



T.C.

AKDENİZ ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ
EĞİTİM BİLİMLERİ
ANA BİLİM DALI

DOKTORA
TEZİ

FARKLI YETENEK KESTİRİM
YÖNTEMLERİYLE BİLGİSAYAR ORTAMINDA
BİREYE UYARLANMIŞ TESTLERİN ÇEŞİTLİ
KOŞULLARA GÖRE İNCELENMESİ

İBRAHİM HAKKI TEZCİ

EĞİTİMDE ÖLÇME VE DEĞERLENDİRME
BİLİM DALI

Antalya, 2025

AKDENİZ ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ
EĞİTİM BİLİMLERİ ANA BİLİM DALI
EĞİTİMDE ÖLÇME VE DEĞERLENDİRME DOKTORA PROGRAMI

FARKLI YETENEK KESTİRİM YÖNTEMLERİYLE BİLGİSAYAR ORTAMINDA
BİREYE UYARLANMIŞ TESTLERİN ÇEŞİTLİ KOŞULLARA GÖRE
İNCELENMESİ
DOKTORA TEZİ

İbrahim Hakkı TEZCİ

Danışman

Prof. Dr. BAYRAM BIÇAK

Antalya, 2025

DOĞRULUK BEYANI

Doktora tezi olarak sunduğum bu çalışmayı, bilimsel ahlak ve geleneklere aykırı düşecek bir yol ve yardıma başvurmaksızın yazdığımı, yararlandığım eserlerin kaynakçada gösterilenlerden oluştuğunu ve bu eserleri her kullanışımında alıntı yaparak yararlandığımı belirtir; bunu onurumla doğrularım. Enstitü tarafından belirli bir zamana bağlı olmaksızın tezimle ilgili bu beyana aykırı bir durumun saptanması hâlinde ortaya çıkacak tüm ahlâkî ve hukukî sonuçlara katlanacağımı bildiririm.

28.02.2025

İBRAHİM HAKKI TEZCİ



T.C.

AKDENİZ ÜNİVERSİTESİ

EĞİTİM BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜNE

İbrahim Hakkı TEZCİ'nin bu çalışması 28.02.2025 tarihinde jürimiz tarafından **Eğitim Bilimleri** Ana Bilim Dalı **Eğitimde Ölçme ve Değerlendirme** Doktora Programında **Doktora Tezi** olarak **oy birliği** ile kabul edilmiştir.

Başkan: Prof. Dr. Selahattin GELBAL

Hacettepe Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü

Üye: Doç. Dr. Eren Can AYBEK

Pamukkale Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü

Üye: Doç. Dr. Güçlü ŞEKERCİOĞLU

Akdeniz Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü

Üye: Doç. Dr. Hakan KOĞAR

Akdeniz Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü

Üye (Danışman): Prof. Dr. Bayram BIÇAK

Akdeniz Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü

DOKTORA TEZİNİN ADI: Farklı Yetenek Kestirim Yöntemleriyle Bilgisayar Ortamında Bireye Uyarlanmış Testlerin Çeşitli Koşullara Göre İncelenmesi

ONAY: Bu tez, Enstitü Yönetim Kurulunca belirlenen yukarıdaki jüri üyeleri tarafından uygun görülmüş ve Enstitü Yönetim Kurulunun tarihli ve sayılı kararıyla kabul edilmiştir.

Doç. Dr. Güçlü ŞEKERCİOĞLU

Enstitü Müdürü

TEŞEKKÜR

Bu doktora tezini tamamlarken, öncelikle bu topraklarda bilimin, aklın ve ilerlemenin ışığını yakan Gazi Mustafa Kemal Atatürk'ü saygı, minnet ve şükranla anmak isterim.

Bu çalışmayı tamamlarken bana destek olan kıymetli danışmanım Prof. Dr. Bayram BIÇAK'a,

Tez sürecinde değerli katkılarıyla bana yol gösteren, kıymetli görüşleri sayesinde çalışmamın bilimsel gücünü artıran doktora tez izleme üyeleri hocalarım Doç. Dr. Eren Can AYBEK ve Doç. Dr. Hakan KOĞAR'a,

Gerek lisansüstü eğitim kapsamımda gerekse doktora tez savunma sürecinde katkılarını bulunan değerli hocalarım Prof. Dr. Selahattin GELBAL ve Doç. Dr. Güçlü ŞEKERCİOĞLU'na,

Tüm akademik hayatım boyunca desteklerini hissettiğim tüm hocalarıma

Tüm tez sürecimde desteklerini hep hissettiğim, yardımlarını sorgusuz ve koşulsuz sunan değerli dostlarıma,

Her daim yanımda olan, sevgileri ve fedakârlıklarıyla bana güç veren canım annem, babam ve ablama,

Tüm eğitim sürecim boyunca her zaman yanımda duran, desteğini hep hissettiğim bu süreci benimle yaşayan sevgili eşim yol arkadaşım Dilek ÇALIŞKAN TEZCİ'ye ve hayatımıza girdiği andan itibaren huzur ve mutluluk kaynağımız oğlum Tuna'ya çok teşekkür ederim.

Türk milletinin bağımsızlık ruhuna ve milli düşünceye olan inancımla, bu çalışmanın ülkemizin bilimsel gelişimine katkı sağlamasını dilerim. Büyük Önder Atatürk'ün işaret ettiği muasır medeniyet seviyesine ulaşma yolunda, bilimin ve aklın ışığında yürümeye devam edeceğim.

İBRAHİM HAKKI TEZCİ

ÖZET

FARKLI YETENEK KESTİRİM YÖNTEMLERİYLE BİLGİSAYAR ORTAMINDA BİREYE UYARLANMIŞ TESTLERİN ÇEŞİTLİ KOŞULLARA GÖRE İNCELENMESİ

TEZCİ, İbrahim Hakkı

Doktora Tezi, Eğitim Bilimleri Ana Bilim Dalı

Tez Danışmanı: Prof. Dr. Bayram BIÇAK

Şubat 2025, xiv + 185 sayfa

Bu çalışma, bilgisayar ortamında bireye uyarlanmış testlerin (BOBUT) farklı koşullarda ve yetenek kestirim yöntemleri altında gösterdiği performans analiz etmeyi amaçlamaktadır. Araştırma kapsamında, BOBUT sisteminin post-hoc simülasyon sonuçları kullanılarak farklı yetenek kestirim yöntemlerinin doğruluk, güvenilirlik ve kullanılabilirlik üzerindeki etkileri değerlendirilmiştir. Elde edilen bulgular, canlı BOBUT uygulamalarında yetenek kestirim yöntemlerinin performans özelliklerini karşılaştırmak amacıyla kullanılmıştır. Ayrıca çalışmada BOBUT ve doğrusal testler çeşitli kriterler açısından karşılaştırılarak her iki test türünün avantajları ve sınırlılıkları ortaya konmuştur. Tüm bu uygulamalar, BOBUT'un farklı yetenek kestirim yöntemleriyle gerçek ve simülasyon koşullarında test edilmesini sağlamak amacıyla temel araştırma modeline dayalı olarak tasarlanmıştır.

Araştırma kapsamında iki farklı çalışma grubu oluşturulmuştur. İlk çalışma grubunda madde test kalibrasyonu ve post-hoc simülasyonlar gerçekleştirilirken ikinci çalışma grubunda canlı BOBUT ve doğrusal test uygulamaları yürütülmüştür. Okuduğunu anlama becerisini ölçmeye yönelik 155 maddelik bir madde havuzu geliştirilmiş ve 1502 ortaokul öğrencisinden elde edilen verilerle çeşitli başlama kuralları ($\theta=0$, $\theta=-0,5:0,5$), madde seçim yöntemleri (MEI, MFI, KL), yetenek kestirim yöntemleri (ML, EAP, MAP) ve sonlandırma kuralları (standart hata ve sabit madde sayısı) kombinasyonları değerlendirilmiştir. Post-hoc simülasyon sürecinin ardından 263 öğrenciye üç farklı yetenek kestirim yöntemi kullanılarak canlı BOBUT ve 25 maddelik doğrusal test uygulamaları yapılmış, bu yöntemlerin performansları karşılaştırılmıştır.

Sonuçlar, başlama kurallarının test performansı üzerinde simülasyon süresi ve madde örtüşme oranları dışında minimal bir etkisi olduğunu göstermiştir. Madde seçim yöntemleri

arasında hesaplama süresi bakımından en kullanışlı yöntem MFI, olarak belirlenirken MEI'nın diğer iki yönteme göre önemli ölçüde daha uzun işlem süreleri gerektirdiği gözlenmiştir. MEI, MFI ve KL yöntemlerine kıyasla daha kısa test uzunlukları sağlamıştır. Tüm madde seçim yöntemleri benzer korelasyon değerleri göstermiş olup bu durum, yetenek kestirim doğruluğu açısından yöntemlerin birbirine yakın performans sergilediklerine işaret etmektedir. Bununla birlikte KL yöntemi daha düşük RMSE değerleri sağlayarak daha yüksek ölçme hassasiyeti sunmuştur.

Sonlandırma kurallarına ilişkin analizler, standart hata ile sabit madde sayısı kombinasyonunun en doğru sonuçları sağladığını, ancak daha uzun simülasyon süreleri gerektirdiğini göstermiştir. 0,50 standart hata kriteri ile 25 madde kombinasyonu, doğruluk ve kullanışlılık arasında optimum bir denge sunmuştur. Yetenek kestirim yöntemleri arasında MAP en hızlı yöntem olurken, ML en uzun hesaplama süresine ihtiyaç duymuştur. EAP, en düşük bias ve RMSE değerlerini sağlayarak özellikle kısa testlerde daha yüksek güvenilirlik sunmuştur. MAP ve EAP, ayrıca daha düşük madde örtüşme oranları göstererek madde havuzunun daha etkin kullanımına olanak tanımıştır.

Canlı BOBUT ve doğrusal test karşılaştırmalarında, özellikle MAP ve EAP yöntemlerini kullanan BOBUT'un daha güvenilir ve kullanışlı yetenek kestirimleri sağladığı belirlenmiştir. ML yöntemi, uç yetenek düzeylerinde daha fazla hata üretirken, MAP yöntemi daha dengeli ve kararlı kestirimler sunmuştur. Doğrusal testler, yetenek kestirimlerinde daha yüksek değişkenlik göstermiş ve daha uzun test sürelerine ihtiyaç duymuştur. Buna karşın uygun başlama kuralları, madde seçim stratejileri ve sonlandırma kriterleri ile optimize edilen BOBUT'un doğruluk, güvenilirlik ve kullanışlılık açısından önemli avantajlar sunduğu görülmüştür. Ayrıca BOBUT doğrusal testlere kıyasla daha yüksek ölçme bilgisi sağlamış ve özellikle orta yetenek düzeylerinde daha hassas ölçümler gerçekleştirmiştir. MAP yöntemi en yüksek test bilgisini sunarken ML yöntemi görece daha düşük bilgi düzeyi sağlamıştır.

Canlı BOBUT uygulamalarının yetenek kestirimleri arasında genel olarak güçlü korelasyonlar bulunurken BOBUT ile doğrusal test arasındaki ilişkilerin nispeten daha zayıf olduğu görülmüştür. ML ve EAP yöntemleri en yüksek korelasyonu gösterirken doğrusal test ile en düşük korelasyon MAP, en yüksek korelasyon ise EAP yöntemleriyle gözlemlenmiştir. Bu bulgular, BOBUT'un kendi içinde daha tutarlı sonuçlar verdiğini, ancak doğrusal test ile kıyaslandığında farklılıklar olabileceğini ortaya koymaktadır.

Test süresi ve madde sayısı açısından, MAP yöntemi en kullanışlı sonuçları sağlarken ML yöntemi en uzun test süresi ve en fazla madde kullanımına ihtiyaç duymuştur. BOBUT, daha az madde kullanarak ölçme sürecini optimize etmiş ve bireylerin yetenek düzeyine uygun

madde seçimini daha etkin bir şekilde gerçekleştirmiştir. Doğrusal testlerin sabit madde yapısı nedeniyle her bireye aynı maddeleri sunması, ölçme sürecinde esnekliği azaltırken, BOBUT'un adaptif yapısı bireylerin yetenek düzeylerine uygun sorular sunarak ölçme doğruluğunu artırmıştır.

Anahtar Kelimeler: *Bilgisayar ortamında bireye uyarlanmış testler, madde tepki kuramı, yetenek kestirim yöntemleri, post-hoc simülasyon, doğrusal test.*



ABSTRACT

EXAMINING COMPUTERIZED ADAPTIVE TESTING UNDER VARIOUS CONDITIONS USING DIFFERENT ABILITY ESTIMATION METHODS

TEZCİ, İbrahim Hakkı

Ph.D. Thesis, Department of Educational Sciences

Supervisor: Prof. Dr. Bayram BIÇAK

February 2025, xiv + 185 pages

This study aims to analyze the performance of computerized adaptive tests (CAT) under different conditions and ability estimation methods. Within the scope of the research, the effects of various ability estimation methods on accuracy, reliability, and efficiency were evaluated using post-hoc simulation results of the CAT system. The findings obtained were used to compare the performance characteristics of ability estimation methods in live CAT applications. Additionally, the study compared CAT and linear tests in terms of various criteria, highlighting the advantages and limitations of both test types. All these applications were designed based on a fundamental research model to test CAT under real and simulation conditions with different ability estimation methods.

Within the scope of the research, two different study groups were formed. In the first study group, item test calibration and post-hoc simulations were conducted, while in the second study group, live CAT and linear test applications were carried out. A 155-item pool was developed to measure reading comprehension skills, and data obtained from 1502 middle school students were used to evaluate various starting rules ($\theta=0$, $\theta=-0.50:0.50$), item selection methods (MEI, MFI, KL), ability estimation methods (ML, EAP, MAP), and termination rules (standard error and fixed item number) combinations. Following the post-hoc simulation process, live CAT and a 25-item linear test were administered to 263 students using three different ability estimation methods, and the performance of these methods was compared.

The results showed that starting rules had a minimal effect on test performance, except for simulation time and item overlap rates. In terms of item selection methods, MFI was identified as the most efficient method in terms of computation time, while MEI required significantly longer processing times. MEI provided shorter test lengths compared to MFI and KL methods. All item selection methods showed similar correlation values, indicating that they

performed similarly in terms of ability estimation accuracy. However, the KL method provided lower RMSE values, offering higher measurement precision.

Analyses of termination rules showed that the combination of standard error and fixed item number provided the most accurate results but required longer simulation times. The combination of a 0.50 standard error criterion and 25 items offered an optimal balance between accuracy and efficiency. Among the ability estimation methods, MAP was the fastest method, while ML required the longest computation time. EAP provided the lowest bias and RMSE values, offering higher reliability, especially in short tests. MAP and EAP also showed lower item overlap rates, allowing for more effective use of the item pool.

In comparisons between live CAT and linear tests, it was found that CAT, particularly when using MAP and EAP methods, provided more reliable and efficient ability estimations. The ML method produced more errors at extreme ability levels, while the MAP method offered more balanced and stable estimations. Linear tests showed higher variability in ability estimations and required longer test durations. In contrast, CAT, optimized with appropriate starting rules, item selection strategies, and termination criteria, offered significant advantages in terms of accuracy, reliability, and efficiency. Additionally, CAT provided higher measurement information compared to linear tests and performed more precise measurements, especially at medium ability levels. The MAP method provided the highest test information, while the ML method offered relatively lower information levels.

Strong correlations were generally observed among live CAT applications, while the relationships between CAT and linear tests were relatively weaker. The ML and EAP methods showed the highest correlations, while the lowest correlation with the linear test was observed with the MAP method, and the highest correlation was observed with the EAP method. These findings indicate that CAT provides more consistent results within itself but may show differences when compared to linear tests.

In terms of test duration and item number, the MAP method provided the most efficient results, while the ML method required the longest test duration and the highest number of items. CAT optimized the measurement process by using fewer items and more effectively selecting items appropriate to individuals' ability levels. The fixed item structure of linear tests, which presents the same items to every individual, reduces flexibility in the measurement process, while the adaptive structure of CAT increases measurement accuracy by presenting questions tailored to individuals' ability levels.

