametrelerinin Dağılımı Grafiği

MC simülasyonu için madde ve birey parametrelerinin türetilmesinden sonra 3000 bireyin 500 maddeye verdiği cevaplar (madde cevap örüntüsü) R ortamında Nydick tarafından (2014) yazılan "catIrt" paketindeki "simIrt" komutu yardımıyla simüle edilmiştir. Türetilen madde havuzu, incelenen koşullara uygun olmak üzere birçok özelliği bir arada gösterecek şekilde manipüle edildiğinden, bir başka deyişle hem kesme noktasında hem de yetenek ölçeği boyunca yüksek bilgi veren çok sayıda maddeden oluşması sağlandığından, simülatif madde cevap örüntüsünün MTK varsayımlarına uygunluğu aşağıda detaylıca incelenmiştir.

3.2.1.1. MC Veri Setinin Tek Boyutluluk Varsayımının İncelenmesi

Bu çalışmanın MC simülasyonu bölümünde veri tek boyutlu MTK modellerine dayalı olarak R ortamında seçkisiz simüle edilmiştir. Lau (1996) ile Spray ve diğerleri (1997) çalışmalarında, sınıflama kriterlerinin tek boyutluluk varsayımının sağlanamadığı durumlarda bile oldukça başarılı performans gösterdiklerini ifade etmektedir. Bununla birlikte BBST simülasyonuna geçmeden simülatif veri setinin boyutluluğu açımlayıcı faktör analiziyle R'da ve doğrulayıcı faktör analiziyle de LISREL programında incelenmiştir.

Tek boyutluluk, maddelerin ölçtüğü ve bireylerin cevaplama performanslarının altında yatan tek bir örtük özellik olması anlamına gelmektedir. Bir başka deyişle madde cevapları arasındaki varyansın tek bir örtük özellik tarafından açıklanmasıdır (Hambleton ve Swaminathan, 1985). Tek boyutluluğun incelenmesinde alanyazında birçok deneysel ve istatistiksel yöntem önerilmiştir.

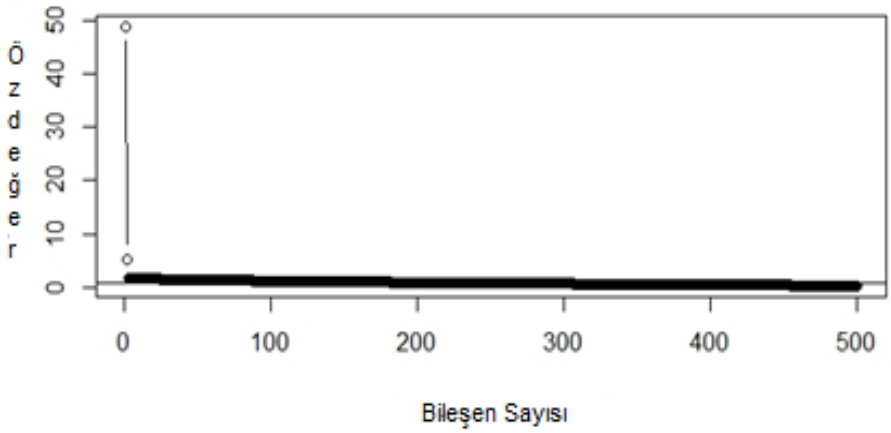
Bu çalışmada tek boyutluluk varsayımının incelenmesi amacıyla, türetilen veri setine R ortamında Revelle (2015) tarafından yazılan “psych” paketindeki “irt.fa” komutuyla tetrakorik korelasyonlara dayanan temel bileşenler faktör analizi uygulanmıştır. Faktör analizi öncesinde aynı paketteki “cortest.bartlett” komutu yardımıyla Bartlett küresellik testi sınanmış ve “KMO” fonksiyonuyla da KMO değeri hesaplanmıştır.

Analiz sonuçlarına göre hesaplanan korelasyon matrisine ait Bartlett testi sonucu manidar (Ki-Kare= 270695,2; sd= 124750; p=0,00) ve hesaplanan KMO değeri 0,96’dır. Bartlett’e ait Ki-kare değerinin manidar çıkması verilerin çok değişkenli normallik varsayımını sağladığı anlamına gelmektedir. KMO testi sonucuna dayanarak da (KMO = 0,96 > 0,60) veri setinin faktörleşebilirlik varsayımını sağladığı yorumu yapılabilir. Faktör analizi sonucu açıklanan varyans ve özdeğerler aşağıda Tablo 3.1.’de yer almaktadır.

Tablo 3.1: MC Veri Setinin Faktör Analizi Sonucu Açıklanan Varyans ve Özdeğerleri

Değişken	*Özdeğer*	*Varyans Oranı*	*Kümülatif Varyans Oranı*
1	81,46	0,16	0,16
2	6,16	0,01	0,17
3	2,09	0,00	0,17
4	1,75	0,00	0,17
5	1,64	0,00	0,17

Analiz sonuçlarına göre özdeğeri 1’in üzerinde olan 45 bileşen olsa da Tablo 3.1.’de diğer bileşenlere ait özdeğerlerin birbirine oldukça yakın olmaları sebebiyle ilk beş bileşene yer verilmiştir. Birinci bileşene ait özdeğer 81,46 ve ikinci bileşene ait özdeğer 6,16’dır. İlk bileşenin özdeğerinin ikinci bileşenin özdeğerine oranı 13,20’dir. Bu oranın 3’ten büyük olması verinin tek faktörlü bir yapı gösterdiğine işaret etmektedir (Lord, 1980). Ayrıca bileşenlerin açıklanan varyansa katkıları incelendiğinde ikinci bileşenle birlikte kümülatif varyansın çok az arttığı; ilk bileşenin başat bir şekilde toplam varyansın %16’sını açıkladığı görülmektedir. Aşağıda Şekil 4’te R ortamında aynı paketteki “vss.scree” komutuyla çizilen yamaç-birikinti grafiği incelendiğinde ilk boyuttan sonra keskin bir düşüşün olduğu ve ikinci faktör ile birlikte diğer faktörlerin varyansa ciddi katkı sağlamadıkları görülmektedir. Bu bulgulara dayanarak verinin tek boyutluluk varsayımını sağladığı söylenebilir.



Şekil 3.3. MC Verisine Ait Yamaç-Birikinti Grafiği

DFA sonucunda Ki-kare değeri 142689,787 ve serbestlik derecesi (sd) 124250 olarak hesaplanmıştır. Buna göre Ki-kare değerinin sd'ye oranı 1,15'tir ve bu oranın 2'den küçük olması sebebiyle verinin tek boyutluluğa mükemmel uyum gösterdiği söylenebilir (Tabachnick ve Fidell, 2007). Bunların yanında uyum indekslerinden NNFI = 0,87; CFI = 0,87; GFI = 0,97; AGFI = 0,97 olarak hesaplanmıştır. Uyum indekslerine göre de veri setinin tek boyutluluk varsayımını mükemmele yakın düzeyde karşıladığı görülmektedir. Ayrıca hesaplanan Alfa güvenirlik katsayısı 0,98'dir ve buna dayanarak da güvenirliğin yüksek olduğu yorumu yapılabilir. Ancak Alfa güvenirlik katsayısının madde sayısının artışına bağlı olarak yüksek değerler verebileceği unutulmamalıdır (Cortina, 1993). Bu bulgular dikkate alındığında veri setinin tek boyutluluk varsayımını sağladığı söylenebilir.

3.2.1.2. MC Veri Setinin Yerel Bağımsızlık Varsayımının İncelenmesi

Yerel bağımsızlık varsayımı, farklı maddelere verilen öğrenci cevaplarının istatistiksel olarak birbirinden bağımsız olması olarak düşünülebilir. Bir öğrencinin bir maddedeki performansı testteki diğer maddeleri iyi ya da kötü etkilememelidir. Bir madde diğer bir maddenin cevabına ipucu olmamalıdır. Bir başka deyişle sabit bir yetenek düzeyinde iki maddenin birbirinden bağımsız olması gerekmektedir (Hambleton ve Swaminathan, 1985). Ayrıca tek boyutluluk yerel bağımsızlığa ilişkin bir kanıt olarak da kabul edilmektedir (Embretson ve Reise, 2000).

Buna göre 3000 öğrenci toplam puanlarına göre 4 yetenek grubuna ayrılmıştır. Yerel bağımsızlığın sağlanabilmesi için yüksek ve düşük yetenek grubu oluşturulup yetenek sabitlendiğinde, maddeler arasında yüksek ve manidar ilişki bulunmaması gerekmektedir. Tüm yetenek grupları için hesaplanan Pearson korelasyon katsayılarına göre maddeler arasındaki korelasyonların büyük bir çoğunluğunun anlamsız; manidar olanların ise düşük düzeyde ilişkili olduğu görülmektedir. Buna göre maddeler arasında yerel bağımsızlık varsayımının büyük oranda sağlandığı söylenebilir.

Yerel bağımsızlık varsayımı Q3 ve G2 gibi istatistiklerle test edilebilmektedir. Ancak 500 maddelik bir test için maddelerin ikili kombinasyonlarının farklı yetenek düzeyine sahip öğrenciler için tekrarlı hesaplanması gerekmektedir. McDonald'a (1980) göre tek boyutluluğun anlamlı bir açıklığı, yerel bağımsızlık ilkesini temel alır ve eğer benzer yetenek düzeyindeki öğrenciler için maddeler arasındaki kovaryans sıfıra eşitse test tek boyutludur (Akt: Hambleton ve Swaminathan, 1985, s. 25). Bir başka deyişle eğer bir test anlamlı bir açıklıkla tek boyutluluk özelliği gösteriyorsa bu testte yer alan maddelerin yerel bağımsızlık özelliğine de sahip olduğu söylenebilir. Buna göre yapılan faktör analizi sonucunda testin tek boyutluluk özelliği göstermesi, aynı zamanda testte yer alan maddelerin yerel bağımsız olduğu şeklinde yorumlanabilir.

3.2.1.3. Testin Hız Testi Olmaması

Testin hız testi olup olmaması cevaplanmamış madde sayısı, testi bitirememiş birey sayısının yanında paralel test uygulaması gibi yöntemlerle kontrol edilebilmektedir (Hambleton ve Swaminathan, 1985, s. 157-161). Veri seti incelendiğinde, verinin simülatif olması sebebiyle, cevaplanmamış madde olmadığı ve testi bitirememiş birey olmadığı görülmektedir. Buna göre testin hız testi olmadığı, buna dayanarak en temel varsayım olan tek boyutluluk varsayımını etkileyecek bir durum olmadığı ve veri setinin MTK varsayımlarını karşıladığı söylenebilir.

Çalışmada MTK varsayımlarının analizi sonrasında BBST koşullarının simülasyonuna geçilmiştir. Simülasyonda Nydick (2014) tarafından yazılan catirt paketinden yararlanılmıştır. BBST simülasyonu R ortamında, oluşturulan 48 koşulun her biri için 25 tekrarla for döngüsü yazılarak sonlandırılmıştır.

Araştırmanın bağımlı değişkenleri olan ortalama test uzunluğu, ortalama sınıflama doğruluğu, gerçek ve kestirilen yetenekler arasındaki korelasyon, yanlılık, RMSE, ortalama mutlak hata yine R yazılımında, araştırmacı tarafından yazılan fonksiyon ve döngülerle elde edilmiştir.

3.2.2. Post Hoc (PH) Simülasyonu İçin Madde ve Yetenek Parametrelerinin Elde Edilmesi

Çalışmanın PH simülasyonu kısmında, gerekli izin alınarak Kezer'in (2013) gerçek bir uygulama sonucu elde etmiş olduğu madde cevap örüntüsü kullanılmıştır. Veri, Kezer (2013) tarafından Ankara Üniversitesi Yabancı Diller Yüksekokulu bünyesinde, 2012 – 2013 eğitim öğretim yılında hazırlık sınıfında öğrenim görmekte olan öğrencilerden toplanmıştır. Bu veri seti, İngilizce dilindeki okuduğunu anlama yeterlik alanına yönelik 80 maddelik madde havuzu ve 994 bireye ait 1-0 madde cevap örüntüsünden oluşmaktadır. PH simülasyonu öncesinde bu veri seti üzerinden 994 bireye ve 80 maddeye ait parametreler 3 PLM temelinde BSD yetenek kestirim yöntemi ile kestirilmiştir. BBST simülasyonunda ise bu parametrelerin yanında, gerçek madde cevap örüntüsünden yararlanılmış; tekrar bir veri seti üretmeye gerek duyulmamıştır.

Kezer (2013) çalışmasında hazırlık sınıfı öğrencilerine uygulanan bu test maddelerinin madde ayırt edicilik indekslerinin 0,36 ile 0,77; madde güçlük indekslerinin ise 0,05 ile 0,89 arasında olduğu sonucuna ulaşmıştır. Bu çalışmada PH simülasyonuna hazırlık aşamasında Klasik Test Kuramı'na dayalı şekilde maddelerin ayırt edicilik ve güçlük parametreleri tekrar hesaplanmıştır. Buna göre madde ayırt edicilik değerlerinin 0,33 ile 0,74 arasında; madde güçlük değerlerinin ise 0,08 ile 0,89 arasında olduğu görülmüştür.

Lau (1996) ile Spray ve diğerleri (1997) çalışmalarında, sınıflama kriterlerinin tek boyutluluk varsayımının sağlanamadığı durumlarda bile oldukça başarılı performans gösterdiklerini ifade etmektedir. Bununla birlikte PH simülasyon çalışmasına başlanmadan veri seti MTK varsayımları bakımından incelenmiştir. Bu amaçla verinin önce tek boyutluluk durumu R'da faktör analiziyle sınanmış; yerel bağımsızlık varsayımı için aynı yetenek düzeyindeki bireylere ait veri setleri üzerinden maddeler arası ilişkiler incelenmiş; testin hız testi olup olmama durumu irdelenmiştir. Ardından model-veri uyumunda -2 Log Likelihood (-2LL) değerleri

arasındaki fark Ki-kare ile test edilmiş; madde ve yetenek parametrelerinin değişmezlikleri incelenmiştir.

3.2.2.1. PH Veri Setinin Tek Boyutluluk Varsayımının İncelenmesi

Kezer (2013) verinin tek boyutluluk varsayımını Açımlayıcı Faktör Analizi (AFA) ve Doğrulayıcı Faktör Analizi (DFA) ile test etmiştir. AFA sonuçlarına göre, başat bir faktöre ait açıklanan varyans değerinin %30'dan büyük olması, yamaç birikinti grafiğinde bileşenlerin ivmelerine göre farkların anlamsız hale gelmesi, özdeğerler arasındaki farkın 1/3 oranından büyük olması gibi kriterler dikkate alınarak verinin tek boyutlu olduğuna karar verilmiştir (Lord, 1980).

AFA'nın ardından tek boyutlu yapının doğruluğunu test etmek amacıyla uygulanan DFA sonuçlarına göre ise uyum indekslerinin tek boyutluluğu onayladığı sonucuna ulaşılmıştır (Ki-Kare: 8732,30 ($p < 0,01$); Ki-Kare / sd oranı: 2,84; RMSEA: 0,043; CFI: 0,96; NFI: 0,92; NNFI: 0,96; GFI: 0,82 ve AGFI: 0,81). Bu çalışmanın PH simülasyonu bölümünde veri tek boyutlu MTK modellerine dayalı olarak simüle edilecektir. Bu nedenle veri setinin boyutluluğu tekrar incelenmiştir.

Tek boyutluluk, maddelerin ölçtüğü ve bireylerin cevaplama performanslarının altında yatan tek bir örtük özellik olması anlamına gelmektedir. Bir başka deyişle madde cevapları arasındaki varyansın tek bir örtük özellik tarafından açıklanmasıdır. Elbette kişilik, motivasyon, test alma becerisi gibi faktörler de bireyin test performansında etkilidir. Ancak tek boyutlulukla kastedilen maddelerin ölçtüğü başat bir boyutun olmasıdır (Hambleton ve Swaminathan, 1985). Bu çalışmada tek boyutluluk varsayımının incelenmesi amacıyla veriye R ortamında Revelle tarafından yazılan (2015) psych paketindeki "irt.fa" komutuyla tetrakorik korelasyonlara dayanan temel bileşenler faktör analizi uygulanmıştır. Faktör analizine geçmeden veri setinin analize uygunluğu incelenmiştir:

- Örneklem büyüklüğü açısından Comrey ve Lee (1992) faktör analizinde 200 deneğin uygun ve 500 deneğin oldukça yeterli olduğunu belirtmişlerdir (Akt: Tabachnick ve Fidell, 2007, s.613). Ayrıca genel bir kural olarak faktör analizi uygulanacak veri en az 300 kişiden oluşmalıdır (Tabachnick ve Fidell, 2007, s. 613). Bu çalışmada 994 deneğe ait cevaplar yer almaktadır. Buna göre örneklem büyüklüğünün analiz için yeterli olduğu söylenebilir.