Keywords: *Computerized adaptive testing, item response theory, ability estimation methods, post-hoc simulation, linear test.*

İÇİNDEKİLER

TEŞEKKÜR.....	i
ÖZET	i
ABSTRACT	iv
TABLolar LİSTESİ	xi
ŞEKİLLER LİSTESİ	xii
KISALTMALAR LİSTESİ	xiv
BÖLÜM I	1
GİRİŞ.....	1
1.1. Problem Durumu.....	1
1.2.Araştırmanın Amacı ve Alt Problemler	6
1.3. Araştırmanın Önemi	7
1.4. Araştırmanın Sınırlılıkları.....	8

BÖLÜM II

KURAMSAL ÇERÇEVE VE İLGİLİ ARAŞTIRMALAR

2.1. Klasik Test Kuramı.....	9
2.2. Geleneksel Kâğıt Kalem-Doğrusal (Lineer) Testler	12
2.3. Madde Tepki Kuramı.....	15
2.3.1. Madde Tepki Kuramının Temel Özellikleri.....	15
2.3.2 Madde Tepki Kuramının Matematiksel Modellemesi	18
2.3.3. Madde Tepki Kuramının Varsayımları	22
2.3.4. Madde Tepki Kuramının Avantaj Dezavantajları	26
2.4.Bilgisayar Ortamında Bireye Uyarlanmış Testler	28
2.4.1. Bilgisayar Ortamında Bireye Uyarlanmış Testlerin Tarihçesi	28
2.4.2. Bilgisayar Ortamında Bireye Uyarlanmış Test Süreci	29

2.4.3. Başlangıç Yetenek Düzeyi	30
2.4.4. Madde Seçim Yöntemleri.....	31
2.4.5. Yetenek Kestirim Yöntemleri	36
2.4.6. Testi Sonlandırma Kuralları	39
2.4.7. Bilgisayar Ortamında Bireye Uyarlanmış Testlerin Avantaj ve Dezavantajları	42
2.5. İlgili Araştırmalar	43
2.5.1. Bilgisayar ortamında bireye uyarlanmış ve geleneksel kalem-kâğıt testlerinin karşılaştırılmasına ait çalışmalar	43
2.5.2. Eğitimde BOBUT kullanımına ilişkin görüş ve deneyimleri içeren çalışmalar.....	45
2.5.3. Farklı test stratejilerinin simülasyonla karşılaştırılmasına ait çalışmalar.....	46
2.5.4. Bilgisayar ortamında bireye uyarlanmış testlerin farklı alanlardaki uygulamaları üzerine çalışmalar.....	48
2.5.5. Hem simülasyon hem de canlı BOBUT uygulaması içeren çalışmalar	49

BÖLÜM III

YÖNTEM

3.1. Araştırmanın Modeli	53
3.2. Çalışma Grubu	54
3.3. Veri Toplama Aracı	55
3.3.1. Test Geliştirme Süreci.....	55
3.3.2. Verilerin düzenlenmesi	56
3.3.3. MTK Varsayımlarının İncelenmesi.....	64
3.3.4. Madde ve Test İstatistikleri	71
3.4. Post-hoc simülasyon süreci.....	77
3.5. Canlı BOBUT ve Doğrusal Test Uygulama Süreci	82
3.6. Çevrimiçi BOBUT Uygulama Platformu-CONCERTO	83
3.7. Verilerin Analizi	86

BÖLÜM IV

BULGULAR

4.1. Post-hoc Simülasyon Sonuçlarına İlişkin Bulgular	87
4.1.1. Post-hoc simülasyon genel performans sonuçları	87
4.1.2. Post-hoc simülasyon tüm koşulların beraber değerlendirilmesine yönelik bulgular	94
4.1.3. Canlı BOBUT uygulaması koşullarının belirlenmesine yönelik bulgular	102
4.2. Canlı BOBUT ve doğrusal test uygulamalarına yönelik bulgular	103
4.2.1. Yetenek kestirimlerine yönelik bulgular	103
4.2.2. Standart hata değerlerine yönelik bulgular.....	106
4.2.3. Güvenirlilik değerlerine yönelik bulgular	108
4.2.4. Test uzunluklarına yönelik bulgular.....	112
4.2.5. Test süresine yönelik bulgular.....	114
4.2.6. Madde kullanım sıklıklarına yönelik bulgular	116

BÖLÜM V

SONUÇ, TARTIŞMA VE ÖNERİLER

5.1. Sonuç ve Tartışma	118
5.1.1. Post-hoc simülasyon sonucu başlama kuralı, madde seçim yöntemi ve sonlandırma kurallarına ait sonuç ve tartışma.....	118
5.1.2. Post-hoc simülasyon sonucu yetenek kestirim yöntemlerinin genel performansları ve başlama kuralı, madde seçim yöntemi ve sonlandırma kurallarıyla olan kombinasyonlarına ait sonuç ve tartışma	121
5.1.3. Canlı BOBUT tüm yetenek kestirim yöntemleri için geçerli ortak kuralların belirlenmesine yönelik sonuç ve tartışma	123
5.1.4. Canlı BOBUT ve doğrusal testlerin yetenek kestirim performanslarına yönelik sonuç ve tartışma.....	124

5.1.5. Yetenek kestirim yöntemlerinin Canlı BOBUT standart hata performanslarına ait sonuç ve tartışma.....	126
5.1.6. Canlı BOBUT ve doğrusal testlerin güvenirlik performanslarına yönelik sonuç ve tartışma.....	127
5.1.7. Canlı BOBUT ve doğrusal testlerin test uzunluklarına yönelik sonuç ve tartışma	128
5.1.8. Canlı BOBUT ve doğrusal testlerin ölçüm süresine yönelik sonuç ve tartışma ...	129
5.1.9. Yetenek kestirim yöntemlerinin Canlı BOBUT madde kullanım durumlarına ait sonuç ve tartışma.....	131
5.2. Öneriler	132
5.2.1. Araştırmacılara Öneriler.....	132
5.2.2. Test Uygulayıcılarına Öneriler.....	134
5.2.3. Kurumlara Öneriler	135
KAYNAKÇA.....	137
EK-1.Kazanımlar.....	155
EK-2. Madde, Kazanım ve Bilişsel Alan Örnekleri	156
EK-3. Veri1-Veri6 Karşılaştırmasına Ait Mantel-Haenszel DIF Analizi Sonuçları.....	162
EK-4. Okuduğunu Anlama Madde Havuzuna Ait Faktör Yükleri.....	164
EK-5. Madde Parametrelerinin Değişmezliği İçin Yapılan DIF Sonuçları.....	165
EK-6 Madde Parametrelerinin Değişmezliği için Rastgele Bölünen İki Grup için Elde Edilen Madde Parametreleri.....	167
EK-7 Madde Parametrelerinin Değişmezliği için Rastgele Bölünen İki Grup için Elde Edilen Madde Karakteristik Eğrileri	169
EK-8. Okuduğunu Anlama Testi Madde Havuzuna Ait Madde Parametreleri ve Uyum Değerleri	171
EK-9. Okuduğunu Anlama Testi Madde Havuzuna Ait Madde Karakteristik Eğrileri	174
EK-10. Okuduğunu Anlama Testi Madde Havuzuna Ait Maddi Bilgi Fonksiyonları.....	175

EK-11.Tüm Simülasyon Koşullarına İlişkin Sonuçlar	176
EK-12. ETİK KURUL VE MEB ARAŞTIRMA İZİN BELGELERİ	179
ÖZ GEÇMİŞ	181
BENZERLİK RAPORU	184
BİLDİRİM.....	185



TABLÖLAR LİSTESİ

Tablo 3.1. Çalışma Gruplarına İlişkin Sınıf ve Cinsiyet Dağılımları	54
Tablo 3.2. Veri1-Veri6 Madde Parametrelerinin Karşılaştırılması	59
Tablo 3.3. Okuduğunu Anlama Testine Ait Test İstatistikleri.....	62
Tablo 3.4. Okuduğunu Anlama Testinin Bileşenlerine Ait Öz Değerler.....	65
Tablo 3.5. Okuduğunu Anlama Madde Havuzuna Ait Faktör Yükleri.....	65
Tablo 3.6. Q3 İstatistik Sonuçları ve Uyum Değerleri	67
Tablo 3.7. Model-Veri Uyum Değerleri	68
Tablo 3.8. Madde Parametrelerinin Değişmezliği için Madde Parametrelerinin Karşılaştırılması.....	69
Tablo 3.9. Yetenek Parametrelerinin Değişmezliği için Oluşturulan Alt Testler.....	70
Tablo 3.10. Yetenek Parametrelerinin Değişmezliği için Oluşturulan Alt Testler Arasında Korelasyon Katsayıları	71
Tablo 4.1. Canlı BOBUT için Seçilen Kriterlerin Post-Hoc Sonuçları	103
Tablo 4.2. Uygulamalar Arası Yetenek Düzeyi Ortalama Karşılaştırma Sonuçları	104
Tablo 4.3. Yetenek Kestirim Yöntemlerinin Standart Hata Değerlerinin Karşılaştırma Sonuçları.....	108
Tablo 4.4. Uygulamalar Arası Test Bilgi Düzeylerinin Karşılaştırma Sonuçları	111
Tablo 4.5. Uygulamalar Arası Test Uzunluklarının Karşılaştırma Sonuçları.....	113
Tablo 4.6. Uygulamalar Arası Test Süreleri Karşılaştırma Sonuçları	114
Tablo 4.7. Kestirim Yöntemlerine Göre Madde Kullanım Sıklığı Değerleri	116

ŞEKİLLER LİSTESİ

Şekil 2.1. Madde Karakteristik Eğrisi	16
Şekil 2.2. Bireye Uyarlanmış Test Adımları	30
Şekil 3.1. a ve b Parametrelerine Ait Dağılım Grafikleri	60
Şekil 3.2. a ve b parametrelerine ait Bland-Altman grafikleri	61
Şekil 3.3. Okuduğunu Anlama Testi Puan Dağılımları	63
Şekil 3.4. Paralel Analiz ve Yamaç Birikinti Grafiği	66
Şekil 3.5. Tüm Maddelere Ait Madde Karakteristik Özellikleri	72
Şekil 3.6. M47, M114 ve M1'e ait madde karakteristik eğrileri	73
Şekil 3.7. Teste ve Maddelere Ait Bilgi Fonksiyonları	75
Şekil 3.8. Madde47 ve M114'e Ait Madde Bilgi Fonksiyonları	76
Şekil 4.1. Post-Hoc Simülasyon Koşullarına Göre Ortalama Simülasyon Süreleri	87
Şekil 4.2. Post-Hoc Simülasyon Koşullarına Göre Ortalama Test Uzunlukları	88
Şekil 4.3. Post-Hoc Simülasyon Koşullarına Göre Gerçek ve Kestirilen Yetenek Arasındaki Ortalama Korelasyon Değerleri	89
Şekil 4.4. Post-hoc simülasyon koşullarına göre ortalama RMSE değerleri	90
Şekil 4.5. Post-Hoc Simülasyon Koşullarına Göre Ortalama Bias Değerleri	91
Şekil 4.6. Post-hoc simülasyon koşullarına göre ortalama madde örtüşme oranları	91
Şekil 4.7. Post-Hoc Simülasyon Koşullarına Göre Ortalama Standart Hata Değerleri	92
Şekil 4.8. Post-Hoc Simülasyon Koşullarına Göre Ortalama Madde Bilgisi Değerleri	93
Şekil 4.9. Yetenek Kestirim Yöntemlerinin Simülasyon Koşullarında Simülasyon Süreleri ...	94
Şekil 4.10. Yetenek Kestirim Yöntemlerinin Simülasyon Koşullarında Ortalama Test Uzunlukları	95
Şekil 4.11. Yetenek Kestirim Yöntemlerinin Simülasyon Koşullarında Gerçek ve Kestirilen Yetenek Arasındaki Korelasyon Değerleri	96
Şekil 4.12. Yetenek Kestirim Yöntemlerinin Simülasyon Koşullarında RMSE Değerleri	97
Şekil 4.13. Yetenek Kestirim Yöntemlerinin Simülasyon Koşullarında BIAS Değerleri	98

Şekil 4.14. Yetenek Kestirim Yöntemlerinin Simülasyon Koşullarında Madde Örtüşme Oranları.....	99
Şekil 4.15. Yetenek Kestirim Yöntemlerinin Simülasyon Koşullarında Ortalama Standart Hata Değerleri.....	100
Şekil 4.16. Yetenek Kestirim Yöntemlerinin Simülasyon Koşullarında Ortalama Madde Bilgi Değerleri	101
Şekil 4.17. Yetenek Kestirimleri Arası İlişki Düzeyleri	104
Şekil 4.18. Yetenek Kestirim Yöntemlerinin Standart Hata Dağılımları.....	106
Şekil 4.19. Yetenek Kestirim Yöntemlerinin Yetenek Düzeylerine Göre Standart Hata Değerleri	107
Şekil 4.20. Yetenek Kestirim Yöntemlerinin Yetenek Düzeylerine Göre Test Bilgi Değerleri	109
Şekil 4.21. Uygulamalara Ait Test Bilgi Düzeyleri	110
Şekil 4.22. Yetenek Kestirim Yöntemlerinin Yetenek Düzeylerine Göre Test Uzunlukları .	112
Şekil 4.23. Yetenek Kestirim Yöntemlerinin Yetenek Düzeylerine Göre Toplam Test Süreleri	114

KISALTMALAR LİSTESİ

AFA: Açımlayıcı Faktör Analizi

BOBUT: Bilgisayar Ortamında Bireye Uyarlanmış Test

DIF: Değişen Madde Fonksiyonu/DMF (Differential Item Functioning)

DOK: Norman Webb'in Bilgi Derinliği (Depth of Knowledge)

EAP: Beklenen A Posteriori Tahmini (Expected A Posteriori Estimation)

KL: Kullback-Leibler (Kullback-Leibler Divergence)

KTK: Klasik Test Kuramı

MAP: Maksimum A Posteriori Tahmini (Maximum A Posteriori Estimation)

MEI: Maksimum Beklenen Bilgi (Maximum Expected Information)

MFI: Maksimum Fisher Bilgisi (Maximum Fisher Information)

MKE: Madde Karakteristik Eğrisi

ML: Maksimum Olabilirlik (Maximum Likelihood)

MTK: Madde Tepki Kuramı

PIRLS: Uluslararası Okuma Okuryazarlığında Gelişim Araştırması (Progress in International Reading Literacy Study)

PISA: Uluslararası Öğrenci Değerlendirme Programı (Programme for International Student Assessment)

RMSE: Kök Ortalama Kare Hatası (Root Mean Square Error)

TIMMS: Uluslararası Matematik ve Fen Eğilimleri Araştırması (Trends in International Mathematics and Science Study)

BÖLÜM I

GİRİŞ

Bu bölümde araştırmanın problem durumu, araştırmanın amacı ve alt problemleri, araştırmanın önemi ve araştırmaya ait sınırlılıklara yer verilmiştir.

1.1. Problem Durumu

Bireylerin gerçek yeteneklerini doğru ve objektif bir şekilde belirlemek, eğitimden iş hayatına kadar birçok süreçte alınacak kararların doğruluğunu ve etkinliğini etkiler. Yeteneklerin doğru bir şekilde ölçülmesi, bireylerin potansiyellerini anlamalarına, güçlü ve zayıf yönlerini tespit etmelerine, dolayısıyla kişisel ve profesyonel başarılarını artırmalarına olanak tanır (O'Boyle Jr & Aguinis, 2012). Eğitimde, bireylerin bireysel farklılıklarını dikkate alarak yönlendirilmesi ve bu farklılıklara uygun öğrenme fırsatlarının sunulması hem birey hem de toplum için büyük faydalar sağlar (Jonassen & Grabowski, 2012). Ayrıca, fırsat eşitliği ilkesi doğrultusunda adaletli bir değerlendirme süreci oluşturulması, bireylerin değerlendirme süreçlerine olan güvenini artırırken ve ayrımcılık ihtimalini azaltır (MacCann & Roberts, 2008). Geleneksel testlerin sabit yapısı ise bu gereksinimlere tam anlamıyla yanıt verememekte ve bireysel farklılıkları yeterince dikkate almamaktadır. Bu durum, bireylere daha kişiselleştirilmiş ve hassas ölçüm yapabilecek yenilikçi test yöntemlerine olan ihtiyacı artırmaktadır.

BOBUT, bireylerin yeteneklerini değerlendirme sürecinde son yıllarda önemli bir alternatif olarak öne çıkmaktadır. BOBUT, bireyin verdiği her yanıtı bağlı olarak testin zorluk seviyesini dinamik bir şekilde uyarlamakta ve bu sayede bireyin gerçek yetenek seviyesini daha doğru bir şekilde tahmin etmektedir (van der Linden & Glas, 2000). Bu adaptif yapı bireylerin farklı bilgi, beceri ve yetenek seviyelerine sahip olduğunu dikkate alarak, sabit ve standart bir test yerine bireyin bireysel özelliklerine uyarlanmış bir test deneyimi sunar. Böylece test, bireyin yeteneklerine uygun bir zorluk seviyesinde ilerlerken bireyin gerçek performansını yansıtmaya potansiyelini de taşır (Hambleton vd., 1991). Bu yaklaşım, bireylerin zorluk seviyesi çok yüksek veya çok düşük sorularla karşılaşarak motivasyon kaybı yaşamalarını önlerken, değerlendirme sürecinin daha etkili ve güvenilir bir şekilde gerçekleştirilmesini sağlar.

BOBUT'un bu adaptif yapısı, bireysel farklılıkları anlamada ve değerlendirmede sağladığı hassasiyet ile geleneksel kâğıt kalem testlerinden ayrılmaktadır. Geleneksel testlerde tüm bireylere aynı sorular sunulurken, BOBUT bireyin yetenek seviyesini tespit ederek bireysel düzeyde daha uygun ve anlamlı bir test deneyimi sunar (Thompson & Weiss, 2019). Örneğin, düşük yetenek düzeyindeki bir birey çok zor sorularla karşılaştığında testin devamına ilişkin motivasyonu azalabilir veya yüksek yetenek düzeyindeki bir birey çok kolay sorularla sınılandığında gerçek potansiyelini yansıtamayabilir. BOBUT ise bu durumu ortadan kaldırarak bireylerin test sürecinde karşılaştıkları zorluk seviyesini optimize eder ve bireyin yetenek seviyesini en doğru şekilde belirlemeyi hedefler.

Ayrıca BOBUT'un dijital platformlarda uygulanması, ölçme ve değerlendirme süreçlerinde geleneksel yöntemlere kıyasla önemli avantajlar sunmaktadır (Embretson & Reise, 2000; Reckase, 2010;). Dijital altyapı sayesinde, bireyler testi zaman ve mekân sınırlaması olmaksızın tamamlayabilir ve sonuçlar gerçek zamanlı olarak analiz edilip raporlanabilir. Özellikle büyük ölçekli uygulamalarda bu özellik, süreçlerin verimliliğini artırmakta ve insan kaynaklı hataları en aza indirmektedir. Ayrıca dijital test ortamlarının sağladığı bir diğer önemli avantaj, bireylere geri bildirim sunma hızıdır. Geleneksel kâğıt kalem testlerinde sonuçların değerlendirilmesi günler, hatta haftalar alabilirken BOBUT ile bireyler test tamamlandıktan hemen sonra sonuçlarına ulaşabilir. Bu hızlı geri bildirim, bireylerin güçlü ve zayıf yönlerini anında fark etmelerine ve öğrenme süreçlerini buna göre yeniden yapılandırmalarına olanak tanır (Hausler & Sommer, 2008).

BOBUT'un bu kişiselleştirilmiş yapısı, özellikle eğitimde bireysel öğrenme eksikliklerini tespit etme ve öğrenme süreçlerini optimize etme açısından büyük bir potansiyele sahiptir. Örneğin, düşük performans sergileyen bireyler için belirli öğrenme alanlarına yönelik daha fazla pratik yapma önerileri sunulabilirken yüksek performans gösteren bireyler için daha zorlu ve gelişim sağlayıcı içerikler önerilebilir. Bu şekilde, BOBUT sadece bir ölçme aracı olarak değil, aynı zamanda bireysel öğrenme süreçlerini yönlendiren bir rehber olarak da kullanılabilir (Weiss & Şahin, 2024).

BOBUT'un başarısında yetenek kestirim yöntemlerinin oynadığı kritik rol göz ardı edilemez. MAP (Maximum A Posteriori), EAP (Expected A Posteriori) ve ML (Maximum Likelihood) gibi yöntemler, bireylerin yeteneklerini tahmin etmek için kullanılan en yaygın yaklaşımlardır (Baker & Kim, 2004; Embretson & Reise, 2000). Her yöntemin temel varsayımları, avantajları ve sınırlılıkları farklıdır. MAP bireyin yetenek seviyesini, gözlemlenen yanıt desenine ve öncül bir yetenek dağılımına dayalı olarak tahmin eder. MAP

yöntemi, özellikle küçük veri setlerinde daha istikrarlı sonuçlar sunabilmesiyle öne çıkar. Ancak bu yöntemin performansı öncül dağılımın doğru bir şekilde belirlenmesine bağlıdır (Baker & Kim, 2004). EAP, MAP yöntemine benzer şekilde, öncül dağılım kullanır ancak tahmini, yanıt desenlerinin beklenen değeri üzerinden gerçekleştirir. Bu yöntem, büyük veri setlerinde ve düşük yetenek varyanslarında etkili sonuçlar sunabilir (van der Linden, 2005). ML ise bireyin yetenek seviyesini, gözlemlenen yanıt desenine dayalı en yüksek olasılıklı tahmin olarak hesaplar. ML, öncül dağılım bilgisine ihtiyaç duymaz ve genellikle büyük veri setlerinde daha doğru sonuçlar sağlar. Ancak, özellikle küçük örneklem gruplarında tahminlerin güvenilirliği azaltırken özellikle hep doğru veya hep yanlış yanıt örüntülerinin kestiriminde problemler yaşamaktadır (Embretson ve Reise, 2000). Belirli test koşulları altında hangi yöntemin daha doğru ve güvenilir sonuçlar sunduğunu anlamak için bu yöntemlerin karşılaştırılması gerekmektedir. Örneğin, bir yöntemin düşük yetenek varyansına sahip bireylerde daha başarılı olması, diğer yöntemlerin ise daha geniş yetenek yelpazesini etkili bir şekilde ölçebilmesi, BOBUT'un farklı uygulama senaryolarında hangi yöntemin tercih edilmesi gerektiğine dair önemli bilgiler sunar.

BOBUT'un etkinliği, yalnızca kullanılan yetenek kestirim yöntemleri ile sınırlı değildir; test başlama kuralları, madde seçim stratejileri ve test sonlandırma kuralları gibi temel parametreler de sürecin başarısını doğrudan etkiler (van der Linden & Glas, 2010). Bu parametreler, testin bireye ne kadar doğru bir şekilde uyum sağlayacağını ve bireyin gerçek yetenek seviyesini ne kadar hassas bir şekilde tahmin edebileceğini belirler. Örneğin, testin başlangıcında kullanılan zorluk seviyesinin uygunluğu, bireyin teste olan motivasyonunu ve test sürecindeki performansını büyük ölçüde etkileyebilir. Bu nedenle, başlama kurallarının bireyin muhtemel yetenek düzeyine uygun şekilde optimize edilmesi, hem bireyin teste olan ilgisini sürdürmesi hem de doğru sonuçlar elde edilmesi açısından kritik bir öneme sahiptir (Hambleton vd., 1991).

Benzer şekilde, madde seçim algoritmaları, testin bireysel farklılıklara ne kadar duyarlı olduğunu belirleyen en önemli bileşenlerden biridir. Bu algoritmalar, bireyin verdiği her yanıttan sonra bir sonraki soruyu seçerken bireyin güncel yetenek tahminine ve testin hedeflediği bilgi kazancına dayanır. Amaç, bireyin yetenek seviyesine en uygun zorluk seviyesindeki soruları seçerek hem bireyin test deneyimini optimize etmek hem de testin doğruluğunu artırmaktır (Kingsbury & Hauser, 2004). Yanlış seçilen bir madde, bireyin yetenek düzeyinin yüksek hata miktarıyla tahmin edilmesine yol açabilir ve testin genel geçerliğini olumsuz etkileyebilir. Örneğin, kolay soruların sürekli seçilmesi, yüksek yetenekli

bireylerin gerçek performansını yansıtmazken, zor soruların sürekli seçilmesi düşük yetenekli bireylerde test sürecini gereksiz yere uzatabilir. Dolayısıyla madde seçim algoritmalarının bireysel farklılıkları dikkate alacak şekilde geliştirilmesi, testin hem birey için daha etkili hem de genel anlamda daha güvenilir olmasını sağlar (Thompson & Weiss, 2019).

Test sonlandırma kuralları da BOBUT'un başarısında önemli bir rol oynar. Sonlandırma kuralları, testin ne zaman tamamlanacağını belirler ve genellikle bireyin yetenek seviyesinin yeterince doğru bir şekilde tahmin edildiği noktada testi sonlandırmayı amaçlar. Bu süreçte genellikle hata marjı, bilgi miktarı veya sabit madde sayısı gibi kriterler kullanılır (Weiss & Kingsbury, 1984). Sonlandırma kurallarının iyi tanımlanması, test sürecinin gereksiz yere uzamasını önlerken bireyin gerçek yetenek seviyesinin doğru bir şekilde tahmin edilmesini de sağlar. Aynı zamanda testin erken sonlandırılması, bireyin yetenek seviyesinin tam olarak yansıtılmasını engeller. Bunun aksine gereğinden uzun süre devam eden bir test, bireyin motivasyonunu kaybetmesine ve testin verimliliğinin düşmesine neden olabilir. Bu nedenle, sonlandırma kurallarının, bireyin yetenek seviyesinin yeterince doğrulukla belirlendiği noktada testi sonlandıracak şekilde tasarlanması büyük önem taşır (van der Linden & Pashley, 2010).

Simülasyon verileri, bu tür yöntemlerin performansını kontrol altında incelemek için güçlü bir araçtır. Simülasyonlar, çeşitli test koşullarını modelleme ve yöntemlerin performanslarını farklı senaryolarda değerlendirme olanağı sunar (Rutten vd., 2012). Simülasyon verilerinin avantajları arasında, değişkenlerin manipüle edilebilir olması, farklı parametre kombinasyonlarının kolayca test edilebilmesi ve sürecin hızlı bir şekilde tamamlanması yer alır (Salas vd., 2009). Ancak simülasyonlar gerçek dünya koşullarını tam anlamıyla yansıtamayabilir. Gerçek veri, bireylerin karmaşık yanıt davranışlarını ve bireysel farklılıklarını daha iyi analiz etme imkânı sunarken simülasyonlar bu etkileşimleri yeterince temsil edemez (Railback & Grimm, 2019). Bu durum, gerçek veri üzerinde yapılan analizlerin önemini artırmaktadır.

Gerçek veri, simülasyonların eksikliklerini tamamlayarak daha geniş ve gerçekçi bir perspektif sunar. Canlı BOBUT uygulamaları, farklı yetenek kestirim yöntemlerinin bireyler üzerindeki gerçek performansını değerlendirmek için kritik bir alan sağlar. Bu tür analizler, yöntemlerin doğruluk ve güvenilirliğini gerçek dünya koşullarında anlamak için benzersiz bir fırsat yaratır. Ayrıca bu yöntemlerin test süresi, test uzunluğu ve bireyler arasındaki farklılıkları yansıtma kapasiteleri gibi ölçütler üzerindeki etkilerinin değerlendirilmesi, testlerin pratikte ne kadar etkili olduğuna dair daha fazla bilgi sağlar. Böylece canlı uygulamalar, yalnızca teorik

modelleri test etmekle kalmaz, aynı zamanda testlerin gerçek dünya bağlamında nasıl performans gösterdiğinin anlaşılmasına yardımcı olur.

Bu çalışma, BOBUT'un performansını farklı koşullar ve yetenek kestirim yöntemleri bağlamında analiz etmeyi amaçlamaktadır. Araştırmada, BOBUT'un post-hoc simülasyon sonuçları incelenerek yetenek kestirim yöntemlerinin doğruluk, güvenilirlik ve kullanılabilirlik üzerindeki etkileri kapsamlı bir şekilde değerlendirilmiştir. Elde edilen bulgular doğrultusunda, farklı yetenek kestirim yöntemlerinin performans özellikleri canlı BOBUT uygulamaları aracılığıyla karşılaştırılmıştır. Bunun yanı sıra, BOBUT ile doğrusal testler, çeşitli ölçütler açısından karşılaştırılarak her iki test türünün avantaj ve sınırlılıkları ortaya konulmuştur. Bu analizlerin, bireylerin bireysel farklılıklarını dikkate alan, daha hassas ve etkili ölçme araçlarının geliştirilmesine yönelik önemli bilgiler sağlaması hedeflenmektedir. Çalışmanın sonuçlarının hem eğitim alanına hem de diğer ölçme ve değerlendirme süreçlerine teorik ve pratik katkılar sunması beklenmektedir.

1.2.Araştırmanın Amacı ve Alt Problemler

Bu çalışma, bilgisayar ortamında bireye uyarlanmış testlerin farklı uygulama koşulları altında gerçekleştirilmesini ve bu koşulların canlı BOBUT, post-hoc simülasyon uygulamaları ve doğrusal test sonuçları ile karşılaştırılmasını amaçlamaktadır. Çalışma kapsamında aşağıdaki sorulara yanıt aranmıştır:

1. Post-hoc simülasyon sonuçları çerçevesinde:
 - Farklı test başlama kuralları, madde seçim yöntemleri, yetenek kestirim yöntemleri ve test sonlandırma kurallarına ait kombinasyonlar yetenek tahmini doğruluğunu nasıl etkilemektedir?
 - Bu kombinasyonların BOBUT uygulama koşullarındaki performansları nasıldır?
2. Post-hoc simülasyon sonuçlarına göre bilgisayar ortamında bireye uyarlanmış testlerde farklı yetenek tahmin yöntemlerine ilişkin:
 - En uygun test başlama kuralı nedir?
 - En uygun madde seçim yöntemi nedir?
 - En uygun test sonlandırma kuralı nedir?
3. Farklı yetenek tahmin yöntemleri kullanılan canlı BOBUT uygulamaları sonucunda:
 - Yetenek kestirimi, standart hata ve güvenilirlik değerleri nasıl değişmektedir?
 - Yetenek kestirimleri arasındaki ilişkiler nasıldır?
 - Yetenek kestirimleri arasında farklar var mıdır?
 - Test uzunlukları arasında farklar var mıdır?
 - Test süreleri arasında farklar var mıdır?
 - Madde kullanım sıklıkları nasıl dağılım göstermektedir?
4. Bilgisayar ortamında bireye uyarlanmış testler ile doğrusal testlerin karşılaştırılması sonucunda:
 - Yetenek kestirimleri arasındaki ilişkiler nasıldır?
 - Yetenek kestirimleri nasıl değişmektedir?
 - Güvenirlik değerleri arasında fark var mıdır?
 - Test uzunlukları arasında fark var mıdır?
 - Test süreleri arasında fark var mıdır?

1.3. Araştırmanın Önemi

Günümüzde dijital teknolojilerin eğitim ve ölçme değerlendirme alanındaki etkisi hızla artmaktadır. Bu teknolojilerden biri olan BOBUT, bireylere yetenek seviyelerine uygun maddeler sunarak ölçme doğruluğunu artırma ve test sürelerini optimize etme potansiyeline sahiptir. Ancak bu yöntemin etkili bir şekilde kullanılabilmesi için test başlama kuralları, madde seçim algoritmaları ve test sonlandırma kriterlerinin farklı koşullar altında nasıl performans gösterdiğinin bilimsel olarak incelenmesi gereklidir. Bu çalışma hem simülasyon hem de canlı uygulamalara dayalı olarak bu parametrelerin sistematik bir değerlendirmesini sunarak BOBUT'un daha etkin kullanılmasına katkıda bulunacağı düşünülmektedir.

Araştırma, BOBUT'un farklı yetenek kestirim yöntemleri ile uygulama sonuçlarını karşılaştırmalı olarak ele alması bakımından önemli görülmektedir. Bu durum, sadece testlerin doğruluk ve güvenilirliğini değerlendirmekle kalmayıp farklı test uygulama türleri (canlı BOBUT, post-hoc simülasyonlar ve doğrusal testler) arasında kapsamlı bir kıyaslama yapma imkânı sunmaktadır. Bu kapsamda yapılan incelemelerin, eğitim bilimleri alanında bilgisayarla bireyselleştirilmiş ölçme sistemlerinin geçerliği ve güvenilirliği konusunda önemli metodolojik bilgiler sağlayacağı düşünülmektedir.

Bu çalışma, BOBUT'un farklı yetenek tahmin yöntemlerinin güvenilirlik ve doğruluk açısından nasıl performans gösterdiğini detaylı bir şekilde incelemektedir. Elde edilen bulguların, testlerin psikometrik özelliklerinin daha iyi anlaşılmasını mümkün kılarak ölçme alanında önemli bir boşluğu dolduracağı düşünülmektedir. Bu durumun hem akademik araştırmalar hem de pratik uygulamalar için temel oluşturabilecek sonuçların üretilmesini sağlayacağı ön görülmektedir.

Araştırmada canlı BOBUT, simülasyon ve doğrusal test sonuçlarının karşılaştırılması, bu test türleri arasındaki ilişkilerin ve farklılıkların bilimsel bir temelde açıklanmasını sağlayacaktır. Özellikle farklı test formatlarının sonuçları nasıl etkilediği ve bireylerin yetenek kestirimlerinin bu formatlara göre nasıl değiştiği konusundaki bulguların, eğitim alanında doğru ölçme stratejileri geliştirmek için bir rehber niteliğinde olacağı düşünülmektedir.

Eğitimde dijitalleşme süreci hızla ilerlerken bireyselleştirilmiş ölçme yöntemleri için geçerli ve güvenilir modeller geliştirilmesi kritik bir ihtiyaçtır. Bu çalışmanın yalnızca mevcut uygulamaları değerlendirmekle kalmayacağı gelecekteki test geliştirme süreçleri için de önemli bir rehber olacağı düşünülmektedir. Özellikle eğitim teknolojisi firmaları ve test geliştiricileri,

bu çalışmadan elde edilen bulgularla BOBUT sistemlerini daha etkili ve kullanıcı dostu bir hâle getirebilirler.