- Örneklem büyüklüğünün yeterliğinden sonra veri setinde kayıp veri olup olmadığına bakılmıştır. Bu amaçla veri setine kayıp veri analizi uygulanmış ve veri setinde kayıp veri olmadığı görülmüştür.
- Kayıp veri analizinden sonra tek değişkenli normallik varsayımı incelenmiştir. Bu amaçla maddelerin betimsel istatistikleri hesaplanmıştır. Maddelerin bir kısmının çarpıklık ve basıklık katsayıları -1 ile +1 aralığının dışındadır. Ancak Tabachnick ve Fidell'e (2007, s. 613) göre faktör analizi gözlenen değişkenler arasındaki ilişkileri açıklamak için kullanıldığında dağılıma ilişkin varsayımlar geçerli değildir; değişkenler normalden sapıyorsa dahi çözüme devam edilmelidir. Buna göre analize devam edilmiştir
- Çok değişkenli normallik varsayımının incelenmesi amacıyla Bartlett küresellik testinden yararlanılmıştır. "psych" paketindeki "cortest.bartlett" komutuyla hesaplanan Bartlett Ki-kare değeri manidardır (Ki-Kare = 18373.37; sd = 3160; p = 0,00). Bartlett'e ait Ki-kare değerinin manidar çıkması verinin çok değişkenli normallik varsayımını sağladığı anlamına gelmektedir.
- Doğrusallık varsayımının kontrolü için tüm maddelere ait ikili saçılım diyagramlarına bakılması gerekmektedir. Ancak 80 maddenin tümüne ait diyagramın çizilmesi esnasında bilgisayarın hata vermesi sebebiyle maddeler arası korelasyonlar incelenmiştir. Çokluk ve diğerlerine göre (2010, s. 209) iki değişken arasındaki doğrusal ilişkinin varlığı doğrusal korelasyon katsayısı ile bulunur ve doğrusallığa ilişkin karar noktası örneklem büyüklüğüne bağlıdır. Örneklem sayısı 100 iken karar noktası 0,196'dır. Buna göre maddeler arası korelasyonların genel olarak manidar ve 0,196'dan büyük olması doğrusallık varsayımının sağlandığına işaret etmektedir. Ayrıca çok değişkenli normallik varsayımının sağlanması değişken çiftleri arasındaki ilişkilerin doğrusal olduğuna işaret eder (Çokluk ve diğerleri, 2010, s. 210). Buna göre veri setinin doğrusallık varsayımını sağladığı söylenebilir.
- Çoklu bağlanım ve tekillik varsayımlarının kontrolü amacıyla yine maddeler arası korelasyonlar incelenmiştir. Çokluk ve diğerlerine göre (2010, s. 210)

0,90 üstü bir korelasyon katsayısı çoklu bağlanım sorununa işaret ederken; 1,00 korelasyon katsayısı da tekillik anlamına gelmektedir. Buna göre veri setinde çoklu bağlanım ve/veya tekillik problemi olmadığı görülmüştür.

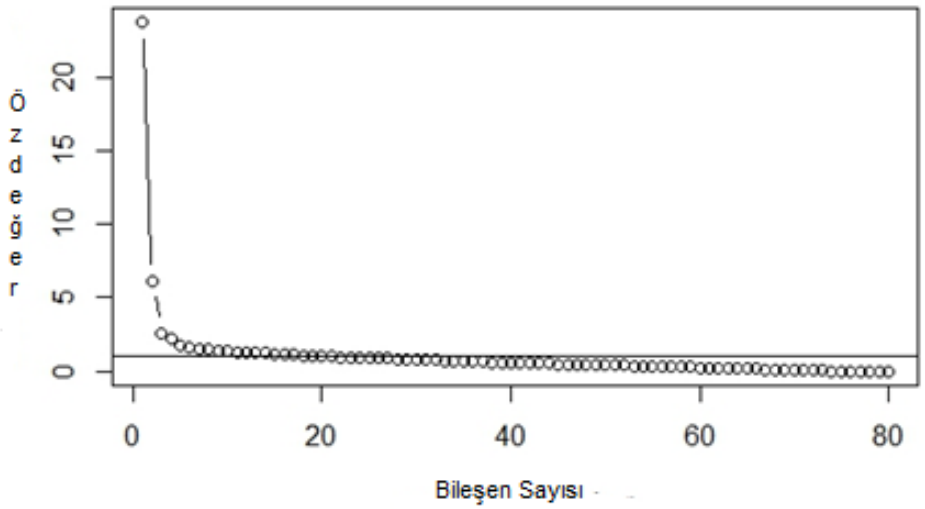
- Faktörleşebilirlik varsayımının kontrolü için hem anti-image korelasyon matrisi hem de "KMO" fonksiyonuyla hesaplanan değerler incelenmiş; veri setinin bu varsayımı da karşıladığı görülmüştür (KMO = 0,94 > 0,60).
- Tek yönlü uç değerlerin kontrolünde Z puanından yararlanılmıştır. Hesaplanan Z değerlerinin (-4,4) aralığında olduğu görüldüğünden bu varsayımın da karşılandığı söylenebilir.
- Çok yönlü uç değerler için Mahalanobis uzaklığı hesaplanmıştır. Madde sayısının bir eksiği olan serbestlik derecesi (79) ve 0,001 hata düzeyindeki kritik Ki-kare değeri 123,594'tür. Veride bu kritik değeri aşan herhangi bir bireyin olmaması ve diğer tüm varsayımların da karşılanması sebebiyle veri setinin tetrakorik korelasyon matrisine dayalı temel bileşenler faktör analizine uygun olduğu düşünülmüş ve analize geçilmiştir. Faktör analizi sonucu açıklanan varyans ve özdeğerler aşağıda Tablo 3.2.'de yer almaktadır.

Tablo 3.2: PH Veri Setinin Faktör Analizi Sonucu Açıklanan Varyans ve Özdeğerleri

Değişken	*Özdeğer*	*Varyans Oranı*	*Kümülatif Varyans Oranı*
1	23,74	0,30	0,30
2	5,39	0,07	0,37
3	1,87	0,02	0,39
4	1,50	0,02	0,39
5	1,02	0,01	0,40

Analiz sonuçlarına göre özdeğeri 1'in üzerinde olan ilk beş bileşendir ve Tablo 3.2.'de bu beş bileşene yer verilmiştir. Tablo 3.2.'ye göre birinci bileşenin özdeğerinden beşinci bileşenin özdeğerine doğru değerlerin hızla düştüğü görülmektedir. Birinci bileşene ait özdeğer 23,74 ve ikinci bileşene ait özdeğer 5,39'dur. İlk bileşenin özdeğerinin ikinci bileşenin özdeğerine oranı 4,59'dur. Bu oranın 3'ten büyük olması verinin tek faktörlü bir yapı gösterdiğine işaret etmektedir (Lord, 1980). Ayrıca bileşenlerin açıklanan varyansa katkıları incelendiğinde ikinci bileşenle birlikte kümülatif varyansın çok az arttığı; ilk bileşenin başat bir şekilde toplam varyansın %30'unu açıkladığı görülmektedir.

Bununla birlikte tek boyut için hesaplanan madde faktör yüklerinin 0,33 ile 0,76 arasında değiştiği görülmüştür (Ek 3). Aşağıda Şekil 5'te R ortamında "psych" paketindeki "vss.scree" komutuyla çizilen yamaç-birikinti grafiği incelendiğinde, ilk boyuttan sonra keskin bir düşüşün olduğu ve ikinci faktör ile birlikte diğer faktörlerin varyansa önemli bir katkı sağlamadıkları görülmektedir. Bu bulgulara dayanarak verinin tek boyutluluk varsayımını sağladığı söylenebilir.



Şekil 3.4. PH Verisine Ait Yamaç-Birikinti Grafiği

Tek boyutluluğun sınanması amacıyla veri setine ayrıca DFA da uygulanmıştır. Analiz sonucunda Ki-kare değeri 13112,72 ve serbestlik derecesi (sd) 3080 olarak hesaplanmıştır. Buna göre Ki-kare/sd oranı 4,25'tir ve bu oran 5'in altında olması sebebiyle orta düzeyde uyuma işaret etmektedir (Kline, 2011). Bunların yanında RMSEA = 0,072; standardize edilmiş RMR = 0,07; NNFI ve CFI değerleri ise 0,92 şeklinde elde edilmiştir. Tüm sonuçlar bir arada düşünüldüğünde verinin model uyumunun orta düzeyde-iyi olduğu yorumu yapılabilir (Çokluk ve diğerleri, 2010). Analiz sonuçlarına göre veri seti tek boyutluluk varsayımını sağlamaktadır.

3.2.2.2. PH Veri Setinin Yerel Bağımsızlık Varsayımının İncelenmesi

Kezer (2013) yerel bağımsızlık varsayımının incelenmesi amacıyla faktör analizlerinden sonra elde edilen 80 maddeye ilişkin artık korelasyon matrisi oluşturmuş ve ikililere ait korelasyon katsayılarının -0,19 ile 0,13 arasında olduğunu göstermiştir. Artık değerlere ilişkin korelasyon katsayılarının sıfıra yakın

çıkması sebebiyle tek boyutluluk varsayımının sağlanmasından dolayı yerel bağımsızlık varsayımının da karşılandığı kabul edilmiştir.

Bu çalışmada ise 994 öğrenci toplam puanlarına göre 4 yetenek grubuna ayrılmıştır. Yerel bağımsızlığın sağlanabilmesi için yüksek ve düşük yetenek grubu oluşturulup yetenek sabitlendiğinde, maddeler arasında yüksek ve manidar ilişki bulunmaması gerekmektedir. Tüm yetenek grupları için hesaplanan Pearson korelasyon katsayılarına göre maddeler arasındaki korelasyonların büyük bir çoğunluğunun anlamsız; manidar olanların ise düşük düzeyde ilişkili olduğu görülmüştür. Buna göre maddeler arasında yerel bağımsızlık varsayımının büyük oranda sağlandığı söylenebilir. Ayrıca yukarıda da bahsedildiği gibi test tek boyutluluk özelliği gösteriyorsa bu testte yer alan maddelerin yerel bağımsızlık özelliğine de sahip olduğu söylenebilir. Buna göre yapılan faktör analizi sonucunda testin tek boyutluluk özelliği göstermesi, aynı zamanda testte yer alan maddelerin yerel bağımsız olduğu seklinde yorumlanabilir.

3.2.2.3. Testin Hız Testi Olmaması

Testin hız testi olup olmaması cevaplanmamış madde sayısı, testi bitirememiş birey sayısının yanında paralel test uygulaması gibi yöntemlerle kontrol edilebilmektedir (Hambleton ve Swaminathan, 1985, s. 157-161). Veri seti incelendiğinde cevaplanmamış madde olmadığı ve testi bitirememiş birey olmadığı görülmektedir. Paralel test uygulaması ise yapılmamıştır. Buna göre testin hız testi olmadığı, buna dayanarak en temel varsayım olan tek boyutluluk varsayımını etkileyecek bir durum olmadığı ve veri setinin MTK varsayımlarını karşıladığı söylenebilir. Bu noktadan sonra model-veri uyumunun incelenmesi gerekmektedir.

3.2.2.4. PH Veri Setinin Model-Veri Uyumunun İncelenmesi

Post-Hoc çalışması için model-veri uyumunun incelenmesi amacıyla veri seti 1 PLM, 2 PLM ve 3 PLM ile analiz edilmiş; -2LL değerleri Ki-kare dağılımı yardımıyla karşılaştırılmıştır. Aşağıda Tablo 3.3.'te 1 PLM, 2 PLM ve 3 PLM'de hesaplanan -2LL değerleri yer almaktadır.

Tablo 3.3: Modellere Ait -2LL Değerleri

Model	*-2LL*
1 PLM	73880,0808
2 PLM	73158,4672
3 PLM	72824,8051

Tablo 3.3.'e göre modellerin -2LL değerleri 1 PLM'den 3 PLM'e doğru düşmektedir. Buna göre öncelikle 1 PLM – 2 PLM arasındaki fark Ki-kare ile test edilmiştir:

- $X^2 = -2LL_{1PLM} – (-2LL_{2PLM}) = 73880,0808 – 73158,4672 = 721,6136$ ve serbestlik derecesi modele eklenen parametre sayısı olan 80'dir. Ki-kare tablosundan kritik $X^2_{0,001;\ 80} = 124,839$'dur. Buna göre hesaplanan ki-kare değeri, kritik değerden büyük olduğu için 2 PLM'in daha uygun olduğu söylenebilir. Bir başka deyişle 1 PLM'e a parametresi eklenmesi durumunda (modelin 2 PLM olması durumunda) eklenen parametreler anlamlı bir farklılık yaratmaktadır (Embretson ve Resie, 2000, s. 74).

Hangi modelin daha uyumlu olduğuna karar verebilmek amacıyla 2 PLM ile 3 PLM arasındaki fark da test edilmelidir:

- $X^2 = -2LL_{2PLM} – (-2LL_{3PLM}) = 73158,4672 – 72824,8051 = 333,6621$ ve serbestlik derecesi modele eklenen parametre sayısı olan 80'dir. Ki-kare tablosundan kritik $X^2_{0,001;\ 80} = 124,839$'dur. Buna göre hesaplanan ki-kare değeri, kritik değerden büyük olduğu için 3 PLM'in daha uygun olduğu söylenebilir. Bir başka deyişle 2 PLM'e c parametresi eklenmesi durumunda (modelin 3 PLM olması durumunda) eklenen parametreler anlamlı bir farklılık yaratmaktadır (Embretson ve Resie, 2000, s. 74).

Bu sonuçlara dayanarak veri setinin 3 PLM'e daha uygun olduğu söylenebilir. Bu çalışmada da veri 3 PLM temelinde simüle edilmiştir. 3 PLM ile kestirilen madde parametrelerinin betimsel özellikleri Ek 4'te yer almaktadır. Ek 4'e göre madde ayırt edicilik (a) parametresi değerleri 0,63 ile 2,08 arasında, ortalaması 1,12 ve standart sapması 0,30; madde güçlük (b) parametresi değerleri -1,69 ile 2,55 arasında, ortalaması 0,91 ve standart sapması 1,11; şans başarısı (c) parametresi değerleri 0,05 ile 0,25 arasında, ortalaması 0,11 ve standart sapması 0,05'tir. 3 PLM ile kestirilen 994 bireye ait yetenek parametrelerinin betimsel özellikleri ise Ek 5'te yer almaktadır. Ek 5'e göre kestirilen en düşük yetenek düzeyi -2,58; en yüksek yetenek düzeyi 3,13; yetenek düzeyleri ortalaması -0,02 ve standart sapması 0,99'dur.

3.2.2.5. PH Veri Setinin Madde ve Yetenek Parametrelerinin Değişmezliğinin İncelenmesi

Madde parametrelerinin değişmezliği, öğrencilere ait farklı alt gruplar oluşturularak incelenmektedir. Alt gruplar, kız-erkek öğrencilerden oluşabileceği gibi seçkisiz oluşturulmuş gruplarla veya yüksek-düşük performans gösteren öğrencilerle de oluşturulabilmektedir (Hambleton ve Swaminathan, 1985, s. 158). Bu çalışmada madde parametrelerinin değişmezliği seçkisiz atanan gruplarda incelenmiştir. Bu amaçla öncelikle 994 birey iki alt gruba seçkisiz olarak atanmış; 3 PLM ile madde parametreleri ayrı ayrı kestirilmiş; kestirilen parametrelerin betimsel özellikleri dikkate alınarak Pearson Momentler Çarpımı Korelasyon Katsayısı ile parametreler arasındaki ilişki belirlenmiştir. Aşağıda Tablo 3.4.'te seçkisiz atanan bu iki alt gruptan ayrı ayrı kestirilen madde parametreleri arasındaki ilişkiyi gösteren korelasyon katsayıları yer almaktadır.

Tablo 3.4: Seçkisiz Atanan İki Alt Gruptan Kestirilen Parametreler Arasındaki İlişki

	a parametresi	***b parametresi***	***c parametresi***
Grup 1 – Grup 2	0,752*	0,974*	0,872*

*p < 0,01

Tablo 3.4.'te seçkisiz atanan iki grup üzerinden kestirilen madde parametreleri arasında, üç parametre için de 0,01 hata ile anlamlı ve kabul edilebilir düzeyde ilişkiler olduğu görülmektedir. İki alt gruptan kestirilen a (ayırt edicilik) parametreleri arasında 0,75 korelasyon değeriyle orta-üstü düzeyde manidar bir ilişki; b (güçlük) parametreleri arasında 0,97 korelasyon değeriyle oldukça yüksek düzeyde manidar bir ilişki ve c (şans) parametreleri arasında ise 0,87 korelasyon değeriyle yine yüksek düzeyde manidar bir ilişki olduğu görülmektedir. Bu bulgu seçkisiz gruplardan kestirilen madde parametrelerinin iki grup için de benzer olduğunu göstermektedir.