Bu araştırmanın, bireye uyarlanmış testlerin teorik temelleri olan madde tepki kuramı ve yetenek tahmini yöntemleri üzerine yapılan bilimsel çalışmalara önemli bir katkı sunacağı düşünülmektedir. Özellikle yetenek tahmini algoritmalarının performansı, madde seçim süreçlerinin etkinliği ve test sürelerinin optimizasyonu gibi konular, psikometrik araştırmaların temel soruları arasında yer almaktadır. Bu bağlamda elde edilen bulgular hem yerel hem de uluslararası düzeyde ölçme biliminde yapılan çalışmalara katkı sağlayacaktır.

1.4. Araştırmanın Sınırlılıkları

1. Araştırmada toplanan olan veriler Antalya ilinde 2023-2024 eğitim-öğretim yılında okuyan 7. ve 8. sınıf ortaokul öğrencileriyle
2. Uygulama kapsamında ilk aşamada yapılan uygulamalar kâğıt kalem testi aracılığıyla olacağı için madde havuzu 155 soruyla
3. Araştırmanın ölçme aracını Türkçe okuduğunu anlama testi oluşturmuş olup ölçme aracının içeriği, kazanım ve soru dağılımları ortaöğretim 7. ve 8. sınıf okuduğunu anlama kazanımlarıyla
4. Post-hoc simülasyon koşullarında başlama kuralı olarak iki farklı yöntem ($\theta=0$ ve $\theta=-0,5:0,5$), madde seçim yöntemi olarak üç farklı yöntem (MFI, MEI ve KL), yetenek kestirim yöntemi olarak üç farklı yöntem (ML, EAP ve MAP) ve sonlandırma kuralı olarak beş farklı yöntem (standart hata = 0,50, standart hata = 0,30, sabit 25 madde, hem standart hata = 0,50 hem sabit 25 madde, hem standart hata = 0,30 hem sabit 25 madde) ile sınırlı tutulmuştur.

BÖLÜM II

KURAMSAL ÇERÇEVE VE İLGİLİ ARAŞTIRMALAR

2.1. Klasik Test Kuramı

Klasik Test Kuramı (KTK), ölçme biliminin temel taşlarından biri olup, bireylerin gözlenen test puanlarının gerçek yetenek düzeyleri ile ölçme hatalarının toplamı olduğu varsayımına dayanır (Crocker & Algina, 1986; Turgut & Baykul, 2010). Eğitim, psikoloji ve insan kaynakları gibi alanlarda yaygın olarak kullanılan bu kuram, test geliştirme süreçlerinde uzun süre hâkim olmuş ve bireylerin performansını toplam puanlar üzerinden değerlendiren bir yöntem sunmuştur (Crocker & Algina, 1986). KTK, yalnızca bireylerin genel performansını ölçmekle kalmayıp, test maddelerinin analiz edilmesi ve testlerin geliştirilmesine de katkı sağlamıştır. Thorndike (1971), KTK'nın yetenek ölçümünde önemli bir çerçeve sunduğunu ve testlerin geçerliliğini artırmada toplam puan yaklaşımının kullanılmasını savunmuştur. Basit ve anlaşılır yapısı sayesinde, modern ölçme uygulamalarında hala referans alınan bir model olarak kabul edilmektedir.

Gözlenen, Gerçek ve Hata Puanları

KTK'nın merkezinde, gözlenen puan (X), gerçek puan (T) ve hata puanı (E) arasındaki ilişkiye dayalı basit ama güçlü bir model yer alır. Bu model, gözlenen bir test puanını, bireyin gerçek yetenek düzeyinin bir tahmini ve ölçüm sürecindeki rastlantısal hataların bir toplamı olarak görür. Crocker ve Algina (1986), bu modeli şu şekilde ifade etmişlerdir:

$$X = T + E$$

Gerçek puan, bireyin ölçülebilir yeteneğini temsil ederken, hata puanı ölçme sürecindeki rastlantısal sapmaları ifade eder (Baker, 2001). Örneğin, bir öğrencinin sınav performansı gerçek bilgi düzeyini yansıtsa da dikkat dağınıklığı veya kaygı gibi faktörler hata puanına katkıda bulunabilir. Hata puanlarının rastlantısal olduğu ve sıfır ortalamalı dağıldığı varsayımı, KTK'nın temel dayanaklarından biridir (Crocker & Algina, 1986). Ancak, sistematik hatalar bu varsayımı ihlal ederek test sonuçlarının geçerliliğini etkileyebilir. KTK'nın sunduğu şeffaf yapı, ölçme süreçlerinde yaygın kullanımını sağlamış ve hata analizleriyle geçerlik ve güvenirlik çalışmalarında önemli bir araç olmuştur.

Güvenirlik ve Ölçme Hataları

Güvenirlik, KTK'nın temel unsurlarından biri olup ölçme sürecinin tutarlılığını ifade eder ve testin farklı zamanlarda veya uygulamalarda benzer sonuçlar üretme yeteneğini gösterir (Algina & Swaminathan, 2015). Kuder ve Richardson (1937), güvenilirlik katsayısını hesaplamak için KR-20 formülünü geliştirerek KTK'ya önemli bir katkı sağlamışlardır. Güvenirlik katsayısı 0 ile 1 arasında değişir ve yüksek değerler testin güvenilirliğinin yüksek olduğunu gösterir. Cronbach (1951), güvenilirliğin bireylerin gerçek yetenek düzeylerini tahmin etmede kritik bir rol oynadığını vurgulamıştır. Ancak, güvenilir bir testin mutlaka geçerli olduğu anlamına gelmez (De Wellis, 2006). KTK'nın, rastlantısal hataların sıfır ortalamalı ve simetrik dağıldığı varsayımı güvenilirlik hesaplamalarında temel alınır (Brennan, 2010), ancak sistematik hatalar bu varsayımı zayıflatabilir. Bu nedenle, güvenilirlik analizi, ölçme hatalarını anlamak ve kontrol etmek açısından büyük önem taşır.

Geçerlik: Testlerin Ölçme Amaçlarına Uygunluğu

Geçerlik, bir testin ölçmek istediği özelliği ne kadar doğru ölçtüğünü belirleyen önemli bir unsurdur (Brown, 2013; Crocker & Algina, 1986). Cronbach (1951), geçerliği ölçme biliminde temel bir gereklilik olarak tanımlamış ve test sonuçlarının nasıl kullanıldığına da dikkat çekmiştir. KTK kapsamında geçerlik; kapsam geçerliği, ölçüt geçerliği ve yapı geçerliği olmak üzere üç temel başlıkta ele alınır (Brown, 2013). Kapsam geçerliği, test maddelerinin ölçülmek istenen alanı ne kadar iyi kapsadığını değerlendirir. Örneğin, bir matematik testi müfredattaki tüm konuları eşit kapsamazsa düşük kapsam geçerliğine sahip olabilir. Ölçüt geçerliği, test sonuçlarının bir dış kriterle ilişkisini belirler. Örneğin, bir işe alım sınavının adayların iş performansını doğru tahmin etmesi ölçüt geçerliğine bağlıdır. Yapı geçerliği ise testin teorik bir özelliği veya yeteneği ne kadar iyi ölçtüğünü değerlendirir. Ancak KTK, geçerlik analizlerini çoğunlukla toplam test puanına dayandırdığı için bazı sınırlılıklar taşır. Özellikle maddelerin tek tek analiz edilmemesi, testin gerçekten hedeflenen özelliği ne kadar iyi ölçtüğünü belirlemede yetersiz kalabilir (Hambleton vd., 1991).

Toplam Test Puanına Bağımlılık

KTK'nın önemli sınırlılıklarından biri, bireylerin yetenek düzeylerini yalnızca toplam test puanına dayanarak değerlendirmesidir. Bu yaklaşım, tüm test maddelerinin yeteneği eşit derecede ölçtüğünü varsayarak, maddeler arasındaki güçlük ve ayırt edicilik farklılıklarını göz ardı eder. Lord ve Novick (1968), toplam test puanına bağımlılığın ölçme sürecini basitleştirdiğini ancak yetenek seviyelerinin ayrıntılı analizini zorlaştırdığını belirtmiştir.

Örneğin, kolay ve zor maddelerin eşit puan değerine sahip olması, bireylerin gerçek yeteneklerini tam olarak yansıtmada eksikliklere yol açabilir. Ayrıca, toplam puanlar bireylerin hangi maddelerde başarılı veya başarısız olduğunu göstermediğinden, öğretim süreçlerinde geri bildirim sağlama açısından sınırlılıklar içerir. Bu durum, bireyselleştirilmiş öğrenme yaklaşımlarına uyum sağlamayı zorlaştırırken, modern ölçme yaklaşımları madde bazlı analizlere odaklanarak bu eksikliği gidermeye çalışmaktadır. Ancak, KTK hâlâ toplam test puanına dayalı bir çerçeve sunmaya devam etmektedir.

Maddelerin Eşit Öneme Sahip Olması Varsayımı

KTK'nın bir diğer sınırlılığı, testteki tüm maddelerin bireylerin yetenek düzeylerini eşit derecede ölçtüğünü varsaymasıdır. Bu varsayım, maddeler arasındaki güçlük veya ayırt edicilik farklılıklarını dikkate almaz (Embretson & Reise, 2000). Örneğin, bir testin bazı maddeleri bireylerin yetenek düzeylerini daha iyi ayırt edebilirken, diğer maddeler bu yetenek düzeylerini belirlemede yetersiz kalabilir. Gleser vd. (1965), KTK'nın bu varsayımının, ölçme süreçlerinde doğruluk ve hassasiyet eksikliğine yol açabileceğini ifade etmişlerdir. Ayrıca, düşük ayırt ediciliğe sahip maddelerin toplam puana olan etkisi, bireylerin gerçek yetenek düzeylerini yansıtmada sorunlara neden olabilir. Modern psikometri yaklaşımları, bu eksiklikleri gidermek için madde bazlı analizlere odaklanmış ve her bir maddenin ölçme gücünü ayrı ayrı değerlendiren yöntemler geliştirmiştir. Ancak KTK, tüm maddeleri eşit kabul eden yapısıyla bu tür ayrıntılı analizlere imkân tanımaz.

Klasik Test Kuramının Sağladığı Avantajlar

KTK, ölçme süreçlerinde sağladığı basitlik ve uygulanabilirlik nedeniyle birçok avantaj sunar. Öncelikle, KTK'nın matematiksel yapısı kolayca anlaşılabilir ve uygulanabilir olduğu için, test geliştirme ve analiz süreçlerinde yaygın olarak kullanılmaktadır. Kuder ve Richardson (1937) gibi bilim insanlarının geliştirdiği güvenilirlik katsayısı hesaplama yöntemleri, ölçme araçlarının kalite kontrolünde KTK'nın etkili bir araç olmasını sağlamıştır. Ayrıca, KTK, toplam test puanına dayalı bir yaklaşım sunarak, bireylerin performansını hızlı ve kolay bir şekilde değerlendirme olanağı tanır (Crocker & Algina, 1986). Bu, özellikle eğitim sistemlerinde geniş ölçekli testlerin uygulanmasında önemli bir avantajdır. Ayrıca, KTK, norm-referanslı değerlendirme sistemlerine uyumlu yapısıyla bireylerin performansını bir grup içinde sıralama olanağı sunar (Algina & Swaminathan, 2015). Bu özellik, bireylerin yetenek düzeylerini görece olarak anlamayı mümkün kılar. Bununla birlikte, KTK'nın sade yapısı, testlerin geniş popülasyonlarda uygulanabilirliğini artırır ve maliyetleri düşürür. Bu avantajlar,

KTK'nın eğitim, psikoloji ve sağlık gibi birçok alanda yaygın bir şekilde kullanılmasını sağlamıştır.

Klasik Test Kuramının Sağladığı Dezavantajları ve Eleştiriler

KTK'nın sağladığı avantajlar, aynı zamanda bazı sınırlılıklar ve eleştirilerle birlikte gelir. En temel dezavantajlarından biri, toplam puana aşırı bağımlılığıdır. Tüm maddelerin bireylerin yetenek düzeyine eşit katkıda bulunduğu varsayımı, gerçek dünyadaki testlerde genellikle geçerli değildir. Örneğin, bir testin bazı maddeleri yüksek ayırt ediciliğe sahipken, diğerleri düşük ayırt edicilik gösterebilir. Bu durum, bireylerin toplam puanlarının yetenek düzeylerini tam olarak yansıtmamasına neden olabilir (Embretson & Reise, 2000). Ayrıca, KTK, madde parametrelerinin örneklemden bağımsız olmadığını varsayar, bu da farklı popülasyonlarda aynı test sonuçlarının geçerliliğini zayıflatabilir. Sistemik hataların hesaba katılmaması, özellikle önyargılı maddeler içeren testlerde büyük bir sorun oluşturur. Örneğin, bir test, belirli bir kültürel gruba avantaj sağlıyorsa, KTK bu önyargıyı tespit etmekte yetersiz kalabilir. Bu sınırlılıklar, daha modern psikometrik yaklaşımların, özellikle Madde Tepki Kuramı'nın (MTK), KTK'ya alternatif olarak geliştirilmesine yol açmıştır. MTK, maddeleri bağımsız bir şekilde analiz ederek, KTK'nın eksikliklerini gidermeyi hedefler. Ancak bu eleştirilere rağmen, KTK, ölçme biliminin temel taşlarından biri olarak varlığını sürdürmekte ve birçok alanda kullanılmaya devam etmektedir.

2.2. Geleneksel Kâğıt Kalem-Doğrusal (Lineer) Testler

Standartlaştırılmış testlerin uygulama biçimlerinden biri olan geleneksel kâğıt kalem testleri, bireylerin bilgi, beceri ve yeteneklerini ölçmek için kullanılan en eski ölçme araçlarından biridir (Cohen vd., 1996). Bu testler, önceden belirlenmiş soruların yazılı veya sözlü olarak yöneltilmesi ve cevapların ölçülmesi yoluyla yapılır. Bu testler, öğrencilerin başarısını değerlendirmek, öğrenme eksikliklerini belirlemek ve öğretmenlerin eğitim programını geliştirmek için kullanılır (Crocker & Algina, 1986). Geleneksel testlerin en büyük avantajı, yaygın olarak kullanılması ve herkesin aynı koşullarda test edilmesini sağlamasıdır (Thorndike, 1971). Geleneksel kâğıt kalem testleri genellikle çoktan seçmeli sorular, doğru/yanlış soruları ve kısa cevaplı sorular gibi çeşitli soru türlerini içerir. Bu soruların tasarımı ve soru sayısı, öğrencilerin belirli bir konuda ne kadar bilgiye sahip olduğunu değerlendirmek için tasarlanır. Örneğin, bir matematik testinde, öğrencilerin aritmetik işlemler yapabilme yetenekleri, problem çözme becerileri ve matematiksel kavramlar hakkındaki bilgileri test edilebilir. Geleneksel kâğıt kalem testleri, özellikle akademik becerilerin

ölçülmesinde yaygın olarak kullanılır. Bu testler, öğrencilerin okuma, yazma, matematik ve bilim gibi konulardaki bilgilerini ölçmek için kullanılır. Bunun yanı sıra, bilişsel beceriler, kişilik özellikleri ve mesleki ilgi alanları gibi diğer alanlarda da kâğıt kalem testleri kullanılmaktadır. Bu testlerin, standart ölçme araçları olmaları ve belirli bir sınav formülasyonu kullanmaları nedeniyle geçerli, güvenilir ve objektif oldukları düşünülmektedir (Kaplan & Saccuzzo, 2017; Linn, 1989).

Geleneksel kâğıt kalem testleri uzun bir zamandır uygulama alanı olan ve sahip olduğu özellikler sayesinde çeşitli avantajları içeren testlerdir. Geleneksel kâğıt kalem testlerinin avantajlarını şu şekilde sıralayabiliriz (Anastasi & Urbina, 1997; Brookhart & Nitko, 2011; Crocker & Algina, 1986):

- *Güvenilirlik:* Geleneksel kâğıt kalem testleri, ölçüm sonuçlarına yüksek güvenilirlik sağlar. Bu testler, standart soruların kullanılması nedeniyle sonuçların tekrarlanabilirliğini ve güvenilirliğini artırır.
- *Kolaylık:* Geleneksel kâğıt kalem testleri, kolayca uygulanabilir ve sınavın sonucu hızlı bir şekilde elde edilebilir. Bu testler, özellikle büyük gruplarda kullanıldığında, sonuçların hızlı bir şekilde elde edilmesine olanak tanır.
- *Esneklik:* Geleneksel kâğıt kalem testleri, farklı konularda ve farklı zorluk seviyelerinde kullanılabilir. Bu testler, ölçülen konuya göre özelleştirilebilir ve farklı zorluk seviyelerinde sorular içerebilir.
- *Uygun Maliyet:* Geleneksel kâğıt kalem testleri, diğer ölçme araçlarına kıyasla düşük maliyetlidir. Bu testler, basit bir şekilde hazırlanabilir ve çoğu kez sadece bir kâğıt, kalem ve cevap anahtarı gerektirir.
- *Yaygın Kabul:* Geleneksel kâğıt kalem testleri, yaygın kabul gören bir ölçme yöntemidir. Bu testler, özellikle okullar, iş yerleri ve askeri kurumlarda sıklıkla kullanılır.

Bununla birlikte, kâğıt kalem testlerinin dezavantajları da vardır. Bu dezavantajlarını çeşitli araştırmacılardan hareketle şu şekilde özetleyebiliriz (Brookhart, 2010; Brown & Hirschfeld, 2008; McMillan & Hellsten, 2010; Nitko, 2013; Pellegrino, Chudowsky & Glaser, 2001; Sheppard, 2000)

- *Yetersiz Ölçüm:* Bu uygulamalar genellikle sınırlı sayıda soruya dayanır ve sadece belirli bir konu veya yetenek alanını kapsar. Bu durum, öğrencinin gerçek yeteneklerini tam olarak yansıtmayabilir ve performanslarını tam olarak değerlendiremeyebilir.
- *Yüzeysel Değerlendirme:* Genellikle nesnel ve ölçülebilir sonuçlar sağlar. Ancak, öğrencinin derinlemesine anlayışını, yaratıcı düşünme yeteneklerini ve problem çözme becerilerini tam olarak değerlendirmekte sınırlı olabilir. Bu tür testler, öğrencilerin sadece bilgiyi ezberleme ve tekrarlama yeteneklerini ölçmeye odaklanabilir.
- *Sınav Kaygısı:* Sınav kaygısı gibi faktörler nedeniyle öğrenciler üzerinde baskı oluşturabilir. Bu, öğrencilerin gerçek potansiyellerini göstermelerini engelleyebilir. Bazı öğrenciler, sınav stresi altında çalışma performanslarını olumsuz etkileyebilir.
- *Dilde ve Kültürdeki Farklılıklar:* Uluslararası bir ortamda, farklı dil ve kültürlerden gelen öğrenciler arasında kâğıt kalem testlerinin adil bir şekilde uygulanması zor olabilir. Dil bariyerleri veya kültürel farklılıklar, öğrencilerin soruları anlamalarını veya beklenen yanıtları verememelerini etkileyebilir.
- *Sınırlı Geri Bildirim:* Genellikle nihai bir puanlama veya not verme sürecini içerir, ancak öğrencilere gerçek zamanlı ve ayrıntılı geri bildirim sağlama imkânı sınırlı olabilir. Bu da öğrencilerin zayıf yönlerini tanımlayıp geliştirmelerini engelleyebilir.
- *Sınırlı Bağlam:* Genellikle kontrollü bir ortamda yapılır ve gerçek dünya bağlamından uzak olabilir. Bu durum, öğrencilerin gerçek hayatta karşılaştıkları durumları veya sorunları çözme becerilerini tam olarak değerlendiremeyebilir. Uygulamada kullanılan bilgileri ve becerileri test etmek için daha gerçekçi bağlamlar gerekebilir.
- *Kopya Riski:* Geleneksel kâğıt kalem testlerinde kopya çekme riski mevcuttur. Öğrenciler arasında yanıtları paylaşma veya hile yapma potansiyeli, test sonuçlarının doğruluğunu etkileyebilir. Bu durum, ölçülen başarıyı tam olarak yansıtmayabilir ve adil bir değerlendirme sağlamayabilir.
- *Tek Boyutluluk:* Geleneksel kâğıt kalem testleri, genellikle tek bir değerlendirme boyutuna odaklanır. Örneğin, sadece doğru yanıtların sayısını ölçebilir ve derinlemesine anlayışı veya çoklu becerileri değerlendirmekte sınırlı kalabilir. Başarının çok yönlü ve kapsamlı bir şekilde ölçülmesi için daha çeşitli değerlendirme yöntemleri ve araçları kullanılması gerekebilir.

- *Standartlaşma Sorunları:* Geleneksel kâğıt kalem testlerinin standartlaştırılması ve güvenilirlikleri bazen sorunlara neden olabilir. Farklı testler arasında tutarlılık sağlamak, testlerin her öğrenci için eşit ve adil bir şekilde uygulanması ve sonuçların güvenilir bir şekilde yorumlanması zor olabilir.
- *Özgün Yetenekleri Ölçememe:* Bu testler genellikle genel yetenekleri ve standart becerileri ölçmeye odaklanırken, öğrencilerin özgün yeteneklerini veya ilgi alanlarını tam olarak yansıtmayabilir. Örneğin, sanatsal, müziksel veya spor yetenekleri gibi alanlarda öğrencilerin potansiyelini değerlendirmekte sınırlı olabilir.

Ayrıca, test sonuçlarının objektifliğinin garantisi yoktur ve farklı öğretmenlerin veya sınav görevlilerinin puanlama farklılıkları nedeniyle sonuçlar farklılık gösterebilir. Bunun yanı sıra öğrencilerin sınavlarda stres veya kaygı yaşamaları ve bu nedenle performanslarının gerçek potansiyellerini yansıtmaması da mümkündür (Bridgeman, 2009). Diğer bir dezavantajı ise, kâğıt kalem testlerinin sadece belirli bir zamanda ve belirli bir yerde yapılabiliyor olmasıdır. Yani, testlerin sonuçları bireylerin o anda ne kadar iyi bir performans sergilediğini yansıtabilir, ancak bireylerin performansı zaman içinde değişebilir. Bu nedenle, kâğıt kalem testlerinin ölçtüğü beceri veya yeteneklerin sabit olduğu varsayımı doğru değildir ve performansın zaman içinde takip edilmesi gerekmektedir (Bugbee, 1996).

2.3. Madde Tepki Kuramı

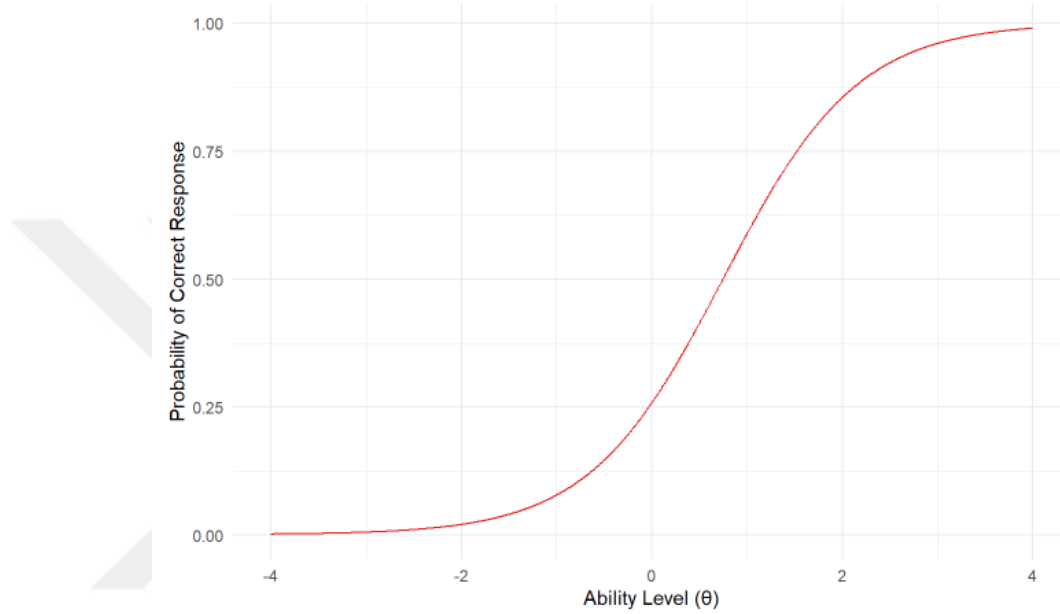
2.3.1. Madde Tepki Kuramının Temel Özellikleri

20. yüzyılın ortalarında KTK'nın sınırlılıklarına bir çözüm olarak ortaya çıkan MTK, bireylerin yanıtlarına dayalı olarak yeteneklerini ve maddelerin özelliklerini analiz etmeyi mümkün kılar (Embretson & Reise, 2000). Bu kuramın temelinde, bir bireyin bir test maddesine doğru cevap verme olasılığının, bireyin yetenek düzeyine ve maddenin psikometrik özelliklerine bağlı olduğu varsayımı yer alır (Hambleton vd., 1991). MTK, geleneksel test yaklaşımlarının aksine, toplam test puanından ziyade her bir maddenin yanıtını detaylı bir şekilde ele alır. Bu özellik, ölçümlerin daha hassas yapılmasını sağlar. KTK'da tüm bireylerin aynı maddelere maruz kalması ve puanlamanın toplam puana dayalı olması, bireyler arasındaki gerçek farklılıkları ölçme konusunda sınırlılıklar yaratabilir (Hambleton & Swaminathan, 1985). Ancak MTK, bireyin yetenek düzeyi ile madde özelliklerini ilişkilendirerek, her bir maddeye verilen yanıtın bireysel ölçümde nasıl bir katkı sunduğunu detaylı bir şekilde analiz

eder. MTK'nın bu yenilikçi yaklaşımı, özellikle eğitim, sağlık ve insan kaynakları gibi alanlarda bireyselleştirilmiş ölçüm sistemleri oluşturulmasına olanak tanımıştır.

Madde Karakteristik Eğrisi (MKE) ve Birey-Madde İlişkisi

MTK'nın önemli bir bileşeni olan madde karakteristik eğrisi (MKE), bir bireyin bir maddeye doğru cevap verme olasılığını bireyin yetenek düzeyi ile ilişkilendirir (Baker, 2001; Hambleton vd., 1991; Koğar, 2021). Örnek bir MKE Şekil 1.1'de gösterilmiştir.



Şekil 2.1. Madde Karakteristik Eğrisi

MKE, bu ilişkinin grafiksel bir temsili olarak, bir maddenin performansını görselleştirmeyi mümkün kılar. MKE'nin yatay eksenini bireyin yetenek düzeyini (θ), dikey eksenini ise doğru cevap verme olasılığını temsil eder (Hambleton vd., 1991). Eğrinin eğimi, maddenin ayırt edicilik gücünü, eğrinin konumu ise maddenin güçlük seviyesini yansıtır (Birnbaum, 1968). MKE, test geliştiricilere bir maddenin hangi yetenek düzeylerindeki bireyler için daha etkili olduğunu analiz etme imkânı sunar. Örneğin, eğrisi daha dik olan maddeler, yetenek düzeylerini daha keskin bir şekilde ayırt edebilirken, düz bir eğri düşük ayırt ediciliğe işaret eder. Ayrıca, eğrinin başlangıç noktasındaki sapmalar, şans faktörünün etkisini temsil eder. Bu, özellikle çoktan seçmeli testlerde düşük yetenek düzeyine sahip bireylerin doğru cevabı rastlantısal olarak seçme olasılığını göstermede önemlidir. MKE, yalnızca bireyin yetenek düzeyini belirlemekle kalmaz, aynı zamanda maddelerin testin genel geçerliliğine ve güvenilirliğine katkısını değerlendirir (Segall, 2004). Bu analiz, testlerin hangi maddelerle güçlendirilmesi veya hangi maddelerin çıkarılması gerektiğine dair önemli bilgiler sağlar.

Bilgi Fonksiyonu ve Test Doğruluğu

MTK'nın bilgi fonksiyonu, bir testin veya maddenin belirli bir yetenek düzeyinde ne kadar bilgi sağladığını ölçen kritik bir araçtır (Hambleton vd., 1991). Bu fonksiyon hem testlerin genel doğruluğunu hem de maddelerin bireysel ölçüm gücünü değerlendirme imkânı sunar (Samejima, 1969). Bilgi fonksiyonu, özellikle bireye uyarlanmış testlerde, testlerin verimliliğini artırmada önemli bir rol oynar. Bireyin yetenek düzeyine uygun maddeler seçilerek hem test süresi kısaltılır hem de ölçüm doğruluğu artırılır. Örneğin, yetenek düzeyi düşük olan bir bireye kolay maddeler, yetenek düzeyi yüksek olan bir bireye ise zor maddeler sunularak, bireylerin gerçek yetenek düzeylerine daha doğru bir şekilde ulaşılabilir. Bilgi fonksiyonunun analizi, bir testin hangi yetenek seviyelerinde daha hassas ölçüm yaptığını ortaya koyar ve bu bilgiler, testlerin tasarımı ve optimizasyonu için temel bir rehber niteliğindedir (Embretson & Reise, 2000; Hambleton vd., 1991).

Madde Tepki Kuramının Klasik Test Kuramına Göre Avantajları

MTK, KTK'ya kıyasla daha modern ve esnek bir yaklaşım sunar. KTK'nın temel varsayımlarından biri, bir testte elde edilen toplam puanın, bireyin yetenek düzeyini yansıttığıdır. Ancak bu yaklaşım, testin özelliklerinin bireylerin yetenek düzeyi üzerindeki etkisini dikkate almaz. Lord ve Novick (1968), KTK'nın toplam test puanlarına olan bağımlılığının, farklı zorluk seviyelerine sahip testler arasındaki karşılaştırmaların güvenilirliğini sınırladığını ifade etmişlerdir. MTK ise bireyin yetenek düzeyini ve test maddelerinin özelliklerini bağımsız olarak ele alarak bu sınırlamayı aşar. Örneğin, Hambleton ve Swaminathan (1985), MTK'nın her bir maddenin ayrı ayrı analiz edilmesine izin verdiğini ve bu sayede hem bireysel hem de grup düzeyindeki ölçümler için daha hassas sonuçlar sağladığını belirtmiştir. Ayrıca, MTK'nın bireylerin yetenek düzeyine uygun maddeler seçerek ölçüm doğruluğunu artırması, BOBUT için önemli bir avantaj sunar (Weiss & Kingsbury, 1984). Bu nedenle, MTK'nın esnek yapısı hem ölçüm süreçlerinin geçerliliğini artırır hem de farklı test uygulamaları arasında karşılaştırma yapılmasını mümkün kılar.

Madde Parametrelerinin Gruptan Bağımsızlığı ve Karşılaştırılabilirlik

MTK'nın en önemli avantajlarından biri madde parametrelerinin gruptan bağımsız olmasıdır. KTK'da, bir maddenin zorluk seviyesi veya ayırt ediciliği, o maddenin uygulandığı gruba bağlı olarak değişebilir (Embretson & Reise, 2000). Ancak MTK, madde parametrelerinin sabit kaldığını ve farklı gruplar arasında karşılaştırılabilir olduğunu varsayar. Baker (2001), bu özelliğin, uluslararası sınavlar ve kültürlerarası karşılaştırmalar gibi

durumlarda MTK'yı vazgeçilmez bir araç haline getirdiğini ifade etmiştir. Örneğin, PISA (Uluslararası Öğrenci Değerlendirme Programı) ve TIMSS (Uluslararası Matematik ve Fen Eğilimleri Araştırması) gibi uluslararası standartlaştırılmış testlerde, farklı ülkelerde uygulanan test maddelerinin aynı psikometrik özelliklere sahip olması, sonuçların güvenilirliğini ve geçerliliğini artırır. Aynı zamanda, madde parametrelerinin sabitliği, testlerin farklı zorluk seviyelerine sahip gruplar üzerinde uygulanmasını mümkün kılar ve adil bir ölçüm süreci sağlar. Bu özellik, özellikle eğitimde eşitlik ilkesinin sağlanmasında önemli bir rol oynar.

Geçerlik ve Güvenirlik Açısından MTK

MTK, testlerin geçerlik ve güvenirlik analizlerinde KTK'ya kıyasla daha ayrıntılı bilgiler sunar. MTK hata tahminlerini grup geneline yönelik bir değer olarak değil, bireysel düzeyde yapılabilir (Embretson & Reise, 2000). Güvenirlik, yetenek puanları (ya da yanıt örüntüleri) arasında farklılık gösterebilir. Hata ya da güvenirlik tahminleri yapılırken, bir maddenin yanıtlayıcının yetenek seviyesinde sağladığı bilgi esas alınır. Bu yaklaşım, test odaklı değil, madde odaklı bir güvenirlik anlayışını ifade eder. Eğer test, yanıtlayıcının yetenek seviyesinde daha fazla bilgi sağlıyorsa, o seviyede testin güvenirliliği de o kadar yüksek kabul edilir (Hambleton & Swaminathan, 2013). Lord (1980), MTK'nın güvenirlik analizlerinde bilgi fonksiyonunu kullanarak, bireylerin farklı yetenek düzeylerinde testin ne kadar güvenilir olduğunu belirlemeye olanak tanıdığını belirtmiştir. Geçerlik açısından, MTK, madde parametrelerinin analizine dayalı olarak testlerin hedeflenen özellikleri ölçüp ölçmediğini değerlendirir (DeMars, 2010). Ayrıca, MTK'nın bireyselleştirilmiş ölçüm sistemlerinde kullanılması, testlerin her bir birey için geçerli sonuçlar üretmesini sağlar.