Parametrelerden a parametresi diğer parametrelere kıyasla daha düşük bir korelasyon değerine sahiptir. Bu durumun sebebi, a parametresinin puanların normalliğinden diğer parametrelere kıyasla daha fazla etkileniyor oluşuyla açıklanabilir (Kelecioğlu, 2011). b parametresi için hesaplanan korelasyonun diğer parametreler için elde edilen korelasyon değerlerinden daha yüksek bir değer alması; bir başka deyişle alt gruplardan kestirilen b parametreleri arasındaki ilişkinin oldukça yüksek olması ise, bu parametrenin puan dağılımından

etkilenmediğine işaret etmektedir (Gelbal, 1994). Sonuç olarak seçkisiz gruplardan elde edilen bu bulguya dayanarak 3 PLM ile madde parametrelerinin değişmezliğinin sağlandığı söylenebilir.

Bu çalışmada yetenek parametrelerinin değişmezliği, maddelerin seçkisiz iki gruba ayrılması ve bireylerin yeteneklerinin bu gruplardan ayrı ayrı kestirilmesi yoluyla incelenmiştir. Bu amaçla 80 madde seçkisiz olarak iki alt gruba atanmış; birey yetenekleri bu alt gruplardan 3 PLM ile analiz yapılarak ve BSD yetenek kestirim yöntemi kullanılarak kestirilmiş; kestirilen yeteneklerin dağılımı dikkate alınarak aralarındaki ilişkiye Pearson Momentler Çarpımı Korelasyon Katsayısı ile bakılmıştır. Analiz sonuçlarına göre iki madde grubundan kestirilen yetenekler arasında 0,01 hata düzeyinde ve 0,88 değerinde yüksek bir ilişki olduğu görülmüştür. Elde edilen bu bulguya göre yetenek parametrelerinin değişmezliğinin de sağlandığı söylenebilir.

3.3. BBST Simülasyonu Koşulları

Çalışmada verinin türetilmesi-elde edilmesinden ve MTK varsayımları bakımından incelenmesinden sonra her bir alt problem kapsamında yer alan koşullar için BBST simülasyonuna geçilmiştir. BBST simülasyonunda Nydick (2014) tarafından yazılan catirt paketinden yararlanılmıştır. Simülasyonda tüm koşullarda başlama noktası, yetenek düzeyi 0 olarak belirlenmiş ve her koşul için bu değer sabit tutulmuştur. Ayrıca tüm koşullar için 25 tekrar yapılmıştır.

Araştırmada MC ve PH olmak üzere iki farklı simülasyon çalışması yapılmış ve buna bağlı olarak iki farklı veri seti kullanılmıştır. Bunun dışında madde seçme yöntemleri, yetenek kestirim yöntemleri ve sınıflama kriterleri çalışmanın amacına uygun şekilde manipüle edilmiştir. Çalışmanın odak noktası olan sınıflama kriterleri, 0,05 ve 0,10 farksızlık bölgesi değerleri ile Ardışık Olabilirlik Oranı Testi (AOOT) ve Genelleştirilmiş Olabilirlik Oranı (GOO); %70 ve %90 güven aralığı düzeylerini içeren Güven Aralığı (GA) yöntemleridir. Çalışmada yer alan yetenek kestirim yöntemleri Beklenen Sonsal Dağılım (BSD) ve Ağırlıklandırılmış Olabilirlik Kestirim (AOK) yöntemleridir. Çalışmada incelenen madde seçme yöntemleri ise, kestirilen yetenek temelli MFB (MFB-KY), kesme noktası temelli MFB (MFB-KN), kestirilen yetenek temelli KLB (KLB-KY) ve kesme noktası temelli KLB (KLB-KN) şeklinde belirlenmiştir.

3.4. Verilerin İşlenmesi ve Çözümlenmesi

Tekrarlar sonucu, her bir koşula ait ortalama test uzunluğu (OTU) ve ortalama sınıflama doğruluğu (OSD) oranları hesaplanmıştır. Bunların yanında gerçek yetenekler ve kestirilen yetenekler arasındaki ilişki Pearson Momentler Çarpımı Korelasyon Katsayısıyla (r) ve yine her bir koşul için tekrarların ortalaması alınarak elde edilmiştir. Ayrıca yanlılık, RMSE (root mean squared error) ve ortalama mutlak hata (OMH) kriterleri R ortamında fonksiyonlar yazılarak hesaplanmıştır. Sınıflama kriterlerini karşılaştırmada, çalışma kapsamındaki tüm koşullar için yukarıda belirtilen ölçütler esas alınarak ve bunların 25 tekrarının ortalaması alınarak analiz sonuçları yorumlamaya gidilmiştir.

Simülasyon sonuçlarından ortalama test uzunluğu $test_length koduyla; bireylerin simülasyon sonunda atandıkları sınıflar ise $cat_categ koduyla çekilmiştir. Ortalama sınıflama doğruluğunu hesaplayabilmek amacıyla, simülasyondan çekilen sınıflarla bireylerin türetilen gerçek sınıfları arasındaki uyuma Cohen'in Kappa istatistiğiyle bakılmıştır.

Yanlılık, BBST simülasyonu sonucu her birey için kestirilen son yetenek düzeyinin ($\widehat{\theta\imath}$) bireyin gerçek yetenek düzeyinden (θ_i) farkının tüm bireyler üzerinden toplamının öğrenci sayısına oranından elde edilmiştir. Yanlılığın formülü aşağıdaki gibidir (Miller ve Miller, 2004):

$$\text{Yanlılık} = \frac{\sum_{i=1}^{n}(\widehat{\theta\imath} - \theta_i)}{n} \tag{18}$$

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

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{\theta\imath} - \theta_i)^2}{n}} \tag{19}$$

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

$$\text{OMH} = \frac{\sum_{i=1}^{n}|\widehat{\theta\imath} - \theta_i|}{n} \tag{20}$$

4. BULGULAR VE TARTIŞMA

Bu bölümde Monte Carlo ve Post Hoc simülasyon çalışmaları için ayrı ayrı olmak üzere, her bir alt probleme yönelik elde edilen bulgulara ve bulguların alanyazın desteğiyle yorumlanmasına yer verilmiştir.

4.1. Birinci Alt Probleme Ait Bulgular ve Yorumlar

Araştırmanın birinci alt probleminde Monte Carlo simülasyonu BBST uygulamasında, yetenek kestirim yöntemi BSD olduğunda, AOOT sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GOO sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GA sınıflama kriterinin %70 ile %90 güven düzeyleri için sınıflama doğruluğu, test uzunluğu ve ölçme kesinliğinin madde seçme yöntemlerinden MFB-KY, MFB-KN, KLB-KY ve KLB-KN'ye göre nasıl değiştiği incelenmiştir.

Aşağıda Tablo 4.1'de bu alt problemde belirtilen koşulları karşılaştırmada kullanılan bağımlı değişkenler olan ortalama test uzunluğu (OTU), ortalama sınıflama doğruluğu (OSD), bireylerin gerçek yetenek düzeyleri ile kestirilen son yetenek düzeyleri arasındaki korelasyon (r), yanlılık, RMSE ve ortalama mutlak hataya (OMH) ait değerler yer almaktadır. Değerlerin tümü her koşul için 25 tekrarın ortalaması alınarak hesaplanmıştır.

Tablo 4.1'de madde seçme yöntemi fark etmeksizin ortalama test uzunluğu (OTU) bakımından testi sonlandırmada, bir başka deyişle bireyleri sınıflamada, en az madde gerektiren sınıflama kriterinin GA yönteminin %70 düzeyi olduğu; bunu takiben sırasıyla GA yönteminin %90 güven düzeyinin; GOO FB: 0,10 yönteminin; GOO FB: 0,05 yönteminin ve AOOT FB: 0,10 yönteminin geldiği; en fazla madde gerektiren sınıflama kriterinin ise AOOT FB: 0,05 yönteminin olduğu görülmektedir. Bu bulguya göre, madde seçme yöntemi ile sınıflama kriterlerinin farksızlık bölgesi değerleri ya da güven düzeyleri fark etmeksizin, OTU bakımından en iyi performansı GA sınıflama kriteri göstermiş; bunu sırasıyla GOO ve AOOT sınıflama kriterleri takip etmiştir.

Tablo 4.1: Yetenek Kestirim Yöntemi BSD Olduğunda Koşullara Ait OTU, OSD, r, Yanlılık, RMSE ve OMH Değerleri

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

Tablo 4.1'de sınıflama kriterlerinden GOO ve AOOT yöntemlerinin farksızlık bölgesi değerleri küçüldükçe ve GA yönteminin güven düzeyi yükseldikçe bireyleri sınıflamada gerekli ortalama madde sayısı olan OTU değerlerinin arttığı görülmektedir. Araştırmanın bu bulgusu Reckase (1983), Lau ve Wang (1999), Thompson ve Ro (2007) ve Thompson'ın (2011) araştırma sonuçlarıyla örtüşmektedir.

Tablo 4.1'de, madde seçme yöntemi fark etmeksizin, ortalama sınıflama doğruluğu (OSD) bakımından GA yöntemi dışındaki sınıflama kriterlerinin tümünün öğrencileri geçti-kaldı kategorilerine sınıflamada benzer performans gösterdiği ve sınıflama doğruluğunun oldukça yüksek (0,95 ile 0,97 aralığında) olduğu görülmektedir. GA yönteminin %70 güven düzeyinin (0,95) ve %90 güven düzeyinin (0,96) diğer yöntemlere kıyasla daha düşük ancak yine de yüksek bir sınıflama katsayısına sahip olduğu görülmektedir.

Tablo 4.1'de bireylerin simülasyon öncesi türetilen gerçek yetenek düzeyleri ile BBST simülasyonu sonucu kestirilen son yetenek düzeyleri arasındaki korelasyon (r) bakımından en yüksek ilişkinin hesaplandığı yöntemin, madde seçme yöntemi fark etmeksizin, AOOT FB: 0,05 yöntemi olduğu; bunu takiben sırasıyla AOOT FB: 0,10 yönteminin, GOO yöntemlerinin ve GA %90 yönteminin geldiği; en düşük ilişkinin ise GA %70 yöntemi ile hesaplandığı görülmektedir. Bu korelasyonlar madde seçme yöntemlerine göre ayrı ayrı incelendiğinde en iyi performansı MFB-KY ve KLB-KY yöntemlerinin verdiği görülmektedir. Bu iki yöntem için hesaplanan korelasyonlar 0,86 ile 0,98 değerleri arasında değişmekte; bu durum da gerçek yeteneklerle kestirilen yetenekler arasındaki korelasyonların bu iki madde seçme yöntemi kullanıldığında oldukça yüksek olduğunu göstermektedir. Diğer madde seçme yöntemlerine ait korelasyon değerleri ise MFB-KN için 0,75 ile 0,92 aralığında ve KLB-KN için 0,75 ile 0,91 aralığında değişmektedir. Bu bulguya dayanarak, bireylerin gerçek yetenek düzeyleri ile kestirilen son yetenek düzeyleri arasındaki korelasyon bakımından, kestirilen yetenek (KY) temelli madde seçme yöntemlerinin kesme noktası (KN) temelli madde seçme yöntemlerine kıyasla daha iyi performans gösterdiği yorumu yapılabilir.

Tablo 4.1'de sınıflama kriterlerinin oldukça düşük yanlılık değerlerine sahip olduğu ve performanslarının madde seçme yöntemlerine göre farklılık gösterdiği görülmektedir. Madde seçme yöntemi MFB-KY olduğunda sınıflama kriterlerinin

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

Tablo 4.1'e göre, yanlılık ile birlikte kestirimin standart hatasını da dikkate alan RMSE ve ortalama mutlak hata (OMH) değerlerine göre, madde seçme yöntemi fark etmeksizin, en iyi performans gösteren sınıflama kriteri AOOT FB: 0,05 yöntemidir. Bunu takiben sırasıyla AOOT FB: 0,10; GOO FB: 0,05; GOO FB: 0,10; GA %90 ve GA %70 yöntemlerinin geldiği görülmektedir. Farksızlık bölgesi ve güven düzeyleri fark etmeksizin RMSE ve OMH bakımından en iyi performansı AOOT yöntemi göstermiştir. AOOT yönteminin ardından GOO yöntemi gelmiş ve göreli olarak en kötü performansı ise GA yöntemi göstermiştir.

Tablo 4.1'de görülüğü üzere, Monte Carlo simülasyonu BBST uygulamasında yetenek kestirim yöntemi BSD olduğunda, madde seçme yöntemi fark etmeksizin, test etkililiği (ortalama test uzunluğu ve ortalama sınıflama doğruluğu) bakımından GA %70 yönteminin diğerlerine kıyasla oldukça başarılı bir performans gösterdiği ortaya çıkmıştır. Bunu sırasıyla GA %90; GOO FB: 0,10; GOO FB: 0,05; AOOT FB: 0,10 ve AOOT FB: 0,05 sınıflama kriterleri izlemektedir. Bireyler geçti-kaldı kategorilerine GA %70 sınıflama kriteri ile ortalama 8 madde ve ortalama 0,95 sınıflama doğruluğuyla sınıflanabilirken; diğer yöntemlerin tümünde ortalama sınıflama doğruluğu 0,97 olmak üzere GA %90 sınıflama kriteri ile ortalama 14 madde; GOO FB: 0,10 sınıflama kriteri ile ortalama 16 madde; GOO FB: 0,05 sınıflama kriteri ile ortalama 17 madde; AOOT FB: 0,10 sınıflama kriteri ile ortalama 32 madde ve AOOT FB: 0,05 sınıflama kriteri ile de ortallama 47 maddeyle testin sonlanabildiği ve bireylerin kategorilere yerleştirilebildiği görülmüştür. Test etkililiği açısından bakıldığında GA ve GOO sınıflama kriterlerinin AOOT'ye kıyasla başarılı performans gösterdikleri görülmektedir. Bununla birlikte kestirilen ve gerçek yetenek düzeyleri arasındaki ilişki (r), yanlılık, RMSE ve OMH değerleri bakımından AOOT sınıflama kriterinin göreli olarak diğer yöntemlerden daha iyi performans gösterdiği; ancak tüm koşullar içinden bu göreli değerlerin en düşük olduğu MFB-KY madde seçme yönteminin ve AOOT FB: 0,05

sınıflama kriterinin birlikte ele alındığı koşulda, GA %70 sınıflama kriterinin neredeyse 6 katı maddede testin sonlandığı-bireylerin sınıflanabildiği dikkat çekmektedir.

4.2. İkinci Alt Probleme Ait Bulgular ve Yorumlar

Araştırmanın ikinci alt probleminde Monte Carlo simülasyonu BBST uygulamasında, yetenek kestirim yöntemi AOK olduğunda, AOOT sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GOO sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GA sınıflama kriterinin %70 ile %90 güven düzeyleri için sınıflama doğruluğu, test uzunluğu ve ölçme kesinliğinin madde seçme yöntemlerinden MFB-KY, MFB-KN, KLB-KY ve KLB-KN'ye göre nasıl değiştiği incelenmiştir.

Aşağıda Tablo 4.2'de bu alt problemde belirtilen koşulları karşılaştırmada kullanılan bağımlı değişkenler olan ortalama test uzunluğu (OTU), ortalama sınıflama doğruluğu (OSD), bireylerin gerçek yetenek düzeyleri ile kestirilen son yetenek düzeyleri arasındaki korelasyon (r), yanlılık, RMSE ve ortalama mutlak hataya (OMH) ait değerler yer almaktadır. Değerlerin tümü her koşul için 25 tekrarın ortalaması alınarak hesaplanmıştır.