2.3.2 Madde Tepki Kuramının Matematiksel Modellemesi

MTK, bireylerin bir teste verdikleri yanıtların, bireylerin yetenek düzeyi (θ) ve maddelerin psikometrik özellikleri (madde güçlüğü, ayırt edicilik, şans faktörü, yanıt değişkenliği) ile açıklanabileceğini varsayan güçlü bir kuramsal çerçevedir. Bu kuram, farklı düzeylerde detay sağlayan bir dizi model sunar: Rasch modeli, 1 Parametre Lojistik Modeli (1PL), 2 Parametre Lojistik Modeli (2PL), 3 Parametre Lojistik Modeli (3PL) ve 4 Parametre Lojistik Modeli (4PL). Her model, farklı ölçüm ihtiyaçlarına göre uygulanabilir. Aşağıda, bu modeller detaylı bir şekilde açıklanmıştır:

Rasch Modeli

Rasch modeli, MTK'nın en temel ve yaygın kullanılan modellerinden biridir. Bu model, sadece madde güçlüğü (b) parametresine odaklanır ve tüm maddelerin ayırt edicilik değerlerini sabit ($a=1$) kabul eder. Matematiksel olarak model şu şekilde ifade edilir:

$$P(\theta) = \frac{e^{(\theta-b)}}{1 + e^{(\theta-b)}}$$

Rasch modelinin temel varsayımlarından biri, madde parametrelerinin bireylerden bağımsız olmasıdır (De Ayala, 2018). Yani, bir maddenin güçlüğü, teste katılan grubun özelliklerinden etkilenmez. Rasch (1960), bu modelin ölçüm sürecine getirdiği basitlik ve şeffaflık nedeniyle, eğitim ve sağlık araştırmalarında standartlaştırılmış testlerin geliştirilmesinde yaygın bir şekilde kullanılmasını savunmuştur. Ayrıca, bu model, verinin modellenen yapıya ne kadar iyi uyduğunu değerlendirme imkânı sunarak testlerin geçerliliğini artırır (Wright & Stone, 1979). Ancak Rasch modeli, ayırt edicilik veya şans faktörlerini hesaba katmadığı için daha karmaşık testlerde yetersiz kalabilir.

1 Parametrelili Lojistik Modeli (1PL)

1PL modeli, Rasch modeline benzer şekilde yalnızca madde güçlüğü (b) parametresini içerir ancak ayırt edicilik parametresinin (a) sabit olması gerektiğini varsayar. Model şu şekilde ifade edilir:

$$P(\theta) = \frac{1}{1 + e^{-D(\theta-b)}}$$

Burada D ifadesi ölçekleme sabiti ifade ederken genellikle 1.70 olarak alınır. Diğer semboller ise Rasch modeli ile aynıdır. Hambleton & Swaminathan (1985), 1PL modelinin, madde güçlüğüne odaklanarak test geliştirme sürecini basitleştirdiğini ve özellikle homojen bir madde kümesine sahip standart testlerde etkili olduğunu belirtmiştir. Ancak, tüm maddelerin eşit ayırt ediciliğe sahip olduğu varsayımı, heterojen madde kümesi içeren testlerde modelin geçerliliğini sınırlayabilir.

Rasch modeli ile 1PL model arasındaki temel fark, iki modelin dayandığı felsefi ve matematiksel yaklaşımlar ile ölçme süreçlerine yönelik varsayımlarından kaynaklanır. Rasch modeli, ölçme işlemini standartlaştırılmış ve mutlak bir yapı olarak görür; bu nedenle, tüm

maddeler için ayırt edicilik (a) sabit olarak 1 kabul edilir ve ölçme süreci yalnızca madde güçlüğü (b) parametresine dayanır (De Ayala, 2018). Rasch modelinin temel hedefi, bireylerin yetenek düzeylerini ve maddelerin güçlük seviyelerini, teste katılan grubun özelliklerinden bağımsız bir şekilde karşılaştırılabilir hale getirmektir (Rasch, 1960). Bu model, veri uyumunu ön planda tutar ve verinin modele uygunluğu sağlanamazsa, model değil veri yenilenir (Embretson & Reise, 2000). Öte yandan, 1PL modeli, Rasch modeli ile aynı matematiksel formüle sahip olmasına rağmen, pragmatik bir yaklaşıma sahiptir. 1PL modelinde, ayırt edicilik parametresi (a) teorik olarak sabit kabul edilse de Rasch modelden farklı olarak 1,00 dışında değerler alabilir.

2 Parametrelili Lojistik Modeli (2PL)

2PL modeli, hem madde güçlüğü (b) hem de ayırt edicilik (a) parametrelerini dikkate alır. Maddeler arasında farklı ayırt edicilik seviyelerinin bulunabileceği varsayımıyla, bireylerin yetenek düzeyini daha hassas ölçmeyi amaçlar:

$$P(\theta) = \frac{1}{1 + e^{-D \cdot a(\theta - b)}}$$

Burada, (a) maddenin ayırt edicilik parametresini ifade ederken, diğer semboller 1PL modelindekiyle aynıdır. Ayırt edicilik parametresi (a), bir maddenin, farklı yetenek düzeylerini ne kadar iyi ayırabileceğini ifade eder (Baker, 2001). Ayırt ediciliği yüksek maddeler, bireylerin yetenek düzeylerini daha keskin bir şekilde sınıflandırabilir. Baker (2001), 2PL modelinin, test geliştirme süreçlerinde hem bireysel hem de grup ölçümlerinde daha doğru sonuçlar verdiğini vurgulamıştır.

2PL ile 1PL modeller arasındaki temel fark, 2PL modelinin hem madde güçlüğü hem de madde ayırt ediciliğini dikkate almasıdır. 1PL modeli, tüm maddelerin eşit ayırt ediciliğe sahip olduğunu varsayar ve bu nedenle maddeler arasındaki ayırt edicilik farklılıklarını göz ardı eder. Ancak gerçek dünyada, her madde bireylerin yetenek düzeylerini aynı ölçüde ayırt edemez; bazı maddeler daha belirgin bir şekilde bireyler arasındaki farkları yansıtabilir. Bu durumu ele almak için 2PL modeli, ayırt edicilik parametresini modele dahil eder ve her bir maddenin, bireylerin yetenek düzeylerini ayırt etme kapasitesini ölçer (Hambleton & Swaminathan, 1985). Ayırt ediciliği yüksek maddeler, yetenek düzeylerinde küçük farkları dahi belirgin hale getirirken, düşük ayırt edicilikteki maddeler bu farkları yeterince yansıtamaz. Bu

durum, 2PL modelini, heterojen madde kümesine sahip testlerde daha etkili kılar (Embretson & Reise, 2000).

3 Parametrelili Lojistik Modeli (3PL)

3PL modeli, madde güçlüğü (b) ve ayırt edicilik (a) parametrelerine ek olarak, şans faktörü (c) adı verilen bir parametreyi de içerir. Model, özellikle çoktan seçmeli testlerde, düşük yetenek düzeyine sahip bireylerin doğru cevabı rastlantısal olarak seçme olasılığını açıklamayı hedefler:

$$P(\theta) = c + (1 - c) \cdot \frac{1}{1 + e^{-D \cdot a(\theta - b)}}$$

Burada (c) şans faktörü yani rastlantısal olarak doğru cevabı seçme olasılığını ifade ederken, diğer semboller önceki modellerle aynıdır. 3PL model, 2PL modelinin üzerine, özellikle çoktan seçmeli testlerde karşılaşılan şans faktörünü (c) ekler. Bu model, düşük yetenek düzeyine sahip bireylerin doğru cevabı tamamen rastlantısal olarak seçme olasılığını hesaba katar. Lord (1980), 3PL modelinin, özellikle çoktan seçmeli testlerde adil bir ölçüm süreci sağlamak için daha uygun olduğunu belirtmiştir. 3PL modelinin eklediği şans faktörü, bireylerin yetenek düzeyleri ile maddelerin psikometrik özellikleri arasındaki ilişkinin daha doğru bir şekilde modellenmesine olanak tanır. Ancak bu model, daha karmaşık bir matematiksel yapıya sahiptir ve daha büyük örneklem boyutları ile güçlü hesaplama araçları gerektirir. Veerkamp & Berger (1997), 3PL modelinin, testlerin genel doğruluğunu artırdığını ancak modelin uygulama maliyetlerini ve teknik zorluklarını da artırdığını ifade etmişlerdir.

4 Parametrelili Lojistik Modeli (4PL)

4PL modeli, madde güçlüğü (b), ayırt edicilik (a) ve şans faktörü (c) parametrelerine ek olarak üst asimptot (d) adı verilen bir parametreyi de içerir. Bu model, doğru cevap olasılığının yalnızca şansa bağlı olarak sıfırdan farklı bir başlangıç seviyesine sahip olmasının yanı sıra, belirli bir maksimum seviyeyi de aşmamasını dikkate alır:

$$P(X = 1|\theta) = c + (d - c) \cdot \frac{1}{1 + e^{-D \cdot a(\theta - b)}}$$

Burada (d) üst asimptot parametresini yani, doğru cevap olasılığının ulaşabileceği maksimum seviyeyi temsil eder (örneğin, dikkatsizlik ya da diğer faktörler nedeniyle en yüksek

yetenek seviyesindekiler dahi her zaman doğru cevap veremeyebilir). Üst asimptot parametresi doğru cevap olasılığının ulaşabileceği maksimum seviyeyi gösterir ve bu, en yüksek yetenek düzeyine sahip bireylerin bile her zaman doğru cevap veremeyebileceğini dikkate alır. Bu modelin özellikle sınav sürecine dair gerçekçi bir bakış sunduğu ve yüksek düzeyde esneklik sağladığı belirtilmiştir (Embretson & Reise, 2000; Lord, 1980). 4PL modelinin en dikkat çekici özelliği, üst asimptot parametresiyle (d) doğru cevap olasılığının 1'den daha düşük bir maksimum değere ayarlanabilmesidir. Lord (1980), bu parametrenin, bireylerin dikkat eksikliği, yanlış anlama ya da diğer dışsal faktörler nedeniyle en yüksek yetenek seviyelerinde dahi hata yapabileceklerini gösterdiğini ifade etmiştir. Ayrıca, Embretson & Reise (2000) 4PL modelinin, bireysel farklılıkları daha hassas şekilde yansıtmaya yeteneğiyle hem düşük hem de yüksek yetenek seviyelerindeki bireyler için daha anlamlı sonuçlar sunduğunu belirtmişlerdir. Ancak, bu modelin matematiksel karmaşıklığı ve büyük örneklem gereksinimi, uygulama maliyetlerini artırabilir (Veerkamp & Berger, 1997).

2.3.3. Madde Tepki Kuramının Varsayımları

Tek Boyutluluk Varsayımı

MTK'nın temel varsayımlarından biri olan tek boyutluluk, bir testin yalnızca tek bir yeteneği veya özelliği ölçtüğünü ifade eder (Hambleton & Swaminathan, 1985; Embretson & Reise, 2000). Bu varsayım, test maddelerinin tümünün aynı temel yetenek veya özelliği değerlendirdiğini garanti eder. Örneğin, bir okuma anlama testinde, tüm maddelerin bireylerin okuma anlama becerilerini ölçtüğü varsayılır. Bu durumda, maddelerin ölçtüğü yeteneğin dışında kalan dışsal faktörlerin test sonuçlarını etkilemediği kabul edilir (Crocker & Algina, 1986; Hambleton vd., 1991). Reckase (1979), tek boyutluluk varsayımının, MTK modellerinin geçerliliği için kritik öneme sahip olduğunu vurgulamıştır. Bu varsayım ihlal edildiğinde, modelin tahmin ettiği birey yetenekleri (θ) yanıltıcı olabilir ve testin genel geçerliliği sorgulanabilir.

Tek boyutluluk varsayımının sağlanıp sağlanmadığını kontrol etmek için çeşitli istatistiksel yöntemler kullanılır. Faktör analizi tek boyutluluğun test edilmesinde en yaygın kullanılan yöntemlerden biridir. Bu analiz, test maddelerinin tek bir faktör etrafında toplandığını ve bu faktörün tüm maddeler için ortak bir varyans kaynağı oluşturduğunu kanıtlamaya yarar (Hambleton & Swaminathan, 1985). Analiz sonucu eğer baskın bir boyut söz konusuyla tek boyutluluk sağlanmış olur. Eğer analiz sonucunda birden fazla faktör ortaya

çıkarsa, bu durum tek boyutluluğun ihlal edildiğine işaret eder ve testin yeniden tasarlanması veya farklı modelleme yaklaşımlarının (örneğin, çok boyutlu MTK) kullanılması gerekebilir.

Yerel Bağımsızlık Varsayımı

Yerel bağımsızlık, MTK'nın temel varsayımlarından biri olup, bir testteki maddelere verilen yanıtların yalnızca bireyin yetenek düzeyi tarafından belirlenmesi gerektiğini ifade eder (Hambleton vd., 1991; Koğar, 2021). Bu varsayıma göre, bir maddeye verilen yanıt, diğer maddelere verilen yanıtlardan bağımsız olmalıdır (Lord & Novick, 1968). Eğer maddeler arasında sistematik bir ilişki varsa, yerel bağımsızlık ihlal edilmiş olur (Hambleton & Swaminathan, 1985).

Yerel bağımsızlığın sağlanıp sağlanmadığını belirlemek için Q3 indeksi gibi yöntemler kullanılır. Yen (1984) tarafından geliştirilen Q3 indeksi, maddeler arasındaki yanıt korelasyonlarını analiz eder ve yüksek korelasyonlar yerel bağımsızlık ihlaline işaret eder. Örneğin, bir hikâyeye dayalı birkaç soru içeren bir okuma anlama testinde, maddeler arasındaki bağımlılık yüksek olabilir ve bu durum testin geçerliliğini zayıflatır (Hambleton vd., 1991).

Yerel bağımsızlık ihlalleri, bireylerin yetenek tahminlerini yanıltıcı hale getirebilir ve ölçme doğruluğunu düşürebilir (Reckase, 2006). Maddeler arasındaki bağımlılık arttığında, testin sağladığı bilgi azalır ve yetenek tahmini üzerindeki belirsizlik artar. Tek boyutluluk ve yerel bağımsızlık varsayımları yakından ilişkilidir; tek boyutluluk sağlandığında yerel bağımsızlık da sağlanır, ancak yerel bağımsızlık sağlansa bile testin tek boyutlu olduğu kesinleşmez. Bu nedenle, iki varsayımın birlikte test edilmesi testlerin genel geçerliliğini artırmada kritik bir rol oynar (Hambleton & Swaminathan, 1985).

Veri-model uyumu

MTK'da veri-model uyumu, bir testten elde edilen verilerin seçilen modelin varsayımlarına uygun olup olmadığını değerlendirir. Bu uyum, modelin birey yeteneklerini (θ) ve madde parametrelerini (madde güçlüğü, ayırt edicilik vb.) doğru tahmin etme kapasitesini belirler (Hambleton & Swaminathan, 1985). Eğer model veriyle yeterince uyumlu değilse, yetenek tahminleri ve madde parametreleri yanıltıcı olabilir. Örneğin, 1PL modeli tüm maddelerin aynı ayırt edicilik gücüne sahip olduğunu varsayar, ancak bu varsayım ihlal edilirse modelin geçerliliği zayıflar. Bu nedenle, test geliştirme sürecinde veri-model uyumu dikkatle değerlendirilmelidir.

Veri-model uyumsuzluğu, test sonuçlarının geçerliliğini ve bireyler arası karşılaştırmaların doğruluğunu olumsuz etkileyebilir (Lord & Novick, 1968). Yerel bağımsızlık ihlali gibi durumlar, modelin yetenek düzeylerini yanlış tahmin etmesine neden olabilir. Ayrıca, testin ölçüm doğruluğunu da düşürebilir ve bilimsel geçerliliğini zayıflatabilir (Reckase, 2006).

Veri-model uyumunu değerlendirmek için ki kare uyum testleri, madde uyum indeksleri ve Akaike Bilgi Kriteri (AIC) ve Bayes Bilgi Kriteri (BIC) gibi genel uyum ölçütleri kullanılır (Swaminathan vd., 2006). Bu yöntemler, modelin tahmin ettiği olasılıklar ile gözlenen yanıtlar arasındaki farkları değerlendirerek, modelin test verilerine ne kadar iyi uyduğunu analiz eder (Hambleton & Swaminathan, 1985). Eğer uyum yetersizse, modelin yeniden gözden geçirilmesi veya daha uygun bir modelin seçilmesi gerekebilir.

Madde Parametrelerinin Değişmezliği

Madde parametrelerinin değişmezliği, MTK'nın temel varsayımlarından biri olup, modelin geçerliliği ve uygulanabilirliği açısından kritik bir öneme sahiptir. Bu varsayım, bir maddeye ait güçlük, ayırt edicilik ve şans parametrelerinin, ölçülen grup veya bireylerin yetenek seviyelerinden bağımsız olması gerektiğini ifade eder (Hambleton & Swaminathan, 1985). Başka bir deyişle, farklı yetenek seviyelerine sahip bireylere aynı maddeler uygulandığında, madde parametreleri değişmemelidir. Bu ilke, MTK'nın genelleştirilebilirliğini ve özellikle test eşitleme ve BOBUT uygulamalarında tutarlılığı sağlar (Lord, 1980; Embretson & Reise, 2000).

Madde parametrelerinin değişmezliğini incelemek için farklı istatistiksel teknikler kullanılır. Ortak madde (anchor item) kullanımı, farklı gruplar arasında madde parametrelerinin sabit kalıp kalmadığını test etmek için yaygın bir yöntemdir (Kolen & Brennan, 2014). Ayrıca, madde uyum indeksleri ve modellik uygunluk testleri, madde parametrelerinin örnekleme bağımlı olmadan tahmin edilip edilmediğini değerlendirir (Embretson & Reise, 2000). Çok gruplu madde tepki modelleri, farklı gruplar için tahmin edilen madde parametrelerinin tutarlılığını inceleyerek bu varsayımın test edilmesini sağlar (Kim & Yoon, 2011).

Özellikle bilgisayarlı bireyselleştirilmiş testler (BOBUT) gibi dinamik testlerde, madde parametrelerinin değişmezliği büyük önem taşır. Test performansı, bireyin yetenek seviyesine en uygun maddelerin seçilmesine dayanır ve bu seçim, madde parametrelerinin bireylerden bağımsız olarak belirlenmesi gerektiği varsayımına dayanır (van der Linden & Pashley, 2010). Eğer madde parametreleri değişkenlik gösterirse, test skorlarının geçerliliği ve güvenilirliği ciddi şekilde zarar görebilir (Baker, 2001). Bu nedenle, madde parametrelerinin değişmezliği,

MTK'nın güvenilir bir ölçme modeli olarak sürdürülebilmesi için temel bir gereklilik olarak kabul edilmektedir.

Yetenek Parametrelerinin Değişmezliği

Yetenek parametrelerinin değişmezliği, MTK'nın temel varsayımlarından biri olup bireylerin yetenek seviyelerinin testte yer alan maddelerden bağımsız olması gerektiğini ifade eder (Lord & Novick, 1968). Bu varsayım, test eşitlenebilirliği ve adillliği açısından kritik olup, bireylerin farklı testlerdeki performanslarının karşılaştırılabilir olmasını sağlar (Hambleton vd., 1991). Araştırmalar, bu varsayımın test maddelerinin içeriğinin homojenliği, test formlarının eşitlenebilirliği ve gruplar arası tutarlılıkla yakından ilişkili olduğunu göstermektedir (Bock, 1972; van der Linden & Glas, 2010). Ancak, bu varsayım ihlal edilirse, test sonuçlarının geçerliliği ve güvenilirliği ciddi şekilde etkilenebilir.

Yerel bağımsızlık, yetenek parametrelerinin değişmezliği ile doğrudan ilişkilidir. Eğer test maddeleri bireylerin yetenek seviyeleri kontrol edildikten sonra bağımsız değilse, değişmezlik varsayımı tehlikeye girer (Hambleton & Swaminathan, 1985; Yen, 1984). Örneğin, belirli bir alt gruba ait maddeler ortak bir özellik taşıyorsa, bu durum yetenek parametrelerinin değişmezliğini ihlal edebilir (Sireci vd., 1991). Böyle bir ihlal, test sonuçlarında yanlılığa yol açarak gruplar arası karşılaştırmaların geçerliliğini zayıflatabilir (Camilli & Shepard, 1994; Mislevy, 1991). Bu nedenle, test geliştirme süreçlerinde yerel bağımsızlık ve değişmezlik varsayımlarını doğrulamak için çeşitli yöntemler kullanılır (MacDonald, 2003).

Bu varsayım, bireyler arası farklılıkların adil bir şekilde değerlendirilmesi açısından da önemlidir. İhlal durumunda, belirli demografik gruplar için test avantajları veya dezavantajları oluşabilir, bu da test adaletini tehdit eder (Messick, 1989). Örneğin, belirli bir gruba avantaj sağlayan test maddeleri, yetenek parametrelerinin değişmezliğini ihlal ederek test sonuçlarının güvenilirliğini zayıflatabilir (Clauser & Mazor, 1998). Değişmezlik varsayımını test etmek için kullanılan bir yöntem, test maddelerinin farklı alt kümeleriyle yetenek tahminleri yaparak bu tahminler arasındaki korelasyonu incelemektir. Yüksek ve anlamlı korelasyonlar, yetenek parametrelerinin madde setlerinden bağımsız olduğunu ve MTK'nın değişmezlik varsayımını desteklediğini gösterir (Lord, 1980).

2.3.4. Madde Tepki Kuramının Avantaj Dezavantajları

Ölçüm Doğruluğu ve Esnekliği

MTK, bireylerin yetenek tahminlerini daha hassas bir şekilde yapabilme yeteneğiyle bilinir. Özellikle heterojen madde havuzlarına sahip testlerde, MTK modelleri bireylerin performansını daha doğru bir şekilde değerlendirebilir. Embretson ve Reise (2000), MTK'nın geleneksel KTK'dan daha üstün olduğunu, çünkü toplam test puanına bağımlı olmadığını ve her bir maddenin bireysel katkısını ayrı ayrı değerlendirebildiğini ifade etmişlerdir. Bu esneklik, testlerin bireyselleştirilmiş bir şekilde tasarlanmasına olanak tanır. Örneğin, BOBUT, MTK modellerinin esnek yapısından faydalanarak her bireyin yetenek seviyesine uygun sorular sunar. Bu durum, test süresini kısaltırken ölçüm doğruluğunu artırır (Ertuna, 2020).

Test Geliştirme Sürecindeki Avantajları

MTK, test geliştirme sürecinde madde analizi için güçlü bir çerçeve sunar. Madde güçlüğü, ayırt edicilik ve şans faktörü gibi parametreler, her bir maddenin testin genel yapısına katkısını değerlendirmek için ayrı ayrı incelenebilir. Lord (1980), MTK'nın madde analizini daha derinlemesine yapabilme yeteneği sayesinde, düşük ayırt ediciliğe sahip maddelerin testten çıkarılmasına olanak tanıdığını ve bu sayede testlerin genel geçerliliğini artırdığını belirtmiştir. Ayrıca, MTK, farklı parametrelere dayalı olarak madde havuzlarının dengeli bir şekilde oluşturulmasını sağlar. Örneğin, bir testin kolay, orta ve zor maddelerden oluşan dengeli bir yapıya sahip olması, MTK'nın sunduğu parametrik analizlerle mümkün hale gelir.

Bilgisayar Tabanlı Testlerdeki Rolü

MTK, özellikle bilgisayar tabanlı bireye uyarlanmış testlerin geliştirilmesinde ve uygulanmasında kritik bir rol oynar. MTK modelleri, bireylerin önceki yanıtlarına dayalı olarak dinamik bir madde seçim süreci sağlar. Weiss ve Kingsbury (1984), MTK'nın bilgisayar tabanlı testlerde test süresini önemli ölçüde kısalttığını ve bireylerin yetenek tahminlerini daha hassas bir şekilde gerçekleştirdiğini belirtmiştir. Ayrıca, MTK, bilgisayar tabanlı sistemlerde adaptif algoritmaların uygulanmasını mümkün kılarak test deneyimini bireyselleştirir. Bu, özellikle dil yeterliği, psikoloji ve eğitim alanlarında büyük bir avantajdır.

Uluslararası Sınavlardaki Avantajları

MTK, uluslararası standart testlerde karşılaştırılabilirlik sağlamak için yaygın olarak kullanılmaktadır. PISA ve TIMSS gibi uluslararası sınavlar, MTK modellerinin sağladığı bireysel bağımsızlık özelliklerinden faydalanarak, farklı kültürel ve dilsel gruplar arasında adil

karşılaştırmalar yapar. Hambleton (1994), MTK'nın uluslararası sınavlarda eşitlik sağlama yeteneğini, ölçme biliminin önemli bir kazanımı olarak değerlendirmiştir. Ayrıca, MTK, testin farklı dillerdeki versiyonlarının madde parametreleri açısından karşılaştırılabilir olmasını sağlayarak, çeviri kaynaklı önyargıları minimize eder.

Karmaşık Yapı

MTK'nın en büyük sınırlılıklarından biri, modellerin matematiksel yapısının KTK'ya kıyasla daha karmaşık olmasıdır. Özellikle 3PL model hem madde gücüğü hem ayırt edicilik hem de şans faktörünü içermesi nedeniyle daha fazla veri ve güçlü hesaplama araçları gerektirir. Reckase (2006), MTK modellerinin uygulanabilirliğinin, özellikle küçük örneklem boyutlarında sınırlı olabileceğini ifade etmiştir. Ayrıca, MTK'nın daha fazla parametre içerdği durumlarda, model tahminlerinin doğruluğunu artırmak için daha büyük örneklemelere ihtiyaç duyulur.

Örneklem Büyüklüğü ve Veri Gereksinimi

MTK, KTK'ya kıyasla daha fazla veri gerektirir ve küçük örneklem boyutlarında geçerliliği azalabilir. Özellikle 2PL ve 3PL gibi daha karmaşık modeller, yeterli sayıda bireyden veri toplanmadığı durumlarda doğru parametre tahminleri sağlayamaz. Hambleton ve Jones (1993), MTK'nın geçerliliği için önerilen minimum örneklem boyutunun 500 ila 1000 arasında olması gerektiğini belirtmiştir. Bu durum, özellikle küçük ölçekli araştırmalar veya sınırlı kaynaklara sahip test uygulamaları için bir dezavantajdır.

Model Uyumu ve Varsayımların Sağlanması

MTK'nın bir diğer sınırlılığı, varsayımlarının sağlanması gerekliliğidir. Tek boyutluluk ve yerel bağımsızlık varsayımları, MTK'nın matematiksel geçerliliği için temel teşkil eder. Eğer bu varsayımlar sağlanamazsa, modelin tahminleri yanıltıcı olabilir. Lord ve Novick (1968), MTK modellerinin varsayımları ihlal edildiğinde, test sonuçlarının geçerliliğinin ciddi şekilde zayıflayabileceğini belirtmişlerdir. Özellikle yerel bağımsızlık varsayımının ihlali, test maddeleri arasındaki bağımlılıkların varlığına işaret eder ve bu durum, testin geçerliliğini olumsuz etkileyebilir (Koğar, 2021).

2.4.Bilgisayar Ortamında Bireye Uyarlanmış Testler

Başarının ölçülmesinde daha çok bireye odaklanan ve daha üst düzey analizler ve programlamalar sayesinde kullanım alanı genişleyen BOBUT, bireylerin yeteneklerini, bilgi düzeylerini veya diğer ölçülebilir özelliklerini belirlemek için kullanılan testlerdir (Meijer & Nering, 1999; Wainer vd., 2000). Bu testler, bilgisayar teknolojisi ve yapay zekâ algoritmalarının kullanımıyla, bireyin test sırasında verdiği yanıtlara göre testin zorluk seviyesini ve soru seçimini kişiselleştirir (Wainer vd., 2000). Bireye uyarlanmış testler, her bireyin test sürecini kendi özelliklerine göre deneyimlemesini sağlar. Test başladığında, bir başlangıç seviyesi belirlenir ve bireyin yanıtlarına göre testin zorluk seviyesi ayarlanır. Eğer birey bir soruyu doğru yanıtlarsa, daha zor bir soruyla karşılaşır. Eğer yanlış yanıtlarsa, daha kolay bir soruyla karşılaşabilir. Bu adaptif süreç, bireyin gerçek yeteneklerini ve sınırlarını daha doğru bir şekilde belirlemeyi amaçlar (van der Linden & Glass, 2000).

BOBUT'un avantajlarından biri, kişiselleştirilmiş ölçüm sunmalarıdır (Chang, 2015; Weiss & Kingsbury, 1984). Her bireye farklı sorular sunulduğu için, test sonuçları bireyin gerçek performansını daha hassas bir şekilde yansıtabilir. Ayrıca, bu testler hızlı ve kullanışlı bir şekilde tamamlanabilir, anında geri bildirim sağlayabilir ve daha esnek bir test deneyimi sunabilirler (Meijer & Nering, 1999). Bu testler, daha doğru ve güvenilir ölçümler elde etmek, test süreçlerini optimize etmek ve bireylere daha etkili geri bildirimler sunmak için önemli bir araç haline gelmiştir (Weiss & Şahin, 2024).

2.4.1. Bilgisayar Ortamında Bireye Uyarlanmış Testlerin Tarihçesi

BOBUT'un tarihçesi, bilgisayar teknolojisinin gelişimiyle paralel olarak ilerlemiştir. Bireye uyarlanmış testlerin tarihçesindeki bazı önemli kilometre taşları şu şekildedir; (Eggen & Verschoor, 2006; van der Linden, 2017; Wainer, 2000):

1950'ler: Bilgisayarların ortaya çıkmasıyla birlikte, testlerin bilgisayar ortamında uygulanması fikri doğdu. İlk bilgisayar tabanlı testler, genellikle çoktan seçmeli formatlarda ve sınırlı ölçüm özellikleriyle kullanılmaktaydı.

1960'lar: Bu dönemde, bilgisayar teknolojisinin gelişimiyle birlikte bireye uyarlanmış testlerin potansiyeli daha fazla keşfedildi. Bu testler, öğrencilerin yeteneklerini belirlemek, eğitim ihtiyaçlarını değerlendirmek veya mesleki alanlarda uygunluklarını ölçmek amacıyla kullanıldı.

1970'ler: Bu dönemde, bireye uyarlanmış testlerin kullanımı yaygınlaşmaya başladı. Bilgisayar teknolojisinin gelişimi ve test algoritmalarının ilerlemesi, testlerin daha doğru ve hızlı bir şekilde bireye uyarlanmasını sağladı. Bu sayede, her bireyin yetenek ve öğrenme seviyesine göre kişiselleştirilmiş testler oluşturulabildi.

1980'ler: Bireye uyarlanmış testlerin kullanımı ve popülaritesi arttı. Özellikle eğitim ve psikoloji alanlarında, bilgisayar tabanlı testlerin avantajları daha fazla fark edildi. Bu dönemde, daha karmaşık test algoritmaları ve ölçme modelleri geliştirildi.

1990'lar: Bilgisayar teknolojisinin ilerlemesiyle birlikte, bireye uyarlanmış testlerin kullanımı daha da yaygınlaştı. Ölçüm modelleri ve algoritmalarındaki gelişmeler, testlerin daha hassas ve güvenilir olmasını sağladı. Ayrıca, grafiksel ve çoklu ortam özelliklerinin eklenmesiyle testlerin kullanıcı deneyimi iyileştirildi.

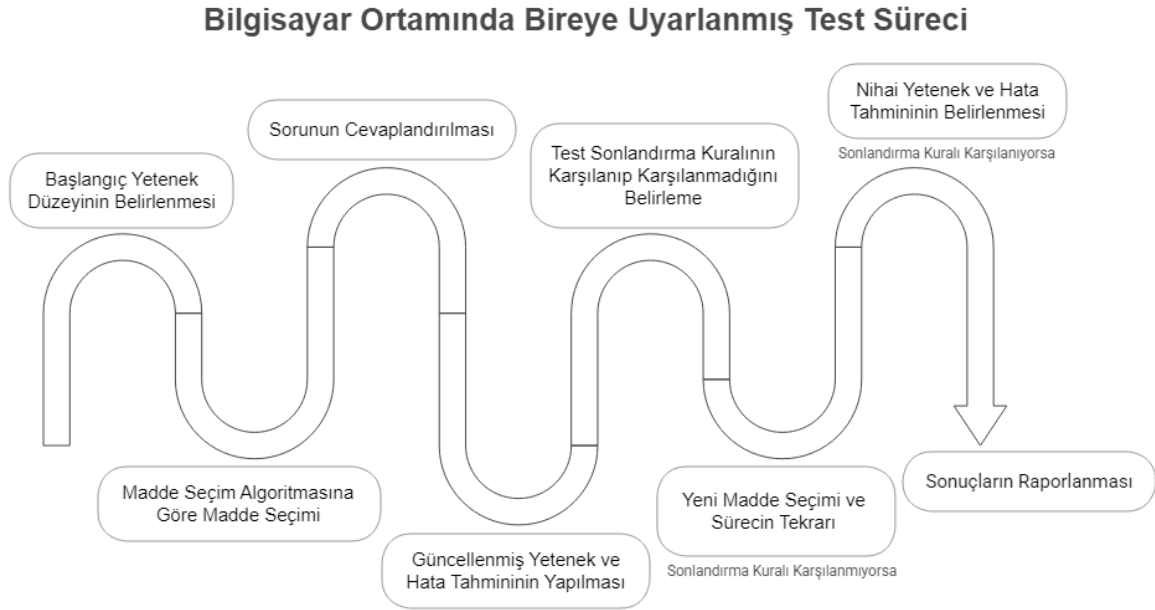
2000'ler: Bireye uyarlanmış testlerin kullanımı, internet ve çevrimiçi testleme imkanlarının artmasıyla daha da yaygınlaştı. Bu dönemde, çevrimiçi test platformları ve adaptif test yazılımları geliştirildi. Mobil teknolojinin yaygınlaşmasıyla birlikte, taşınabilir cihazlar üzerinden de bireye uyarlanmış testlerin kullanımı arttı.