Tablo 4.2'de madde seçme yöntemlerinden MFB-KY ve KLB-KY kullanıldığında ortalama test uzunluğu (OTU) bakımından testi sonlandırmada, bir başka deyişle bireyleri sınıflamada, en az madde gerektiren sınıflama kriterinin GA yönteminin %70 düzeyi olduğu; bunu takiben sırasıyla GA yönteminin %90 güven düzeyinin; GOO FB: 0,10 yönteminin; GOO FB: 0,05 yönteminin ve AOOT FB: 0,10 yönteminin geldiği; en fazla madde gerektiren sınıflama kriterinin ise AOOT FB: 0,05 yönteminin olduğu görülmektedir. Madde seçme yöntemlerinden MFB-KN ve KLB-KN kullanıldığında ise OTU bakımından iyi performans gösteren sınıflama kriterlerinin GA %70; GOO FB: 0,10 yöntemi; GOO FB: 0,05 yöntemi; GA yönteminin %90 güven düzeyi; AOOT FB: 0,10 yöntemi ve AOOT FB: 0,05 yöntemi şeklinde sıralandığı görülmektedir. Bu bulguya göre OTU bakımından en iyi performansı KY temelli madde seçme yöntemleri kullanıldığında GA sınıflama kriteri göstermiş, bunu sırasıyla GOO ve AOOT sınıflama kriterleri takip etmiş iken; KN temelli madde seçme yöntemlerinde GA yönteminin %90 güven düzeyinin GOO yöntemine kıyasla daha kötü performans gösterdiği ve yine bireyleri

sınıflamada en fazla sayıda madde gerektiren sınıflama kriterinin AOOT olduğu görülmüştür.

Tablo 4.2'de sınıflama kriterlerinden GOO ve AOOT yöntemlerinin farksızlık bölgesi değerleri küçüldükçe ve GA yönteminin güven düzeyi yükseldikçe bireyleri sınıflamada gerekli ortalama madde sayısı olan OTU değerlerinin arttığı görülmektedir. Araştırmanın bu bulgusu Reckase (1983), Lau ve Wang (1999), Thompson ve Ro (2007) ve Thompson'ın (2011) araştırma sonuçlarıyla örtüşmektedir.

Tablo 4.2'de madde seçme yöntemi fark etmeksizin ortalama sınıflama doğruluğu (OSD) bakımından sınıflama kriterlerinin tümünün öğrencileri geçti-kaldı kategorilerine sınıflamada benzer performans gösterdiği ve sınıflama doğruluğunun 0,95 ile 0,97 aralığında oldukça yüksek olduğu görülmektedir. Madde seçme yöntemlerinden MFB-KY veya KLB-KY kullanıldığında sınıflama kriterlerinden GA %70 ve GA %90 yöntemlerinin; MFB-KN kullanıldığında GA sınıflama kriterinin %70 güven düzeyinin; KLB-KN kullanıldığında ise GA %70 yöntemi ile GOO FB: 0,10 yönteminin diğer sınıflama kriterlerine kıyasla düşük OSD değerlerine sahip olsalar da sınıflama doğruluklarının yüksek olduğu görülmüştür.

Tablo 4.2'de gerçek yetenek düzeyleri ile kestirilen yetenek düzeyleri arasındaki korelasyon (r) bakımından en yüksek ilişkinin hesaplandığı yöntemin MFB-KY yöntemi kullanıldığında AOOT FB: 0,05 yöntemi olduğu; bunu takiben sırasıyla AOOT FB: 0,10 yönteminin, GOO yöntemlerinin ve GA %90 yönteminin geldiği; en düşük ilişkinin ise GA %70 yöntemi ile hesaplandığı görülmektedir. Tüm sınıflama kriterlerine ait r değerleri KY temelli yöntemler için 0,85 ile 0,98 aralığında ve KN temelli yöntemler için de 0,76 ile 0,91 aralığındadır. Bu bulguya göre r bakımından KY temelli madde seçme yöntemlerinin KN temelli madde seçme yöntemlerine kıyasla daha iyi performans gösterdiği ortaya koyulmuştur.

Tablo 4.2: Yetenek Kestirim Yöntemi AOK Olduğunda Koşullara Ait OTU, OSD, r, Yanlılık, RMSE ve OMH Değerleri

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

Tablo 4.2'de sınıflama kriterlerinin Tablo 4.1'e kıyasla daha yüksek yanlılık değerlerine sahip olduğu; bir başka deyişle yetenek kestirim yöntemi olarak BSD yerine AOK kullanıldığında yanlılık değerlerinin yükseldiği ve yöntemlerin performanslarının madde seçme yöntemlerine göre farklılık gösterdiği görülmektedir. Madde seçme yöntemi fark etmeksizin AOOT sınıflama kriterinin en düşük yanlılık değerlerine sahip olduğu görülmektedir. Bununla birlikte GA yönteminin %90 güven düzeyinin, madde seçme yöntemi MFB-KY veya KLB-KY olduğunda GOO FB: 0,10 yönteminden daha düşük yanlılık değerine; madde seçme yöntemi MFB-KN veya KLB-KN olduğunda ise GOO yönteminin her iki farksızlık bölgesine kıyasla daha az yanlılığa sahip olduğu görülmektedir. Buna dayanarak GA %90 sınıflama kriterinin madde seçme yöntemleriyle birlikte yanlılık bakımından iyi sonuçlar verdiği ve özellikle KN temelli madde seçme yöntemleriyle ele alındığında oldukça düşük yanlılık değeri verdiği söylenebilir.

Tablo 4.2'de yanlılık ile birlikte kestirimin standart hatasını da dikkate alan RMSE ve ortalama mutlak hata (OMH) değerlerine göre madde seçme yöntemi fark etmeksizin, en iyi performans gösteren sınıflama kriteri AOOT FB: 0,05 yöntemidir. Bununla birlikte GA yönteminin %90 güven düzeyinin, madde seçme yöntemi MFB-KN veya KLB-KN olduğunda, GOO yönteminin her iki farksızlık bölgesine kıyasla daha düşük RMSE ve OMH değerine sahip olduğu görülmektedir. Buna dayanarak GA %90 sınıflama kriterinin özellikle KN temelli madde seçme yöntemleriyle ele alındığında oldukça başarılı performans gösterdiği söylenebilir.

Tablo 4.2'de görülüğü üzere BBST simülasyonunda yetenek kestirim yöntemi AOK olduğunda, madde seçme yöntemi fark etmeksizin test etkililiği (ortalama test uzunluğu ve ortalama sınıflama doğruluğu) bakımından GA %70 yönteminin diğerlerine kıyasla oldukça başarılı bir performans gösterdiği görülmektedir. Bunu sırasıyla KY temelli madde seçme yöntemlerinde GA %90; GOO FB: 0,10; GOO FB: 0,05; AOOT FB: 0,10 ve AOOT FB: 0,05 ve KN temelli madde seçme yöntemlerinde ise GOO FB: 0,10; GOO FB: 0,05; GA %90; AOOT FB: 0,10 ve AOOT FB: 0,05 sınıflama kriterleri izlemektedir.

Tablo 4.2'de bireyler geçti-kaldı kategorilerine GA %70 sınıflama kriteri ile ortalama 9 madde ve ortalama 0,96 sınıflama doğruluğuyla sınıflanabilirken; diğer yöntemlerin tümünde ortalama sınıflama doğruluğu 0,96 ya da 0,97 olmak üzere GA %90 sınıflama kriteri ile ortalama 16 madde; GOO FB: 0,10 sınıflama kriteri ile

ortalama 16 madde; GOO FB: 0,05 sınıflama kriteri ile ortalama 17 madde; AOOT FB: 0,10 sınıflama kriteri ile ortalama 32 madde ve AOOT FB: 0,05 sınıflama kriteri ile de ortalama 47 maddeyle testin sonlanabildiği ve bireylerin kategorilere yerleştirilebildiği görülmüştür. Test etkililiği açısından bakıldığında GA ve GOO sınıflama kriterlerinin AOOT'ye kıyasla başarılı performans gösterdikleri söylenebilir. Bununla birlikte kestirilen ve gerçek yetenek düzeyleri arasındaki ilişki (r), yanlılık, RMSE ve OMH değerleri bakımından AOOT sınıflama kriterinin göreli olarak diğer yöntemlerden daha iyi performans gösterdiği; ancak tüm koşullar içinden bu göreli değerlerin en düşük olduğu KLB-KY madde seçme yönteminin ve AOOT FB: 0,05 sınıflama kriterinin birlikte ele alındığı koşulda, GA %70 yönteminin neredeyse 6 katı maddede testin sonlandığı-bireylerin sınıflanabildiği dikkat çekmektedir.

4.3. Üçüncü Alt Probleme Ait Bulgular ve Yorumlar

Araştırmanın üçüncü alt probleminde Monte Carlo simülasyonu BBST uygulamasında, bağımsız değişkenlerden sınıflama kriterleri, ortalama test uzunluğu (OTU), ortalama sınıflama doğruluğu (OSD), korelasyon (r), yanlılık, RMSE ve ortalama mutlak hata (OMH) bakımından incelenmiştir. Aşağıda Tablo 4.3'te sınıflama kriterlerinin bağımlı değişkenler bakımından nasıl değiştiği görülmektedir.

Tablo 4.3: Sınıflama Kriterlerinin OTU, OSD, r, Yanlılık, RMSE ve OMH Değerleri

Bağımsız Değişken	*OTU*	*OSD*	*r*	*Yanlılık*	*RMSE*	*OMH*
AOOT	39,45	0,97	0,93	0,039	0,311	0,257
GOO	16,36	0,97	0,85	0,079	0,475	0,394
GA	11,24	0,96	0,82	0,085	0,523	0,436

Tablo 4.3'e göre sınıflama kriterlerinden, yöntemlerin farklı hata düzeyleri veya farksızlık bölgesi değerleri dikkate alınmaksızın, en az sayıda maddeyle sınıflama yapabilen yöntemin yaklaşık 11 maddeyle GA yöntemi olduğu; bunu 16 maddeyle GOO yönteminin takip ettiği ve en çok sayıda maddeyle sınıflama yapabilen yöntemin de 40 maddeyle AOOT olduğu görülmektedir. Sınıflama kriterlerinin ortalama sınıflama doğruluğu bakımından ise benzer sonuçlar verdikleri görülmektedir. Bu bulgu, test etkililiği bakımından GA ve GOO yöntemlerinin BBST uygulamalarında daha kullanışlı olacağına işaret etmektedir.

Tablo 4.3'te bireylerin türetilen gerçek yetenek düzeyleri ile BBST uygulaması sonucu kestirilen son yetenek düzeyleri arasındaki korelasyon (r) bakımından sınıflama kriterlerinin üçünün de iyi performans gösterdikleri görülmektedir. Korelasyonlar bakımından en iyi sonucu AOOT sınıflama kriteri verirken; GOO ve GA yöntemlerinin benzer şekilde çalıştıkları görülmüştür. Bu sonuç test ekililiği ile birlikte düşünüldüğünde, GOO ve GA sınıflama kriterlerinin uygulamada avantaj sağlayacağı yorumu yapılabilir.

Tablo 4.3'te yanlılık, RMSE ve OMH bakımından GA sınıflama kriterinin diğer iki yönteme kıyasla daha kötü performans gösterdiği; en iyi performansı ise AOOT sınıflama kriterinin sergilediği görülmektedir. Bu sonuçlar test etkililiği ile birlikte düşünüldüğünde, AOOT sınıflama kriteri sonuçlarının hatadan daha arınık olmasına karşın; bu yöntemin test etkililiği bakımından kullanışlı olmadığı dikkat çekmektedir.

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

4.4. Dördüncü Alt Probleme Ait Bulgular ve Yorumlar

Araştırmanın dördüncü alt probleminde Monte Carlo simülasyonu BBST uygulamasında, bağımsız değişkenlerden madde seçme yöntemleri, ortalama test uzunluğu (OTU), ortalama sınıflama doğruluğu (OSD), korelasyon (r), yanlılık, RMSE ve ortalama mutlak hata (OMH) bakımından incelenmiştir. Aşağıda Tablo 4.4'te madde seçme yöntemlerinin bağımlı değişkenler bakımından nasıl değiştiği görülmektedir.

Tablo 4.4'e göre, madde seçme yönteminin hangi ölçütü temele aldığı fark etmeksizin, MFB ve KLB madde seçme yöntemlerinin bağımlı değişkenler bakımından birbirine oldukça benzer performans gösterdiği görülmektedir. Bu bulgu Eggen (1999), Lau ve Wang (1999), Cheng ve Liou (2000) ile Lin ve Spray'in (2000) çalışma sonuçlarına benzerlik göstermekte iken; Spray ve

Reckase (1994) ile Lau ve Wang'ın (1998) araştırma sonuçlarıyla örtüşmemektedir. Alanyazın incelendiğinde bu iki madde seçme yönteminin performansları hakkında araştırmacılar tarafından fikir birliğinin sağlanamadığı görülmektedir.

Tablo 4.4: Madde Seçme Yöntemlerinin OTU, OSD, r, Yanlılık, RMSE ve OMH Değerleri

Bağımsız Değişken		*OTU*	*OSD*	*r*	*Yanlılık*	*RMSE*	*OMH*
MFB		22,36	0,97	0,87	0,067	0,436	0,362
KLB		22,35	0,97	0,87	0,068	0,438	0,363
KY		22	0,97	0,91	0,008	0,384	0,304
KN		22,71	0,97	0,83	0,127	0,49	0,421
MFB	**KY**	22,02	0,97	0,91	0,007	0,382	0,303
MFB	**KN**	22,69	0,97	0,83	0,127	0,489	0,42
KLB	**KY**	21,97	0,97	0,91	0,009	0,385	0,305
KLB	**KN**	22,72	0,97	0,83	0,127	0,49	0,421

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

Tablo 4.4'teki dört madde seçme yöntemi incelendiğinde, madde seçme yöntemlerinden en iyi performansı MFB-KY'nin sergilediği; bunu KLB-KY'nin takip ettiği ve MFB-KN ile KLB-KN'nin ise benzer performans gösterdiği görülmektedir. Bu bulgu Eggen ve Straetmans'ın (2000) çalışma sonuçlarıyla örtüşmekte iken

Eggen'in (1998) sonuçlarıyla örtüşmemektedir. Eggen'in (1998) araştırmasında MFB-KN'nin MFB-KY'ye kıyasla daha iyi performans gösterdiği bulunmuştur.

4.5. Beşinci Alt Probleme Ait Bulgular ve Yorumlar

Araştırmanın beşinci alt probleminde Monte Carlo simülasyonu BBST uygulamasında, bağımsız değişkenlerden yetenek kestirim yöntemleri, ortalama test uzunluğu (OTU), ortalama sınıflama doğruluğu (OSD), korelasyon (r), yanlılık, RMSE ve ortalama mutlak hata (OMH) bakımından incelenmiştir. Aşağıda Tablo 4.5'te yetenek kestirim yöntemlerinin bağımlı değişkenler bakımından nasıl değiştiği görülmektedir.

Tablo 4.5: Yetenek Kestirim Yöntemlerinin OTU, OSD, r, Yanlılık, RMSE ve OMH Değerleri

Bağımsız Değişken	*OTU*	*OSD*	*r*	*Yanlılık*	*RMSE*	*OMH*
BSD	22,04	0,97	0,87	0,001	0,424	0,352
AOK	22,65	0,97	0,87	0,135	0,449	0,373

Tablo 4.5'e göre BSD ve AOK yetenek kestirim yöntemlerinin OTU, OSD ve r bakımından oldukça benzer çalıştıkları ancak yanlılık, RMSE ve OMH değerleri bakımından BSD'nin AOK'a kıyasla göreli olarak daha iyi performans sergilediği görülmektedir. Bu bulgu Wang, Hanson ve Lau'nun (1999) çalışma sonuçlarıyla AOK'un değişken uzunluklu testlerde yüksek yanlılık değerine sahip olması bakımından örtüşmektedir. Ayrıca Yi, Wang ve Ban (2000) araştırmalarında değişken uzunluklu testlerde AOK'un BSD'den daha fazla sayıda madde gerektirdiği sonucuna ulaşmışlardır. Tablo 4.5'te BSD ile test ortalama 22 maddede sonlanırken; yetenek kestirim yöntemi AOK olduğunda bu ortalama 23'e çıkmaktadır. Çalışmanın bu bulgusu da alanyazın ile örtüşmektedir.

4.6. Altıncı Alt Probleme Ait Bulgular ve Yorumlar

Araştırmanın altıncı alt probleminde Post Hoc simülasyonu BBST uygulamasında, yetenek kestirim yöntemi BSD olduğunda, AOOT sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GOO sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GA sınıflama kriterinin %70 ile %90 güven düzeyleri için sınıflama doğruluğu, test uzunluğu ve ölçme kesinliğinin madde seçme yöntemlerinden MFB-KY, MFB-KN, KLB-KY ve KLB-KN'ye göre nasıl değiştiği incelenmiştir.

Aşağıda Tablo 4.6'da bu alt problemde belirtilen koşulları karşılaştırmada kullanılan bağımlı değişkenler olan ortalama test uzunluğu (OTU), ortalama sınıflama doğruluğu (OSD), bireylerin gerçek yetenek düzeyleri ile kestirilen son yetenek düzeyleri arasındaki korelasyon (r), yanlılık, RMSE ve ortalama mutlak hataya (OMH) ait değerler yer almaktadır. Değerlerin tümü her koşul için 25 tekrarın ortalaması alınarak hesaplanmıştır.

Tablo 4.6'da, madde seçme yöntemi fark etmeksizin, ortalama test uzunluğu (OTU) bakımından testi sonlandırmada, bir başka deyişle bireyleri sınıflamada, en az madde gerektiren sınıflama kriterinin GA yönteminin %70 düzeyi olduğu; bunu takiben sırasıyla GA yönteminin %90 güven düzeyinin; GOO FB: 0,10 yönteminin; GOO FB: 0,05 yönteminin ve AOOT FB: 0,10 yönteminin geldiği; en fazla madde gerektiren sınıflama kriterinin ise AOOT FB: 0,05 yönteminin olduğu görülmektedir. Bu bulguya göre madde seçme yöntemi ile sınıflama kriterlerinin farksızlık bölgesi değerleri ya da güven düzeyleri fark etmeksizin OTU bakımından en iyi performansı GA sınıflama kriteri göstermiş, bunu sırasıyla GOO ve AOOT sınıflama kriterleri takip etmiştir.

Tablo 4.6'da madde seçme yöntemi fark etmeksizin ortalama sınıflama doğruluğu (OSD) bakımından tüm sınıflama kriterlerinin öğrencileri geçti-kaldı kategorilerine sınıflamada benzer performans gösterdiği ve sınıflama doğruluğunun oldukça yüksek (0,99) olduğu görülmektedir.

Tablo 4.6'da sınıflama kriterlerinden GOO ve AOOT yöntemlerinin farksızlık bölgesi değerleri küçüldükçe ve GA yönteminin güven düzeyi yükseldikçe bireyleri sınıflamada gerekli ortalama madde sayısı olan OTU değerlerinin arttığı görülmektedir. Araştırmanın bu bulgusu Reckase (1983), Lau ve Wang (1999), Thompson ve Ro (2007) ve Thompson'ın (2011) araştırma sonuçlarıyla örtüşmektedir.

Tablo 4.6: Yetenek Kestirim Yöntemi BSD Olduğunda Koşullara Ait OTU, OSD, r, Yanlılık, RMSE ve OMH Değerleri

Koşullar		*Bağımlı Değişkenler*					
Madde Seçme Yöntemi	***Sınıflama Kriteri***	***OTU***	***OSD***	***r***	***Yanlılık***	***RMSE***	***OMH***
MFB-KY	**AOOT FB: 0,05**	49,96	0,99	0,94	0,007	0,341	0,277
	AOOT FB: 0,10	36,52	0,99	0,93	0,011	0,367	0,297
	GOO FB: 0,05	12,76	0,99	0,80	0,016	0,562	0,462
	GOO FB: 0,10	12,26	0,99	0,80	0,011	0,565	0,462
	GA %70 güven düzeyi	**5,61**	**0,99**	**0,75**	**0,010**	**0,636**	**0,523**
	GA %90 güven düzeyi	7,57	0,99	0,76	0,001	0,618	0,508
MFB-KN	**AOOT FB: 0,05**	49,63	0,99	0,89	0,017	0,420	0,346
	AOOT FB: 0,10	31,81	0,99	0,84	0,021	0,493	0,417
	GOO FB: 0,05	12,78	0,99	0,74	0,020	0,617	0,520
	GOO FB: 0,10	12,15	0,99	0,74	0,017	0,621	0,524
	GA %70 güven düzeyi	**5,74**	**0,99**	**0,70**	**0,015**	**0,671**	**0,570**
	GA %90 güven düzeyi	7,72	0,99	0,71	0,015	0,658	0,558
KLB-KY	**AOOT FB: 0,05**	49,97	0,99	0,94	0,011	0,341	0,277
	AOOT FB: 0,10	36,41	0,99	0,92	0,014	0,369	0,300
	GOO FB: 0,05	12,67	0,99	0,80	0,011	0,561	0,462
	GOO FB: 0,10	12,25	0,99	0,80	0,014	0,570	0,470
	GA %70 güven düzeyi	**5,74**	**0,99**	**0,75**	**0,011**	**0,631**	**0,520**
	GA %90 güven düzeyi	7,63	0,99	0,76	0,007	0,622	0,511
KLB-KN	**AOOT FB: 0,05**	49,62	0,99	0,89	0,016	0,419	0,346
	AOOT FB: 0,10	31,89	0,99	0,84	0,022	0,495	0,418
	GOO FB: 0,05	12,70	0,99	0,74	0,013	0,619	0,522
	GOO FB: 0,10	12,25	0,99	0,74	0,018	0,619	0,522
	GA %70 güven düzeyi	**5,80**	0,99	**0,70**	**0,013**	**0,668**	**0,566**
	GA %90 güven düzeyi	7,60	0,99	0,71	0,009	0,662	0,562

Tablo 4.6'da bireylerin simülasyon öncesi türetilen gerçek yetenek düzeyleri ile BBST simülasyonu sonucu kestirilen son yetenek düzeyleri arasındaki korelasyon (r) bakımından en yüksek ilişkinin hesaplandığı sınıflama kriterinin, madde seçme yöntemi fark etmeksizin, AOOT FB: 0,05 yöntemi olduğu; bunu takiben sırasıyla AOOT FB: 0,10 yönteminin, GOO yöntemlerinin ve GA %90 yönteminin geldiği; en düşük ilişkinin ise GA %70 yöntemi ile hesaplandığı görülmektedir. Bu korelasyonlar madde seçme yöntemlerine göre ayrı ayrı incelendiğinde en iyi performansı MFB-KY ve KLB-KY madde seçme yöntemlerinin verdiği görülmektedir. Bu iki yöntem için hesaplanan korelasyonlar 0,75 ile 0,94 değerleri arasında değişmekte; bu durum da gerçek yeteneklerle kestirilen yetenekler arasındaki korelasyonların bu iki madde seçme yöntemi kullanıldığında oldukça yüksek olduğunu göstermektedir. MFB-KN ve KLB-KN madde seçme yöntemlerine ait korelasyon değerleri ise 0,70 ile 0,89 aralığında değişmektedir. Bu bulguya dayanarak bireylerin gerçek yetenek düzeyleri ile kestirilen son yetenek düzeyleri arasındaki korelasyon bakımından, kestirilen yetenek (KY) temelli madde seçme yöntemlerinin kesme noktası (KN) temelli madde seçme yöntemlerine kıyasla daha iyi performans gösterdiği yorumu yapılabilir.

Tablo 4.6'da sınıflama kriterlerinin, Tablo 4.1'deki Monte Carlo simülasyonu çalışmasındaki değerlere kıyasla daha yüksek yanlılık değerlerine sahip olduğu ve performanslarının madde seçme yöntemlerine göre farklılık gösterdiği görülmektedir. Sınıflama kriterlerinin en düşük yanlılık değeri, madde seçme yöntemi MFB-KY olduğunda elde edilmiştir. Buna göre yanlılık bakımından sınıflama kriterlerinin MFB-KY ile birlikte başarılı bir performans gösterdikleri; bunu takiben KLB-KY madde seçme yönteminin geldiği görülmektedir. KY temelli madde seçme yöntemleriyle KN temelli yöntemlere kıyasla sınıflama kriterlerinin yanlılık değerlerinin daha düşük olduğu görülmektedir. Buna dayanarak sınıflama kriterleriyle KY temelli madde seçme yöntemlerinin ele alınmasının yanlılık bakımından daha iyi performans gösterdiği söylenebilir.

Tablo 4.6'da yanlılık ile birlikte kestirimin standart hatasını da dikkate alan RMSE ve ortalama mutlak hata (OMH) değerlerine göre, madde seçme yöntemi fark etmeksizin, en iyi performans gösteren sınıflama kriteri AOOT FB: 0,05 yöntemidir. Bunu takiben sırasıyla AOOT FB: 0,10; GOO FB: 0,05; GOO FB: 0,10; GA %90 ve GA %70 yöntemlerinin geldiği görülmektedir. Farksızlık bölgesi ve güven düzeyleri

fark etmeksizin RMSE ve OMH bakımından en iyi performansı AOOT yöntemi göstermiştir. AOOT yönteminin ardından GOO yöntemi gelmiş ve göreli olarak en kötü performansı ise GA yöntemi göstermiştir.

Tablo 4.6'da görülüğü üzere gerçek veri seti üzerinden gerçekleştirilen Post Hoc simülasyonu BBST uygulamasında yetenek kestirim yöntemi BSD olduğunda, madde seçme yöntemi fark etmeksizin test etkililiği (ortalama test uzunluğu ve ortalama sınıflama doğruluğu) bakımından GA %70 yönteminin diğer sınıflama kriterlerine kıyasla oldukça başarılı bir performans gösterdiği söylenebilir. Bunu sırasıyla GA %90; GOO FB: 0,10; GOO FB: 0,05; AOOT FB: 0,10 ve AOOT FB: 0,05 sınıflama kriterleri izlemektedir. Bireyler geçti-kaldı kategorilerine GA %70 sınıflama kriteri ile ortalama 6 madde ve ortalama 0,99 sınıflama doğruluğuyla sınıflanabilirken; diğer yöntemlerin tümünde de ortalama sınıflama doğruluğu 0,99 olmak üzere GA %90 sınıflama kriteri ile ortalama 8 madde; GOO FB: 0,10 sınıflama kriteri ile ortalama 12 madde; GOO FB: 0,05 sınıflama kriteri ile ortalama 13 madde; AOOT FB: 0,10 sınıflama kriteri ile ortalama 35 madde ve AOOT FB: 0,05 sınıflama kriteri ile de ortallama 50 madde ile testin sonlanabildiği ve bireylerin kategorilere yerleştirilebildiği görülmüştür. Test etkililiği açısından bakıldığında GA ve GOO sınıflama kriterlerinin AOOT'ye kıyasla başarılı performans gösterdikleri görülmektedir. Bununla birlikte kestirilen ve gerçek yetenek düzeyleri arasındaki ilişki (r), yanlılık, RMSE ve OMH değerleri bakımından AOOT sınıflama kriterinin göreli olarak diğer yöntemlerden daha iyi performans gösterdiği; ancak tüm koşullar içinden bu göreli değerlerin en düşük olduğu MFB-KY madde seçme yönteminin ve AOOT FB: 0,05 sınıflama kriterinin birlikte ele alındığı koşulda, GA %70 yönteminin neredeyse 8 katı maddede testin sonlandığı-bireylerin sınıflanabildiği dikkat çekmektedir.

4.7. Yedinci Alt Probleme Ait Bulgular ve Yorumlar

Araştırmanın yedinci alt probleminde Post Hoc simülasyonu BBST uygulamasında, yetenek kestirim yöntemi AOK olduğunda, AOOT sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GOO sınıflama kriterinin FB: 0,05 ile FB: 0,10 düzeyleri için, GA sınıflama kriterinin %70 ile %90 güven düzeyleri için sınıflama doğruluğu, test uzunluğu ve ölçme kesinliğinin madde seçme yöntemlerinden MFB-KY, MFB-KN, KLB-KY ve KLB-KN'ye göre nasıl değiştiği incelenmiştir.

Aşağıda Tablo 4.7'de bu alt problemde belirtilen koşulları karşılaştırmada kullanılan bağımlı değişkenler olan ortalama test uzunluğu (OTU), ortalama sınıflama doğruluğu (OSD), bireylerin gerçek yetenek düzeyleri ile kestirilen son yetenek düzeyleri arasındaki korelasyon (r), yanlılık, RMSE ve ortalama mutlak hataya (OMH) ait değerler yer almaktadır. Değerlerin tümü her koşul için 25 tekrarın ortalaması alınarak hesaplanmıştır.

Tablo 4.7'de madde seçme yöntemi fark etmeksizin ortalama test uzunluğu (OTU) bakımından testi sonlandırmada, bir başka deyişle bireyleri sınıflamada, en az madde gerektiren sınıflama kriterinin GA yönteminin %70 düzeyi olduğu; bunu takiben sırasıyla GA yönteminin %90 güven düzeyinin; GOO FB: 0,10 yönteminin; GOO FB: 0,05 yönteminin ve AOOT FB: 0,10 yönteminin geldiği; en fazla madde gerektiren sınıflama kriterinin ise AOOT FB: 0,05 yönteminin olduğu görülmektedir. Bu bulguya göre madde seçme yöntemi ile sınıflama kriterlerinin farksızlık bölgesi değerleri ya da güven düzeyleri fark etmeksizin OTU bakımından en iyi performansı GA sınıflama kriteri göstermiş, bunu sırasıyla GOO ve AOOT sınıflama kriterleri takip etmiştir.

Tablo 4.7'de sınıflama kriterlerinden GOO ve AOOT yöntemlerinin farksızlık bölgesi değerleri küçüldükçe ve GA yönteminin güven düzeyi yükseldikçe bireyleri sınıflamada gerekli ortalama madde sayısı olan OTU değerlerinin arttığı görülmektedir. Araştırmanın bu bulgusu Reckase (1983), Lau ve Wang (1999), Thompson ve Ro (2007) ve Thompson'ın (2011) araştırma sonuçlarıyla örtüşmektedir.

Tablo 4.7'de madde seçme yöntemi fark etmeksizin ortalama sınıflama doğruluğu (OSD) bakımından tüm sınıflama kriterlerinin öğrencileri geçti-kaldı kategorilerine sınıflamada benzer performans gösterdiği ve sınıflama doğruluğunun oldukça yüksek (0,99) olduğu görülmektedir.

Tablo 4.7'de gerçek yetenek düzeyleri ile kestirilen yetenek düzeyleri arasındaki korelasyon (r) bakımından en yüksek ilişkinin hesaplandığı yöntemin MFB-KY yöntemi kullanıldığında AOOT FB: 0,05 yöntemi olduğu; bunu takiben sırasıyla AOOT FB: 0,10 yönteminin, GOO yöntemlerinin ve GA %90 yönteminin geldiği; en düşük ilişkinin ise GA %70 yöntemi ile hesaplandığı görülmektedir. Tüm sınıflama kriterlerine ait r değerleri KY (kestirilen yetenek) temelli yöntemler için 0,74 ile 0,93

aralığında ve KN (kesme noktası) temelli yöntemler için 0,69 ile 0,89 aralığındadır. Buna göre r bakımından KY temelli madde seçme yöntemlerinin KN temelli madde seçme yöntemlerine kıyasla daha iyi performans gösterdiği söylenebilir.

Tablo 4.7'de sınıflama kriterlerinin performanslarının madde seçme yöntemlerine göre farklılık gösterdiği görülmektedir. KN temelli madde seçme yöntemlerine kıyasla KY temelli madde seçme yöntemleri kullanıldığında sınıflama kriterlerinin genel olarak daha düşük yanlılık değerine sahip oldukları görülmektedir. Bir başka deyişle yanlılık bakımından MFB-KY ve KLB-KY yöntemleri sınıflama kriterleriyle birlikte daha iyi performans göstermiştir. MFB-KY madde seçme yöntemiyle yanlılık bakımından en iyi performansı GA %90 sınıflama kriteri gösterirken; diğer madde seçme yöntemleri ile birlikte en iyi performansı gösteren sınıflama kriteri AOOT FB: 0,05 yöntemidir.

Tablo 4.7'de yanlılık ile birlikte kestirimin standart hatasını da dikkate alan RMSE ve ortalama mutlak hata (OMH) değerlerine göre madde seçme yöntemi fark etmeksizin, en iyi performans gösteren sınıflama kriteri AOOT FB: 0,05 yöntemidir. Bunu takiben sırasıyla AOOT FB: 0,10 yöntemi; GOO sınıflama kriterleri, GA %70 düzeyi ile GA %90 düzeyleri gelmektedir. KY temelli madde seçme yöntemleri kullanıldığında GOO sınıflama kriterlerinin, farksızlık bölgesi değerleri değişmeksizin, aynı RMSE ve OMH'ye sahip oldukları; KN temelli madde seçme yöntemleri kullanıldığında ise GOO FB: 0,05 sınıflama kriterinin GOO FB: 0,10 yöntemine kıyasla daha düşük RMSE ve OMH değerlerine sahip olduğu görülmektedir.

Tablo 4.7'de görülüğü üzere BBST simülasyonunda yetenek kestirim yöntemi AOK olduğunda, madde seçme yöntemi fark etmeksizin test etkililiği bakımından GA %70 yönteminin diğer sınıflama kriterlerine kıyasla oldukça başarılı bir performans gösterdiği söylenebilir. Bunu sırasıyla GA %90 sınıflama kriteri, GOO yöntemleri ve AOOT FB: 0,10 ve AOOT FB: 0,05 yöntemleri takip etmektedir.

Tablo 4.7: Yetenek Kestirim Yöntemi AOK Olduğunda Koşullara Ait OTU, OSD, r, Yanlılık, RMSE ve OMH Değerleri

Koşullar		*Bağımlı Değişkenler*					
Madde Seçme Yöntemi	***Sınıflama Kriteri***	***OTU***	***OSD***	***r***	***Yanlılık***	***RMSE***	***OMH***
MFB-KY	**AOOT FB: 0,05**	49,99	0,99	0,93	0,008	0,373	0,299
	AOOT FB: 0,10	36,15	0,99	0,92	-0,013	0,408	0,324
	GOO FB: 0,05	12,62	0,99	0,77	-0,011	0,674	0,534
	GOO FB: 0,10	12,11	0,99	0,77	-0,006	0,674	0,534
	GA %70 güven düzeyi	**6,88**	**0,99**	**0,74**	**0,028**	**0,732**	**0,590**
	GA %90 güven düzeyi	9,73	0,99	0,75	0,004	0,699	0,558
MFB-KN	**AOOT FB: 0,05**	49,63	0,99	0,89	0,091	0,447	0,363
	AOOT FB: 0,10	31,80	0,99	0,83	0,183	0,521	0,437
	GOO FB: 0,05	12,71	0,99	0,72	0,258	0,677	0,568
	GOO FB: 0,10	12,30	0,99	0,72	0,261	0,682	0,573
	GA %70 güven düzeyi	**6,93**	**0,99**	**0,69**	**0,289**	**0,731**	**0,613**
	GA %90 güven düzeyi	10,57	0,99	0,71	0,262	0,694	0,585
KLB-KY	**AOOT FB: 0,05**	49,98	0,99	0,93	0,002	0,372	0,299
	AOOT FB: 0,10	35,98	0,99	0,92	-0,006	0,414	0,329
	GOO FB: 0,05	12,61	0,99	0,77	0,028	0,668	0,531
	GOO FB: 0,10	12,16	0,99	0,77	0,032	0,668	0,534
	GA %70 güven düzeyi	**6,85**	**0,99**	**0,74**	**0,067**	**0,728**	**0,582**
	GA %90 güven düzeyi	9,78	0,99	0,75	0,045	0,690	0,550
KLB-KN	**AOOT FB: 0,05**	49,63	0,99	0,88	0,090	0,449	0,364
	AOOT FB: 0,10	31,92	0,99	0,83	0,188	0,519	0,435
	GOO FB: 0,05	12,85	0,99	0,73	0,260	0,675	0,567
	GOO FB: 0,10	12,32	0,99	0,72	0,262	0,679	0,571
	GA %70 güven düzeyi	6,9	0,99	0,69	0,291	0,733	0,614
	GA %90 güven düzeyi	10,64	0,99	0,71	0,262	0,689	0,580

Tablo 4.7'de bireyler geçti-kaldı kategorilerine GA %70 sınıflama kriteri ile ortalama 7 madde ve ortalama 0,99 sınıflama doğruluğuyla sınıflanabilirken; diğer yöntemlerin tümünde ortalama sınıflama doğruluğu da 0,99 olmak üzere GA %90 sınıflama kriteri ile ortalama 11 madde; GOO sınıflama kriterleriyle 13 madde; AOOT FB: 0,10 sınıflama kriteri ile ortalama 34 madde ve AOOT FB: 0,05 sınıflama kriteri ile de ortalama 50 maddeyle testin sonlanabildiği ve bireylerin kategorilere yerleştirilebildiği görülmüştür. Test etkililiği açısından bakıldığında GA ve GOO sınıflama kriterlerinin AOOT'ye kıyasla başarılı performans gösterdikleri görülmektedir. Bununla birlikte kestirilen ve gerçek yetenek düzeyleri arasındaki ilişki (r), yanlılık, RMSE ve OMH değerleri bakımından AOOT sınıflama kriterinin göreli olarak diğer yöntemlerden daha iyi performans gösterdiği; ancak tüm koşullar içinden bu göreli değerlerin en düşük olduğu KLB-KY madde seçme yönteminin ve AOOT FB: 0,05 sınıflama kriterinin birlikte ele alındığı koşulda, GA %70 yönteminin neredeyse 7 katı maddede testin sonlandığı-bireylerin sınıflanabildiği görülmektedir.

4.8. Sekizinci Alt Probleme Ait Bulgular ve Yorumlar

Araştırmanın sekizinci alt probleminde Post Hoc simülasyonu BBST uygulamasında, bağımsız değişkenlerden sınıflama kriterleri, ortalama test uzunluğu (OTU), ortalama sınıflama doğruluğu (OSD), korelasyon (r), yanlılık, RMSE ve ortalama mutlak hata (OMH) bakımından incelenmiştir. Aşağıda Tablo 4.8'de sınıflama kriterlerinin bağımlı değişkenler bakımından nasıl değiştiği görülmektedir.

Tablo 4.8: Sınıflama Kriterlerinin OTU, OSD, r, Yanlılık, RMSE ve OMH Değerleri

Bağımsız Değişken	*OTU*	*OSD*	*r*	*Yanlılık*	*RMSE*	*OMH*
AOOT	41,93	0,99	0,89	0,041	0,422	0,346
GOO	12,47	0,99	0,76	0,075	0,633	0,522
GA	7,61	0,99	0,73	0,083	0,679	0,562

Tablo 4.8'de sınıflama kriterlerinden, farklı hata düzeyleri veya farksızlık bölgesi değerleri, madde seçme ve yetenek kestirim yöntemleri dikkate alınmaksızın, en az sayıda maddeyle bireyleri sınıflayabilen yöntemin yaklaşık 8 maddeyle GA yöntemi olduğu; bunu 13 maddeyle GOO yönteminin takip ettiği ve en çok sayıda maddeyle sınıflama yapabilen yöntemin de 42 maddeyle AOOT olduğu görülmektedir. Buna göre OTU bakımından en iyi performansı GA sınıflama kriteri

göstermiş; bunu GOO yöntemi takip etmiştir. En kötü performansı gösteren yöntem ise AOOT'dir.

Tablo 4.8'deki OTU değerleri ile Tablo 4.3'teki MC simülasyonundan elde edilen OTU değerleri karşılaştırıldığında, sınıflama kriterlerinin OTU bakımından bu alt problemdeki gibi sıralandığı ancak yöntemlerin OTU değerlerinin değiştiği görülmektedir. Bireyleri sınıflamak için AOOT yöntemi MC simülasyonunda 40 madde gerektirirken gerçek veri üzerinden uygulanan PH simülasyonunda bu değer 42'ye çıkmıştır. GOO ve GA sınıflama kriterleri, MC ve PH simülasyon sonuçları bakımından karşılaştırıldığında ise PH simülasyonunda bu iki yöntemin OTU değerlerinin ortalama üç madde azaldığı görülmektedir. Gerçek uygulamaya daha yakın bir simülasyon çalışması olan PH simülasyonunda, GOO ve GA sınıflama kriterleri için MC'ye kıyasla daha düşük OTU değerlerinin elde edilmesi, bu iki yöntemin gerçek bir BBST uygulamasında da daha başarılı performans gösterebileceğine işaret etmektedir.

Tablo 4.8'e göre sınıflama kriterlerinin ortalama sınıflama doğruluğu (OSD) bakımından benzer sonuçlar verdikleri görülmektedir. Tablo 4.3'teki MC simülasyonuna ait OSD değerlerine kıyasla gerçek veri üzerinden uygulanan BBST simülasyonunda sınıflama kriterlerinin daha yüksek sınıflama doğruluğuna sahip oldukları görülmektedir. Buna göre ortalama test uzunluğu ve ortalama sınıflama doğruluğu bakımından GA ve GOO yöntemlerinin gerçek BBST uygulamalarında avantaj sağlayacağı düşünülebilir. Ancak gerçek yetenek düzeyiyle kestirilen yetenek düzeyleri arasındaki korelasyon (r), yanlılık, RMSE ve OMH değerleri bakımından, GA yönteminin diğer iki yönteme kıyasla daha kötü performans gösterdiği; bu ölçütler bakımından AOOT sınıflama kriterinin diğer yöntemlerden göreli olarak daha iyi performans sergilediği görülmektedir.

Tablo 4.8 ile Tablo 4.3'teki RMSE ve OMH değerleri karşılaştırıldığında PH simülasyonunda MC simülasyonunda kıyasla hatanın daha yüksek olduğu görülmektedir. Bunun nedeni PH simülasyonunda gerçek veri setinin kullanılmış olması olabilir. Simülatif MC veri seti tamamen belirlenen koşullara uygun şekilde manipüle edildiğinden BBST sonucu hata değerleri daha küçük hesaplanmıştır. PH simülasyonunun ise gerçek uygulama verisi içermesi sebebiyle, gerçek bir BBST uygulamasına daha yakın hata değerleri vermiş olabileceği düşünülmektedir.

BBST uygulamalarından beklenen bireyleri az sayıda maddeyle yüksek doğrulukta sınıflamaktır. Tablo 4.8'de AOOT'nin ortalama 42 maddeyle testi ancak sonlandırabilmesi, bir başka deyişle bireyleri kategorilere sınıflayabilmesi için ortalama 42 madde gerektirmesi nedeniyle yöntemin kullanışlı olmayacağı görülmektedir. Bu açıdan bakıldığında GOO sınıflama kriterinin diğer yöntemlere kıyasla daha kullanışlı bir seçenek olduğu yorumu yapılabilir. Bu bulgu, Thompson (2011) ile Nydick, Nozawa ve Zhu'nun (2012) çalışmalarının sonuçları ile örtüşmektedir. Bahsedilen çalışmalara göre de GOO'nun madde sayısı ve sınıflama doğruluğu bakımından en uygun sınıflama kriteri olduğu ortaya koyulmuştur.

4.9. Dokuzuncu Alt Probleme Ait Bulgular ve Yorumlar

Araştırmanın dokuzuncu alt probleminde Post Hoc simülasyonu BBST uygulamasında, bağımsız değişkenlerden madde seçme yöntemleri, ortalama test uzunluğu (OTU), ortalama sınıflama doğruluğu (OSD), korelasyon (r), yanlılık, RMSE ve ortalama mutlak hata (OMH) bakımından incelenmiştir. Aşağıda Tablo 4.9'da madde seçme yöntemlerinin bağımlı değişkenler bakımından nasıl değiştiği görülmektedir.

Tablo 4.9: Madde Seçme Yöntemlerinin OTU, OSD, r, Yanlılık, RMSE ve OMH Değerleri

Bağımsız Değişken		*OTU*	*OSD*	*r*	*Yanlılık*	*RMSE*	*OMH*
MFB		20,66	0,99	0,80	0,064	0,579	0,477
KLB		20,67	0,99	0,80	0,070	0,578	0,477
KY		21,00	0,99	0,82	0,013	0,554	0,447
KN		20,33	0,99	0,77	0,121	0,603	0,506
MFB	**KY**	21,01	0,99	0,82	0,006	0,554	0,447
MFB	**KN**	20,31	0,99	0,77	0,121	0,603	0,506
KLB	**KY**	21,00	0,99	0,82	0,020	0,553	0,447
KLB	**KN**	20,34	0,99	0,77	0,120	0,602	0,506

Tablo 4.9'a göre, madde seçme yönteminin hangi ölçütü temele aldığı fark etmeksizin, MFB ve KLB madde seçme yöntemlerinin tüm bağımlı değişkenler bakımından birbirine oldukça benzer performans gösterdiği görülmektedir. Alanyazın incelendiğinde bu iki madde seçme yönteminin performansları hakkında araştırmacılar tarafından fikir birliğinin sağlanamadığı görülmektedir. Bu alt problemin bulgusu Eggen (1999), Lau ve Wang (1999), Cheng ve Liou (2000) ile Lin ve Spray'in (2000) çalışma sonuçlarına benzerlik göstermekte iken; Spray ve

Reckase (1994) ile Lau ve Wang'ın (1998) araştırma sonuçlarıyla örtüşmemektedir.

Tablo 4.9'da yer alan madde seçme yöntemlerinin hangi temele dayandığı incelendiğinde ise ortalama test uzunluğu bakımından kesme noktasına dayanan (KN) madde seçiminin kestirilen yeteneğe dayanan (KY) madde seçimine kıyasla daha iyi performans sergilediği görülmektedir. Alanyazında bu karşılaştırmaya az sayıda çalışmada yer verilmiş ve araştırmalarda bu yöntemlerin etkililiği hakkında ortak bir sonuca ulaşılamamıştır. Bu alt problemde elde edilen bulgu Spray ve Reckase'in (1994) çalışma sonuçlarıyla örtüşmekte iken; Thompson'ın (2007, 2009) araştırma bulgularıyla örtüşmemektedir.

Tablo 4.9'da KY ve KN temelli madde seçme yöntemleri ortalama sınıflama doğruluğu bakımından benzer sonuçlar vermiştir. Bu noktada test etkililiği düşünüldüğünde, bir başka deyişle OTU ve OSD birlikte ele alındığında, küçük bir farkla KN temelli yöntemlerin daha iyi performans gösterdiği görülmektedir. Ancak gerçek yetenek düzeyleriyle kestirilen son yetenek düzeyleri arasındaki korelasyon, yanlılık, RMSE ve OMH değerleri incelendiğinde KY temelli madde seçme yönteminin daha başarılı olduğu dikkati çeken diğer bir bulgudur. İki yöntemin birbirine oldukça yakın OTU ile OSD değerlerine sahip olması ve diğer dört bağımlı değişken bakımından da KY temelli madde seçme yöntemlerinin daha az hata içermesi sebebiyle KY temelli madde seçme yöntemlerini kullanmanın daha avantajlı olduğu söylenebilir.

Tablo 4.9'daki dört madde seçme yöntemine ait sonuçlarda, MFB-KN ve KLB-KN'nin test etkililiği (OTU ve OSD) bakımından iyi ancak diğer değişkenler bakımından daha kötü performans sergilediği görülmektedir. MFB-KN'nin test etkililiği bakımından iyi sonuç vermesi Eggen'in (1998) araştırma sonuçlarıyla örtüşmektedir.

Tablo 4.9 ile MC simülasyon sonuçlarının yer aldığı Tablo 4.4 karşılaştırıldığında ise, madde seçme yöntemlerinin tümü için MC simülasyonuna kıyasla PH simülasyon sonuçlarının, daha az sayıda madde, daha yüksek sınıflama doğruluğu ve daha yüksek korelasyon değerleriyle daha iyi performans gösterdikleri görülmektedir. RMSE ve OMH bakımından ise gerçek veri seti üzerinden gerçekleştirilen PH simülasyonun MC simülasyonuna kıyasla madde

seçme yöntemlerinin daha yüksek hata değerlerine sahip olduğu görülmektedir. Bunun nedeni olarak BBST simülasyonunun, MC veri setinin aksine, PH veri setinde herhangi bir manüpülasyon olmaksızın gerçekleşmiş olması ve bu durumun da simülasyonun gerçek bir uygulamaya daha yakın hata değerleri vermesine sebep olabileceği düşünülmektedir.

4.10. Onuncu Alt Probleme Ait Bulgular ve Yorumlar

Araştırmanın onuncu alt probleminde Post Hoc simülasyonu BBST uygulamasında, bağımsız değişkenlerden yetenek kestirim yöntemleri, ortalama test uzunluğu (OTU), ortalama sınıflama doğruluğu (OSD), korelasyon (r), yanlılık, RMSE ve ortalama mutlak hata (OMH) bakımından incelenmiştir. Aşağıda Tablo 4.10'da yetenek kestirim yöntemlerinin bağımlı değişkenler bakımından nasıl değiştiği görülmektedir.

Tablo 4.10: Yetenek Kestirim Yöntemlerinin OTU, OSD, r, Yanlılık, RMSE ve OMH Değerleri

Bağımsız Değişken	*OTU*	*OSD*	*r*	*Yanlılık*	*RMSE*	*OMH*
BSD	20.38	0.99	0,80	0,013	0,548	0,456
AOK	20.96	0.99	0,79	0,12	0,608	0,497

Tablo 4.10'da BSD ve AOK yöntemlerinin OTU, OSD ve r bakımından oldukça benzer çalıştıkları ancak yanlılık, RMSE ve OMH değerleri bakımından BSD'nin AOK'a kıyasla göreli olarak daha iyi performans sergilediği görülmektedir. Bu bulgu Wang, Hanson ve Lau'nun (1999) çalışmasıyla, AOK'un değişken uzunluklu testlerde yüksek yanlılık değerine sahip olması bakımından örtüşmektedir. Tablo 4.10 ile MC simülasyon sonuçlarının yer aldığı Tablo 4.5 karşılaştırıldığında, PH simülasyonu sonucunda yetenek kestirim yöntemlerinden elde edilen OTU değerlerinin azaldığı ve OSD değerlerinin yükseldiği görülmektedir. Bu bulguya dayanarak gerçek veri seti üzerinden gerçekleştirilen PH simülasyonunda yetenek kestirim yöntemlerinin daha iyi performans gösterdiği söylenebilir. RMSE ve OMH bakımından ise PH simülasyon sonuçlarının daha yüksek değerlere sahip olduğu sonucuna ulaşılmıştır. PH simülasyonunda gerçek veri setinin kullanmanın BBST simülasyonu sonucu hataları artırmış olabileceği düşünülmektedir.

5. SONUÇ ve ÖNERİLER

Bu bölümde araştırma sonuçlarına ve sonuçlara dayalı önerilere yer verilmiştir.

5.1. Sonuçlar

1. Bu araştırmada, BBST uygulamalarındaki sınıflama kriterleri, yetenek kestirim yöntemleri ve madde seçme yöntemleri hem Monte Carlo simülasyonu (MC) hem de Post Hoc (PH) simülasyonu altında incelenmiştir. MC ve PH simülasyon sonuçları birbirine benzerlik gösterdiğinden bir arada ele alınmıştır.
2. Araştırma sonucunda, hem MC hem de PH simülasyonlarındaki tüm koşullarda, madde seçme yöntemi ve yetenek kestirim yöntemi fark etmeksizin, bireyleri sınıflamada en az sayıda madde gerektiren sınıflama kriterinin Güven Aralığı (GA) yöntemi olduğu; bunu takiben Genelleştirilmiş Olabilirlik Oranı (GOO) yönteminin geldiği ve en fazla sayıda madde gerektiren sınıflama kriterinin ise Ardışık Olasılık Oran Testi (AOOT) olduğu ortaya koyulmuştur.
3. Çalışmanın hem MC hem de PH simülasyonlarındaki tüm koşullarında, madde seçme yöntemi ve yetenek kestirim yöntemi fark etmeksizin, sınıflama kriterlerinin farksızlık bölgesi değerleri veya hata düzeyleri değiştiğinde ortalama test uzunluklarının da değiştiği görülmüştür. BBST simülasyonu sonucunda, AOOT ve GOO sınıflama kriterleri için farksızlık bölgesi genişledikçe ve GA sınıflama kriteri için ise hata düzeyi değeri küçüldükçe ortalama test uzunluğunun azaldığı belirlenmiştir.
4. Araştırmanın MC ve PH simülasyonu bölümlerinin her ikisinde de, madde seçme yöntemi ve yetenek kestirim yöntemi fark etmeksizin, sınıflama kriterlerinden elde edilen ortalama sınıflama doğruluklarının birbirine yakın değerler aldığı ve bu değerlerin oldukça yüksek düzeyde sınıflama doğruluğuna işaret ettiği görülmüştür. Buna göre üç sınıflama kriterinin de bireyleri doğru kategoriye atayabilmesi bakımından oldukça iyi performans gösterdikleri belirlenmiştir.

5. Çalışmanın hem MC hem de PH simülasyonlarında, madde seçme yöntemi ve yetenek kestirim yöntemi fark etmeksizin, sınıflama kriterlerinin kestirilen yetenekler ile gerçek yetenekler arasındaki korelasyonların (r) ortalaması bakımından yüksek düzeyde ilişki verdikleri görülmüştür. Buna göre sınıflama kriterlerinin BSD veya AOK ile birlikte yetenek kestiriminde başarılı oldukları belirlenmiştir.
6. Çalışmanın MC ve PH simülasyonları olmak üzere her iki bölümünde de, madde seçme yöntemi ve yetenek kestirim yöntemi fark etmeksizin, yanlılık, RMSE ve ortalama mutlak hata bakımından göreli olarak en iyi performansı AOOT yönteminin gösterdiği; bunu GOO yönteminin takip ettiği ve en kötü performansı ise GA yönteminin gösterdiği belirlenmiştir. GA yönteminden elde edilen yanlılık, RMSE ve ortalama mutlak hata değerlerinin diğer sınıflama kriterlerine kıyasla daha yüksek olduğu ortaya çıkmıştır.
7. Çalışmanın hem MC hem de PH simülasyonu bölümlerinde incelenen madde seçme yöntemleri olan MFB ve KLB'nin, sınıflama kriteri ve yetenek kestirimi yöntemi fark etmeksizin, ortalama test uzunluğu, ortalama sınıflama doğruluğu, kestirilen yetenek ile gerçek yetenek düzeyi arasındaki korelasyon, yanlılık, RMSE ve ortalama mutlak hata bakımından birbirine oldukça benzer çalıştıkları belirlenmiştir.
8. Araştırmanın iki bölümünde de, sınıflama kriteri ve yetenek kestirim yöntemi fark etmeksizin, madde seçme yöntemlerinin dayandığı temel bakımından kestirilen yetenek (KY) ve kesme noktası (KN) temelli yöntemlerin ortalama test uzunluğu ve ortalama sınıflama doğruluğu bakımından birbirine benzer performans gösterdikleri ancak kestirilen yetenek ile gerçek yetenek düzeyi arasındaki korelasyon, yanlılık, RMSE ve ortalama mutlak hata bakımından KY'nin KN'ye kıyasla daha iyi performans gösterdiği belirlenmiştir.
9. Çalışmanın her iki bölümünde de, sınıflama kriteri ve yetenek kestirim yöntemi fark etmeksizin, MFB ve KLB madde seçme yöntemleri ile yöntemlerin dayandıkları temellerin (KY ve KN) çaprazlanması sonucu, MFB-KY yönteminin ortalama test uzunluğu, ortalama sınıflama doğruluğu, kestirilen yetenek ile gerçek yetenek düzeyi arasındaki korelasyon, yanlılık,

RMSE ve ortalama mutlak hata bakımından diğer yöntemlere kıyasla daha iyi performans gösterdiği sonucuna ulaşılmıştır.

10. Yetenek kestirim yöntemlerinin ortalama test uzunluğu, ortalama sınıflama doğruluğu ve kestirilen yetenek ile gerçek yetenek düzeyi arasındaki korelasyon açısından benzer çalıştıkları ancak BSD'nin çok düşük yanlılık değerine; göreli olarak daha küçük RMSE ve ortalama mutlak hata değerine sahip olduğu görülmüştür. Buna göre BBST uygulamalarında BSD'nin AOK üzerindeki gücü ortaya koyulmuştur.

5.2. Öneriler

Bu başlık altında BBST uygulamaları hakkında araştırmacılara ve uygulayıcılara öneriler yer almaktadır.

5.2.1. Araştırmaya Dönük Öneriler

1. Bu çalışmanın bir sınırlılığı, araştırmada içerik dengelemenin (content balancing) incelenmemiş olmasıdır. Buna göre ileriki çalışmalarda birden fazla içeriğe sahip madde havuzu üzerinden içerik dengeleme ile hangi sınıflama kriterinin daha başarılı performans gösterdiği incelenebilir.
2. Bu çalışmanın ikinci sınırlılığı, madde kullanım sıklığının (item exposure) ele alınmamış olmasıdır. Madde kullanım sıklığı kontrolü ile madde havuzundan bireylere yöneltilen maddelerin görünme sıklıkları incelenebilir ve hangi kontrol yönteminin BBST uygulamaları için daha yararlı olacağı araştırılabilir.
3. Sivri, basık vb. özellikte oluşturulan madde havuzlarında, bir başka deyişle test bilgi fonksiyonunun farklı dağılım özellikleri göstermesi durumunda, sınıflama kriterlerinin performansı incelenebilir. Bir başka deyişle sınıflama kriterleri farklı türdeki madde havuzlarıyla çaprazlanabilir.
4. Bu çalışmada sadece ikili puanlanan maddelerden oluşan madde havuzları ele alınmıştır. Benzer çalışmalar çok kategorili veya karma maddelerden oluşan madde havuzları ile yapılabilir.
5. Bu araştırmanın bir diğer sınırlılığı kesme noktasının bir ve sınıf sayısının iki olması durumudur. Bireylerin sınıflanacağı kategori sayısının 2'den fazla

olduğu durumlarda sınıflama kriterlerinin ve diğer yöntemlerin performansı incelenebilir.

6. Aynı madde havuzu üzerinden farklı kesme noktalarına göre BBST simülasyonları yürütülebilir ve madde havuzu özellikleriyle kesme noktaları çaprazlanabilir.
7. Bu çalışmada sadece BSD ve AOK yetenek kestirim yöntemleri kullanılmıştır. Benzer bir çalışma MOK, MSD gibi diğer yetenek kestirim yöntemleriyle yürütülebilir.
8. Bu çalışmada MFB-KY, MFB-KN, KLB-KY ve KLB-KN madde seçme yöntemleri incelenmiştir. İleride yapılacak çalışmalarda diğer madde seçme yöntemleri kullanılabilir.
9. Bu araştırmada madde havuzlarının tek boyutlu olması söz konusudur. Çok boyutlu madde havuzlarında BBST çalışmaları yürütülebilir.
10. Boyutluluk durumunun ve model uyumunun sağlanamadığı durumlarda sınıflama kriterlerinin nasıl performans gösterdiği araştırılabilir.

5.2.2. Uygulamaya Dönük Öneriler

1. BBST uygulamalarında bu çalışmada incelenen sınıflama kriterlerinden ortalama test uzunluğu (OTU) bakımından yüksek değerler vermesi sebebiyle AOOT'nin tercih edilmemesi; bunun yerine çalışma sonucunda test etkililiğini artırdığı görülen GOO veya GA sınıflama kriterlerinin kullanılması önerilebilir.
2. Gerçek veri seti üzerinden gerçekleştirilen PH simülasyon çalışmasında simülatif veri seti üzerinden gerçekleştirilen MC simülasyon çalışmasına kıyasla, AOOT'nin OTU değerlerindeki artışı, gerçek bir BBST uygulamasında da karşılaşılabilecek bir durum olduğundan ve bu durumun test etkililiğini azaltacağından, uygulayıcıların AOOT yerine diğer iki sınıflama kriterinden birini tercih etmeleri önerilebilir.
3. GOO sınıflama kriterine ait farksızlık bölgesi değerinin artışı ve GA sınıflama kriterine ait hata düzeyi değerinin düşmesi sonucunda OTU değerlerinin düştüğü, buna bağlı olarak da test etkililiğinin arttığı dikkate alınması gereken bir durum olarak karşımıza çıkmaktadır. Bir başka deyişle

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

4. Ortalama sınıflama doğruluğu (OSD) bakımından, incelenen sınıflama kriterlerinin birbirine benzer performans göstermeleri ve oldukça yüksek sınıflama doğruluğuna sahip olmaları sebebiyle üç sınıflama kriteri de uygulamada kullanılabilir. Ancak test etkililiği düşünüldüğünde, OSD ile birlikte OTU'nun da dikkate alınması gerekmektedir. Buna göre bir önceki paragrafta açıklanan sebeplerden ötürü uygulamada sınıflama kriterlerinden GOO veya GA'nın kullanılması önerilebilir.
5. Madde seçme yöntemlerinin hemen hemen tüm koşullarda birbirine benzer performans göstermesi sebebiyle uygulamada MFB veya KLB kullanılabilir. MFB'nin göreli olarak daha iyi çalışmış olması sebebiyle özellikle MFB tercih edilebilir.
6. Madde seçiminin dayandığı temel bakımından, kestirilen yetenekte yüksek bilgi veren madde seçimi (KY), kesme noktasında (KN) yüksek bilgi veren maddenin seçimine kıyasla daha kullanışlıdır. Buna göre uygulamada KY'nin tercih edilmesi önerilebilir.
7. BBST uygulamalarında kestirilen yetenekte en yüksek bilgiyi veren Maksimum Fisher Bilgisi (MFB-KY) madde seçme yönteminin, bu çalışmanın bağımlı değişkenleri olan ortalama test uzunluğu, ortalama sınıflama doğruluğu, kestirilen ve gerçek yetenek düzeyleri arasındaki korelasyon, yanlılık, RMSE ve ortalama mutlak hata bakımından diğer yöntemlere kıyasla daha iyi performans göstermesi sebebiyle, uygulamalarda kullanılması önerilebilir.
8. BBST uygulamalarında daha yansız kestirimler yapabilmesi bakımından, bu durumun da test etkililiğini artıracağı göz önünde bulundurularak, yetenek kestirim yöntemlerinden AOK yerine BSD'nin kullanılması önerilebilir.

KAYNAKÇA

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

Bartroff, J., Finkelman, M. & Lai, T. L. (2008). Modern sequential analysis and its applications to computerized adaptive testing. *Psychometrika, 73(3), 473-486*.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.). *Statistical theories of mental test scores, 397-472*. Massachusetts: Addison-Wesley.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm. *Psychometrika, 46, 443-459.*

Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6 (4), 431-444.*

Boyd, A. M. (2003). *Strategies for controlling testlet exposure rates in computerized adaptive testing systems*. (Unpublished Doctoral Dissertation). The University of Texas.

Boztunç Öztürk, N. (2014). *Bireyselleştirilmiş bilgisayarlı test uygulamalarında madde kullanım sıklığı kontrol yöntemlerinin incelenmesi*. (Yayımlanmamış Doktora Tezi). Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Breslow, N. E. & Holubkov, R. (1997). Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Statistics in Medicine, 16, 103-116.*

Bulut, O. & Kan, A. (2012). Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Eurasian Journal of Educational Research, 49, 61-80.*

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

Cortina, J. M. (1993). What is coefficient Alpha? An examination of theory and applications. *Journal of Applied Psychology, 78, 98-104.*

Çokluk, Ö., Şekercioğlu, G. ve Büyüköztürk, Ş. (2010). *Sosyal bilimler için çok değişkenli istatistik*. (1. baskı). Ankara: Pegem Akademi.

Diao, Q. & Reckase, M. (2009). Comparison of ability estimation and item selection methods in multidimensional computerized adaptive testing. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieved [7.12.2015] from www.psych.umn.edu/psylabs/CATCentral/

Dooley, K. (2002). Simulation research methods. In J. Baum (Ed.). *Companion to organizations, 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, 249-261.*

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

Eggen, T. J. H. M. (2004). Contrıbutıons to The Theory And Practıce of Computerızed Adaptive Testing. Cito Reports. [Online: www.cito.nl/~/media/.../cito_dissertatie_theo_eggen.ashx, Accessed date: 10.7.2017.]

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

Eroğlu, M. G. (2013). *Bireyselleştirilmiş bilgisayarlı test uygulamalarında farklı sonlandırma kurallarının ölçme kesinliği ve test uzunluğu açısından karşılaştırılması.* (Yayımlanmamış Doktora Tezi). Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

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.*

Flaugher, R. (2000). Item pools. In H. Wainer (Ed), *Computerized adaptive testing: A primer*, 37-59. NJ: Lawrence Erlbaum Associates.

Gelbal, S. (1994). p madde güçlük indeksi ile Rasch modelinin b parametresi ve bunlara dayalı yetenek ölçülerine üzerine bir karşılaştırma. *Eğitim Fakültesi Dergisi, 10, 85-94.*

Gökçe, S. (2012). *Comparison of linear and adaptive versions of the Turkish pupil monitoring system (pms) mathematics assessment*. (Unpublished Doctoral Dissertation). Middle East Technical University.

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

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

İşeri, A. I. (2002). *Assessment of students' mathematics achievement through computer adaptive testing procedures*. (Unpublished Doctoral Dissertation). Middle East Technical University.

Jiao, H. & Lau, A. C. (2003). *The Effects of Model Misfit in Computerized Classification Test*. The annual meeting of the National Council of Educational Measurement. Chicago, IL, April 2003. [Online: http://iacat.org/sites/default/files/biblio/ji03-01.pdf, date accessed: 17.3.2014.]

Kalender, İ. (2011). *Effects of different computerized adaptive testing strategies on recovery of ability*. (Unpublished Doctoral Dissertation). Middle East Technical University.

Kaptan, S. (1977). *Bilimsel araştırma teknikleri*. (2. baskı). Ankara: Tekışık Matbaası ve Rehber Yayınevi.

Kaptan, F. (1993). *Yetenek kestiriminde bireyselleştirilmiş test uygulaması ile geleneksel kâğıt-kalem testi uygulaması sonuçlarının karşılaştırılması*. (Yayımlanmamış Doktora Tezi). Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Kaskatı, O. T. (2011). *Rasch modelleri kullanarak romatoid artirit hastaları özürlülük değerlendirimi için bilgisayar uyarlamalı test yöntemi geliştirilmesi*. (Yayımlanmamış Doktora Tezi). Ankara Üniversitesi, Sağlık Bilimleri Enstitüsü, Ankara.

Kelecioğlu, H. (2001). Örtük özellikler teorisindeki b ve a parametreleri ile klasik test teorisindeki p ve r istatistikleri arasındaki ilişki. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 20, 104-110.*

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

Kingsbury, G. G. & Weiss, D. J. (1980). *A Comparison of Adaptive, Sequential and Conventional Testing Strategies for Mastery Decisions*. Research Report 80-4. [Online: http://iacat.org/sites/default/files/biblio/ki80-04.pdf, Accessed date: 21.3.2014.]

Kline, R. B. (2011). *Principles and practice of structural equation modeling*. (3rd Edition). New York: Guilford Publications, Inc.

Köklü, N. (1990). *Klasik test teorisine göre geliştirilen tailored test ile grup testi arasında bir karşılaştırma*. (Yayımlanmamış Doktora Tezi). Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

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

Lau, C. A. & Wang, T. (1998). *Comparing and Combining Dichotomous and Polytomous Items with SPRT Procedure in Computerized Classification Testing*. The annual meeting of the American Educational Research Association. San Diego, CA, 13-17 April 1998. [Online: http://iacat.org/sites/default/files/biblio/la98-01.pdf, Accessed date: 27.2.2014.]

Lau, C. A. & Wang, T. (1999). *Computerized Classification Testing under Practical Constraints with a Polytomous Model*. AERA Annual Meeting. Montreal, Canada, April 1999. [Online: https://eric.ed.gov/?id=ED430032, Accessed date: 27.2.2014.]

Lin, C. J. & Spray, J. (2000). *Effects of Item-Selection Criteria on Classification Testing with the Sequential Probability Ratio Test*. ACT Research Report Series 2000-8. [Online: https://eric.ed.gov/?id=ED445066, Accessed date: 26.2.2014.]

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New Jersey: Lawrence Erlbaum Associates Publishers.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48, 233-245.*

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

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

Nydick, S. W., Nozawa, Y. & Zhu, R. (2012). Accuracy and Efficiency in Classifying Examinees Using Computerized Adaptive Tests: An Application to a Large Scale Test. *The Annual Meeting of the National Council on Measurement in Education*. Vancouver, British Columbia, Canada. April 2012. [Online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.476.3381&rep=rep1&type=pdf, Accessed date: 17.3.2014.]

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

Nydick, S. W. (2014). *catirt: An R Package for Simulating IRT-Based Computerized Adaptive Tests*. [Online: https://cran.r-project.org/web/packages/catIrt/catIrt.pdf, Accessed date: 20.5.2015.]

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70, 351-356.*

Öztuna, D. (2008). *Kas-iskelet sistemi sorunlarının özürlülük değerlendiriminde bilgisayar uyarlamalı test yönteminin uygulanması*. (Yayımlanmamış Doktora Tezi). Ankara Üniversitesi, Sağlık Bilimleri Enstitüsü, Ankara.

Penfield, R. D. & Bergeron, J. M. (2005). Applying a weighted maximum likelihood latent trait estimator to the generalized partial credit model. *Applied Psychological Measurement, 29(3), 218–233.*

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

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

Revelle, W. (2015). psych: Procedures for Psychological, Psychometric, and Personality Research. [Online: https://cran.r-project.org/web/packages/psych/psych.pdf, Accessed date: 20.5.2015.]

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, Monograph No. 17.*

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. [Online: https://eric.ed.gov/?id=ED372078, Accessed date: 20.12.2015.]

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.*

Spray, J. A., Abdel-Fattah, A., Huang, C. & Lau, C. A. (1997). *Unidimensional Approximations for a Computerized Classification Test When the Item Pool and Latent Space Are Multidimensional*. ACT Research Report Series. [Online: https://eric.ed.gov/?id=ED414298, Accessed date: 19.1.2016]

Sulak, S. (2013). *Bireyselleştirilmiş bilgisayarlı test uygulamalarında kullanılan madde seçme yöntemlerinin karşılaştırılması*. (Yayımlanmamış Doktora Tezi). Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using multivariate statistics*. USA: Pearson Education Inc.

Tao, J., Shi, N. Z. & Chang, H. H. (2012). Item-weighted likelihood method for ability estimation in tests composed of both dichotomous and polytomous items. *Journal of Educational and Behavioral Statistics, 37(2), 298-315.*

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 [22.3.2014] from www.psych.umn.edu/psylabs/CATCentral/

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

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

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

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

van der Linden, W. J. (1990). Applications of decision theory to test-based decision making. In R. K. Hambleton & J. N. Zaal (Eds.). *Advances in educational and psychological measurement, 129-156.* Massachusetts: Kluwer-Nijhof.

van Groen, M. M., Eggen, T. J. H. M. & Veldkamp, B. P. (2014). Item selection methods based on multiple objective approaches for classifying respondents into multiple levels. *Applied Psychological Measurement, 38(3), 187-200.*

Wainer, H. (2000). *Computerized adaptive testing: a primer*. (2nd edition). New Jersey: Lawrence Erlbaum Associates.

Wainer, H. & Mislevy, R. J. (2000). Item response theory, item calibration and proficiency estimation. In Wainer, H. (Ed.). *Computerized adaptive testing: a primer, 61-100*. New Jersey: Lawrence Erlbaum Associates Publishers.

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

Wang, T. (1997). Essentially unbiased EAP estimates in computerized adaptive testing. *The annual meeting of the American Educational Research Association Conference*. Chicago, USA. [Online: http://iacat.org/sites/default/files/biblio/wa97-01.pdf, Accessed date: 2.12.2015.]

Wang, T. & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35(2), 109-135.*

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

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

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

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

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

Wouda, J. T. & Eggen, T. J. H. M. (2009). Computerized classification testing in more than two categories by using stochastic curtailment. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieved [20.12.2015] from www.psych.umn.edu/psylabs/CATCentral/

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.*

Yaşar, M. (1999). *Bireyselleştirilmiş testler üzerine bir çalışma*. (Yayımlanmamış Doktora Tezi). Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Yi, Q., Wang, T. & Ban, J. (2000). Effects of Scale Transformation and Test Termination Rule on the Precision of Ability Estimates in CAT. ACT Research Report Series, 2000-2. [Online: http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2001.tb01127.x/full, Accessed date: 7.12.2015.]

EKLER DİZİNİ

EK 1. ETİK KOMİSYON İZİN MUAFİYET FORMU

Form: 40

Tez Çalışması Etik Komisyon İzin Muafiyeti Formu

7 / 11 / 2017

Hacettepe Üniversitesi
Eğitim Bilimleri Enstitüsü
Eğitimde Ölçme ve Değerlendirme Anabilim Dalı Başkanlığı'na

Tez Başlığı / Konusu:	Bireyselleştirilmiş Bilgisayarlı Sınıflama Testi Kriterlerinin Sınıflama Doğruluğu ve Test Uzunluğu Açısından Karşılaştırılması

Yukarıda başlığı/konusu gösterilen tez çalışmam:

1. İnsan ve hayvan üzerinde deney niteliği taşımamaktadır,
2. Biyolojik materyal (kan, idrar vb. biyolojik sıvılar ve numuneler) kullanılmasını gerektirmemektedir.
3. Beden bütünlüğüne müdahale içermemektedir.
4. Gözlemsel ve betimsel araştırma (anket, ölçek/skala çalışmaları, dosya taramaları, veri kaynakları taraması, sistem-model geliştirme çalışmaları) niteliğinde değildir.

Hacettepe Üniversitesi Etik Kurullar ve Komisyonlarının Yönergelerini inceledim ve bunlara göre tez çalışmamın yürütülebilmesi için herhangi bir Etik Komisyondan/Kuruldan izin alınmasına gerek olmadığını; aksi durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini saygılarımla arz ederim.

Ceylan GÜNDEĞER
(Öğrencinin Adı Soyadı, İmzası)

Öğrenci Bilgileri

Adı Soyadı	**Ceylan Gündeğer**
Öğrenci No	N11246191
Anabilim Dalı	Eğitim Bilimleri ABD
Programı	Eğitimde Ölçme ve Değerlendirme BD
Statüsü	☐ Yüksek Lisans ☒ Doktora ☐ Bütünleşik Dr.

Danışman Görüşü ve Onayı

Prof. Dr. Nuri DOĞAN
(İmza)
(Danışmanın Unvanı, Adı ve Soyadı)

EK 2. ORİJİNALLİK RAPORU

HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ
YÜKSEK LİSANS/DOKTORA TEZ ÇALIŞMASI ORİJİNALLİK RAPORU

HACETTEPE ÜNİVERSİTESİ

EĞİTİM BİLİMLER ENSTİTÜSÜ

EĞİTİM BİLİMLERİ ANA BİLİM DALI BAŞKANLIĞI'NA

Tarih: 8 / 11 / 2017

Tez Başlığı: Bireyselleştirilmiş Bilgisayarlı Sınıflama Testi Kriterlerinin Sınıflama Doğruluğu ve Test Uzunluğu Açısından Karşılaştırılması

Yukarıda başlığı verilen tez çalışmamın tamamı (kapak sayfası, özetler, ana bölümler, kaynakça) aşağıdaki filtreler kullanılarak **Turnitin** adlı intihal programı aracılığı ile kontrol edilmiştir. Kontrol sonucunda aşağıdaki veriler elde edilmiştir.

Rapor Tarihi	Sayfa Sayısı	Karakter Sayısı	Savunma Tarihi	Benzerlik Endeksi	Gönderim Numarası
8 / 11 / 2017	93	26,498	13 / 10 / 2017	%3	876447964

Uygulanan filtreler:

1- Kaynakça hariç
2- Alıntılar dâhil
3- 5 kelimeden daha az örtüşme içeren metin kısımları hariç

Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü Tez Çalışması Orijinallik Raporu Alınması ve Kullanılması Uygulama Esasları'nı inceledim ve çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini saygılarımla arz ederim.

8.11.2017

Tarih ve İmza

Adı Soyadı: Ceylan Gündeğer

Öğrenci No: N11246191

Anabilim Dalı: Eğitim Bilimleri ABD

Programı: Eğitimde Ölçme ve Değerlendirme BD

Statüsü: ☐ Y.Lisans ☒ Doktora ☐ Bütünleşik Dr.

DANIŞMAN ONAYI

UYGUNDUR.

Prof. Dr. Nuri DOĞAN

HACETTEPE UNIVERSITY
GRADUATE SCHOOL OF EDUCATIONAL SCIENCES
THESIS/DISSERTATION ORIGINALITY REPORT

HACETTEPE UNIVERSITY
GRADUATE SCHOOL OF EDUCATIONAL SCIENCES
TO THE DEPARTMENT OF EDUCATIONAL SCIENCES

Date: 8 / 11 / 2017

Thesis Title: A Comparison of Computerized Adaptive Classification Test Criteria in Terms of Classification Accuracy and Test Length

The whole thesis that includes the *title page, introduction, main chapters, conclusions and bibliography section* is checked by using **Turnitin** plagiarism detection software take into the consideration requested filtering options. According to the originality report obtained data are as below.

Time Submitted	Page Count	Character Count	Date of Thesis Defence	Similarity Index	Submission ID
8 / 11 / 2017	93	26,498	13 / 10 / 2017	%3	876447964

Filtering options applied:

1. Bibliography excluded
2. Quotes included
3. Match size up to 5 words excluded

I declare that I have carefully read Hacettepe University Graduate School of Educational Sciences Guidelines for Obtaining and Using Thesis Originality Reports; that according to the maximum similarity index values specified in the Guidelines, my thesis does not include any form of plagiarism; that in any future detection of possible infringement of the regulations I accept all legal responsibility; and that all the information I have provided is correct to the best of my knowledge.

I respectfully submit this for approval.

8.11.2017
Date and Signature

Name Surname: Ceylan Gündeğer
Student No: N11246191
Department: Educational Sciences
Program: Measurement and Evaluation in Education
Status: ☐ Masters ☒ Ph.D. ☐ Integrated Ph.D.

ADVISOR APPROVAL

APPROVED

Prof. Dr. Nuri DOĞAN

EK 3. MONTE CARLO SİMÜLASYONU İÇİN TÜRETİLEN MADDE PARAMETRELERİNİN BETİMSEL ÖZELLİKLERİ

	a	*b*	*c*
N	500	500	500
Ortalama	1,24	1,05	0,15
Ortanca	1,22	1,08	0,15
Mod	1,49	0,81	0,14
Standart Sapma	0,45	1,48	0,05
Basıklık	-1,24	0,13	-0,07
Basıklığın Standart Hatası	0,22	0,22	0,22
Çarpıklık	0,04	0,08	0,07
Çarpıklığın Standart Hatası	0,11	0,11	0,11
Ranj	1,5	7,02	0,30
Minimum	0,50	-3,61	0,02
Maksimum	2,00	3,41	0,31

EK 4. MONTE CARLO SİMÜLASYONU İÇİN TÜRETİLEN YETENEK PARAMETRELERİNİN BETİMSEL ÖZELLİKLERİ

	θ
N	3000
Ortalama	-0,00
Ortanca	-0,02
Mod	0,04
Standart Sapma	0,99
Basıklık	-0,09
Basıklığın Standart Hatası	0,09
Çarpıklık	-0,01
Çarpıklığın Standart Hatası	0,05
Ranj	6,67
Minimum	-3,59
Maksimum	3,08

EK 5. POST HOC SİMÜLASYONUNDA KULLANILAN MADDELERİN FAKTÖR YÜKLERİNİN BETİMSEL ÖZELLİKLERİ

	Madde Faktör Yükü
N	80
Ortalama	0,54
Ortanca	0,53
Mod	0,54
Standart Sapma	0,10
Basıklık	-1,06
Basıklığın Standart Hatası	0,53
Çarpıklık	0,17
Çarpıklığın Standart Hatası	0,27
Ranj	0,41
Minimum	0,33
Maksimum	0,74

EK 6. POST HOC SİMÜLASYONUNDA KULLANILAN MADDE HAVUZUNUN 3 PLM TEMELİNDE KESTİRİLEN MADDE PARAMETRELERİNİN BETİMSEL ÖZELLİKLERİ

	a	*b*	*c*
N	80	80	80
Ortalama	1,12	0,91	0,11
Ortanca	1,11	1,16	0,10
Mod	1,11	2,01	0,06
Standart Sapma	0,30	1,11	0,05
Basıklık	0,74	-0,19	0,06
Basıklığın Standart Hatası	0,53	0,53	0,53
Çarpıklık	0,80	-0,79	0,74
Çarpıklığın Standart Hatası	0,27	0,27	0,27
Ranj	1,45	4,23	0,20
Minimum	0,63	-1,68	0,05
Maksimum	2,08	2,55	0,25

EK 7. POST HOC SİMÜLASYONUNDA MADDE TEPKİ KURAMINA DAYALI KESTİRİLEN YETENEK PARAMETRELERİNİN BETİMSEL ÖZELLİKLERİ

	θ
N	994
Ortalama	-0,02
Ortanca	0,08
Mod	0,44
Standart Sapma	0,99
Basıklık	-0,32
Basıklığın Standart Hatası	0,16
Çarpıklık	-0,07
Çarpıklığın Standart Hatası	0,08
Ranj	5,71
Minimum	-2,58
Maksimum	3,13

ÖZGEÇMİŞ

Kişisel Bilgiler

Adı Soyadı	Ceylan Gündeğer
Doğum Yeri	Keçiören / Ankara
Doğum Tarihi	3.11.1986

Eğitim Durumu

Lise	Ankara Anadolu Lisesi / Ankara	2004
Lisans	Hacettepe Üniversitesi / Sınıf Öğretmenliği ABD	2008
Yüksek Lisans	Hacettepe Üniversitesi / Eğitimde Ölçme ve Değerlendirme ABD	2011
Yabancı Dil	İngilizce: Okuma (Çok iyi), Yazma (İyi), Konuşma (İyi) Almanca: Okuma (İyi), Yazma (Orta), Konuşma (Orta)	

İş Deneyimi

Stajlar		
Projeler	Hacettepe Üniversitesi Öğrencileri İçin Biçimlendirmeye Dönük Web Tabanlı Değerlendirme Sisteminin Geliştirilmesi ve Uygulanması	2013-2016
Çalıştığı Kurumlar	M.E.V. Gökkuşağı İlköğretim Okulu Aksaray Üniversitesi Hacettepe Üniversitesi	2008-2009 2010-2011 2011-Devam

Akademik Çalışmalar

Yayınlar (Ulusal, uluslararası makale, bildiri, poster vb gibi.)

Doğan, N. & Gündeğer, C. (2016). The Impact of Item Format on Students' Score and Item Statistics, *International Journal of Research in Engineering and Social Sciences, 6 (11), pp.18-25.*

Gerçek, C., Doğan, N., Gündeğer, C. & Yakar, L. (2017). Effect of health warnings on cigarette pockets on behaviour: *Educational perspective, Eurasian Journal of Educational Research, 17 (68), pp.63-80.*

Seminer ve Çalıştaylar

Sertifikalar

İletişim

e-Posta Adresi	cgundeger@gmail.com

Jüri Tarihi	13.10.2017