Günümüz: Bireye uyarlanmış testler, eğitim, istihdam değerlendirmesi ve psikolojik değerlendirme gibi alanlarda yaygın bir şekilde kullanılmaktadır. Bilgisayar teknolojisi ve yapay zekâ algoritmalarının ilerlemesi, bireye uyarlanmış testlerin daha da sofistike hale gelmesini sağlamıştır. Bu testler, bireylerin öğrenme seviyelerini, yeteneklerini, ilgi alanlarını veya psikolojik durumlarını belirlemek için kullanılmaktadır. Günümüzde, bireye uyarlanmış testler, kişiselleştirilmiş geri bildirim ve öğrenme fırsatları sunarak, bireylere daha etkili bir şekilde yardımcı olmaktadır. Ayrıca, büyük veri analizi ve ölçme modellemesi gibi alanlardaki gelişmeler, bireye uyarlanmış testlerin daha doğru, güvenilir ve geçerli sonuçlar sağlamasını desteklemektedir.

2.4.2. Bilgisayar Ortamında Bireye Uyarlanmış Test Süreci

Bilgisayar ortamında bireye uyarlanmış test süreci çeşitli adımlar aracılığıyla gerçekleştirilir. Bu süreç, bireyin başlangıç yetenek düzeyinin belirlenmesiyle başlar ve madde seçim algoritmasına göre uygun bir test maddesinin sunulmasıyla devam eder. Bireyin yanıt doğrultusunda yetenek ve hata tahmini güncellenir ve testin sonlandırma kriterleri değerlendirilir. Eğer sonlandırma kuralları karşılanmazsa, yeni bir madde seçilir ve süreç

tekrarlanır. Nihai yetenek tahmini elde edildiğinde test tamamlanır ve sonuçlar raporlanır. Bu testlerin uygulanmasında izlenen adımlar ise Şekil 2.2’de ayrıntılı olarak sunulmuştur.



Şekil 2.2. Bireye Uyarlanmış Test Adımları

Bilgisayar ortamında bireye uyarlanmış test süreci, adayın başlangıç yetenek düzeyinin belirlenmesiyle başlar. Daha sonra, madde seçim algoritmasına göre uygun bir madde seçilir ve aday bu soruyu cevaplandırır. Cevaplandırma sonrası, adayın yetenek tahmini güncellenir ve hata tahmini yapılır. Ardından, test sonlandırma kuralının karşılanıp karşılanmadığı kontrol edilir. Eğer sonlandırma kriterleri sağlanmışsa, adayın nihai yetenek ve hata tahmini belirlenir ve süreç sonuçların raporlanmasıyla tamamlanır. Eğer sonlandırma kriterleri karşılanmamışsa, yeni bir madde seçilerek süreç aynı döngü içinde devam eder. Bu yöntem, her aday için en uygun soruların seçilmesini sağlayarak değerlendirme sürecinin verimliliğini artırır.

2.4.3. Başlangıç Yetenek Düzeyi

BOBUT uygulamalarında başlangıç yetenek düzeyinin belirlenmesi, testin geçerliliği, güvenilirliği ve ölçüm hassasiyeti açısından kritik bir rol oynar (Segall, 2005). Bu süreçte kullanılan yöntemler, bireyin geçmiş performansı, rastgele tahminleme, sabit başlangıç noktası seçimi, belirli bir yetenek aralığı kullanımı ve Bayesci yaklaşımlar gibi farklı stratejileri içerir (Thissen & Mislevy, 2000).

Geçmiş performansa dayalı başlangıç yetenek tahmini, bireyin önceki sınav sonuçları, akademik başarıları veya diğer ölçüm araçlarından elde edilen bilgilerle yapılır (Embretson & Reise, 2000; Kalender, 2011). Bu yöntem, bireyin gerçek yeteneğine daha yakın bir başlangıç noktası sunarak testin etkinliğini artırır. Ancak, güvenilir geçmiş verilere ihtiyaç duyulması en büyük sınırlılıklarından biridir (Eggen & Verschoor, 2006).

Rastgele tahminleme yöntemi, birey hakkında önceden bilgi bulunmadığında kullanılan bir yaklaşımdır (Kingsbury & Zara, 1989). Bu yöntemde başlangıç yetenek düzeyi için rastgele bir değer atanır. Ancak, başlangıçta bireye uygun olmayan maddeler sunulma riski nedeniyle, yetenek tahmininin test sürecinde hızla güncellenmesi gerekir (Lord, 1980).

Sabit başlangıç noktası seçimi, tüm test alıcılar için belirlenen ortak bir yetenek düzeyini başlangıç noktası olarak kabul eder. Genellikle $\theta=0$ seçilir, çünkü yetenek dağılımının ortalaması olarak kabul edilir (Segall, 2004; van der Linden, 2005). Ancak, bu yöntem bireysel farklılıkları göz ardı edebilir ve heterojen gruplarda ölçüm doğruluğunu sınırlayabilir.

Belirli bir yetenek aralığının kullanılması, başlangıç yetenek seviyesini $\theta=-0,5$ ile $\theta=0,5$ aralığında rastgele belirlemeyi içerir (Thompson & Weiss, 2019). Bu yöntem, sabit başlangıç noktasının sınırlamalarını aşmaya çalışırken, bireyler arasında farklı başlangıç noktaları sunarak ölçüm hassasiyetini artırır. Ancak, başlangıç maddelerinin bireye uygun olmama riski devam etmektedir (Eggen, 1999).

Bayesci yaklaşımlar, bireyin yetenek düzeyini başlangıçta belirli bir ön bilgiye dayalı olarak tahmin eder. Bu yöntem, birey hakkında doğrudan bilgi olmasa bile genel yetenek dağılımlarını kullanarak makul bir başlangıç noktası sağlar (Birnbbaum, 1968). Sürekli güncellenen yetenek tahminleri ile ölçüm doğruluğu artırılır, ancak doğru bir ön bilgiye ihtiyaç duyulması önemli bir kısıtlama olabilir.

Sonuç olarak, başlangıç yetenek düzeyinin belirlenmesi, testin doğruluğunu ve bireyselleştirilmiş ölçüm süreçlerinin etkinliğini doğrudan etkiler. Kullanılacak yöntemin seçimi, testin hedef kitlesi ve ölçüm amacı doğrultusunda dikkatle değerlendirilmelidir.

2.4.4. Madde Seçim Yöntemleri

BOBUT, her bir katılımcıya test sırasında sunulacak soruları, önceki yanıtlara göre seçerek, testin her birey için daha doğru ve kullanışlı olmasını sağlar (Weiss, 1985). Bu süreç, her bireyin beceri düzeyine uygun maddelerle, testin uzunluğunu kısaltarak zaman açısından da verimlilik sağlar. Ancak, bu adaptif testlerin doğru bir şekilde çalışabilmesi için etkili madde

seçim yöntemlerinin uygulanması gerekir. Madde seçim, genellikle madde özellikleri (zorluk, ayırt edicilik gibi) ile testin geçerliliğini ve güvenilirliğini sağlamaya yönelik olarak, katılımcının yetenek seviyesiyle uyumlu maddelerin seçilmesini içerir (Geerlings vd., 2011).

Madde seçim yöntemlerinin çeşitliliği, temel olarak testin amaçlarına ve kullanılan modelin gereksinimlerine dayanır. Birçok BOBUT uygulaması MTK'yı kullanarak, her bir maddenin zorluk düzeyini, ayırt edici gücünü ve diğer özelliklerini dikkate alır. Bu tür testlerde, katılımcının yetenek seviyesi belirlendikçe, testin kalan maddeleri bu seviyeye uyarlanarak sunulur. Bu yöntemlerin her biri, testin her bir aşamasında hangi maddenin seçileceğini belirlemek için farklı yaklaşımlar kullanır. Aşağıda, bu yöntemlerin açıklamaları yer almaktadır:

Maximum Fisher Information-Maksimum Fisher Bilgisi (MFI)

MFI, bilgisayar ortamında bireye uyarlanmış test uygulamalarında sık kullanılan bir yöntemdir. Bu yöntemin amacı, test sorularını katılımcının yetenek düzeyi hakkında en fazla bilgiyi sağlayacak şekilde seçmektir. Fisher bilgisi, bir sorunun katılımcının yetenek seviyesindeki belirsizliği ne kadar azalttığını ölçer (Lord, 1980). Her bir soru, verilen yanıtlarla birlikte katılımcının yetenek seviyesinin daha hassas tahmin edilmesine olanak tanır. MFI yöntemi, özellikle MTK tabanlı testlerde etkili bir yaklaşımdır çünkü maddelerin güçlük (b) ve ayırt edicilik (a) gibi psikometrik özellikleri dikkate alınarak maksimum bilgi sağlayan sorular seçilir (Weiss, 1985). Bu yöntem, testin doğruluğunu artırırken test süresini kısaltarak verimliliği de artırır (Choi & Swartz, 2009). Bununla birlikte, MFI'nin bazı sınırlamaları vardır. Testin ilk aşamalarında, katılımcının yetenek seviyesi tam olarak belirlenmediğinden, düşük bilgi sağlayan maddeler seçilebilir (Geerlings vd., 2011). Ayrıca, yüksek ayırt ediciliğe sahip maddelerin testin başında kullanılması, ilerleyen aşamalarda kalan maddelerin ölçme hassasiyetini azaltabilir (Glas & van der Linden, 2003). Bu durum, test sonlandırma aşamasında optimal madde seçimini zorlaştırabilir. Bir diğer sınırlama, MFI'nin yalnızca Fisher bilgisine odaklanması nedeniyle test içeriği dengesini gözetmemesidir. Belirli bir konuya ait maddeler daha fazla bilgi sağlıyorsa, test bu konuya aşırı odaklanabilir (van der Linden & Glas, 2003).

Bayesian Optimal- Bayesyen Optimumu (bOpt)

bOpt, Bayesci bir madde seçim yöntemi olup, katılımcının önceki yanıtları ve yetenek seviyesi doğrultusunda sonraki soruyu belirler (Hambleton & Swaminathan, 1985). Yetenek tahminleri sürekli güncellenerek daha hassas hale gelir ve güvenilir sonuçlar elde edilir (Sinharay vd., 2006). MTK modellerine dayanan bu yöntem, maddelerin güçlük ve ayırt edicilik

özelliklerini dikkate alarak en uygun soruyu seçer (De Ayala, 1995). Bayesci yöntemlerin avantajı, bireyin yetenek seviyesini doğru belirlemesidir. Ancak, yüksek işlem gücü gerektirmesi zaman açısından dezavantaj oluşturabilir (van der Linden & Pashley, 2010) ve başlangıçtaki öncül dağılımlara duyarlılığı nedeniyle hatalı tahminler yapılabilir (Thissen & Mislevy, 2000). Bu nedenle, uygun öncül dağılımlar ve yeterli madde havuzu sağlanmalıdır.

Threshold Optimal- Eşik Optimumu (thOpt)

thOpt, katılımcının yetenek seviyesini belirli bir eşik değeriyle ilişkilendirerek en uygun soruyu seçen bir madde seçim stratejisidir (Thissen & Mislevy, 2000). Yeterlilik testleri veya belirli bir başarı seviyesinin ölçülmesi gereken sınavlar için uygundur (Veldkamp & Verschoor, 2019). Testin temel amacı, katılımcının bu eşiği aşıp aşmadığını belirlemektir. Her yeni madde seçiminde, katılımcının yetenek seviyesi ile hedef eşik değeri arasındaki fark minimize edilir (Zhang & Li, 2016). Bu strateji, belirli bir başarı düzeyini ölçmede etkili olsa da yalnızca belirli bir seviyeye odaklanması nedeniyle geniş bir yetenek aralığını ölçmede sınırlamalar taşır. Özellikle düşük veya yüksek yetenek seviyesindeki bireyler için ölçme doğruluğu azalabilir (Veldkamp & Verschoor, 2019). Bu nedenle, eşik değerinin doğru belirlenmesi ve test sürecinin dikkatle yönetilmesi kritik öneme sahiptir.

Maximum Likelihood Weighted Information- Maksimum Olabilirlik Ağırlıklı Bilgi (MLWI)

MLWI, maksimum olasılık tahminleri ile ağırlıklı Fisher bilgisini birleştirerek bireyin yetenek seviyesini en doğru şekilde tahmin etmeyi amaçlayan bir madde seçim yöntemidir (Lord, 1980). Katılımcının yetenek seviyesi ve maddenin psikometrik özellikleri dikkate alınarak en fazla bilgi sağlayacak madde seçilir. MLWI, maddeleri ağırlıklandırarak değerlendirdiği için ölçme doğruluğunu ve testin kullanılabilirliğini artırır (Sarı & Raborn, 2018). Testin başlangıcında ve sonunda daha doğru sonuçlar elde edilmesini sağlarken, her yeni maddeyi önceki yanıtlara dayalı olarak belirler (Ueno & Sonmuang, 2010). Ancak, büyük ölçekli testlerde karmaşık hesaplamalar gerektirdiğinden işlem gücünü zorlayabilir ve test hızını etkileyebilir. Buna rağmen, bireyin yetenek seviyesini daha güvenilir ve doğru belirlediği için bireye uyarlanmış testlerde güçlü bir yöntem olarak öne çıkar (Sarı & Raborn, 2018).

Maximum Posterior Weighted Information- Maksimum Artık Ağırlıklı Bilgi (MPWI)

MPWI, maksimum posterior ağırlıklı Fisher bilgisini kullanarak madde seçimi yapan ve Bayesci bir yaklaşıma dayanan bir yöntemdir (Hambleton & Swaminathan, 1985). Katılımcının önceki yanıtlarına göre yetenek tahminlerini günceller ve en fazla bilgi sağlayacak maddeyi seçer (Choi & Swartz, 2009). Bayesci model sayesinde, bireysel özelliklere uygun maddeler

belirlenerek ölçme doğruluğu artırılır. Özellikle testin erken aşamalarında, öncül dağılımlar yetenek tahminlerinin güvenilirliğini artırır (Sarı & Raborn, 2018). Heterojen veri setlerinde etkili sonuçlar sunan MPWI, kişiselleştirilmiş bir test deneyimi sağlar. Ancak, yüksek işlem gücü gerektirmesi hesaplama süresini uzatabilir ve test hızını etkileyebilir. Ayrıca, yanlış öncül dağılım seçimi testin ilk aşamalarındaki tahminlerin doğruluğunu olumsuz etkileyebilir (Kaplan vd., 2015). Sonuç olarak, MPWI bireye uyarlanmış testlerde güçlü bir yöntem sunarken, hesaplama zorlukları ve öncül dağılım seçimi dikkatle ele alınmalıdır.

Maximum Expected Information- Maksimum Beklenen Bilgi (MEI)

Bu yöntemde, madde havuzundaki her bir maddeye bireyin vereceği yanıtların olasılık dağılımı tahmin edilir ve bu olasılık dağılımında birey hakkında en fazla bilgiyi sağlaması beklenen madde seçilir (Aybek & Demirtaşlı, 2017). Bu yöntem, test sürecinde en fazla bilgi sağlayan maddeleri belirleyerek, ölçme doğruluğunu artırmayı amaçlar (Weiss, 1985). MEI'nin temel mantığı, her yeni madde seçildiğinde katılımcının yetenek seviyesindeki belirsizliği minimuma indirerek, en iyi tahminin yapılmasını sağlamaktır. MEI yöntemi, Fisher bilgisi temelinde en fazla ölçme gücüne sahip maddeleri seçerken, testin genel verimliliğini artırmaya da odaklanır (Ueno & Sonmuang, 2010). Özellikle testin başında, katılımcının yetenek seviyesi hakkında kesin bir bilgiye sahip olunmadığında etkili bir strateji olarak öne çıkar, çünkü her yeni madde katılımcının yetenek seviyesinin daha hassas tahmin edilmesine olanak tanır. Bu nedenle, MEI yöntemi, test sürecinin başlangıç aşamalarında yetenek tahmininin hızla iyileştirilmesini sağlayarak adaptif testlerin daha verimli çalışmasına yardımcı olur (Penfield, 2006). MEI yönteminin bazı sınırlamaları da bulunmaktadır. Özellikle daha kısa testlerde veya zaman sınırlamalarının olduğu durumlarda, yöntemin uygulanması daha karmaşık hale gelebilir. Bunun nedeni, MEI'nin her madde seçiminde, beklenen bilgi miktarını hesaplaması gerektiğinden, yüksek işlem gücü gerektirmesidir. Ayrıca, bu yöntemin test içeriği dengesini gözetmemesi, belirli konu alanlarında aşırı madde seçimine neden olabilir (Sarı & Raborn, 2018).

Maximum Expected Posterior Variance- Maksimum Beklenen Artık Varyans (MEPV)

MEPV, katılımcının yetenek seviyesindeki belirsizliği en aza indirerek daha hassas bir ölçüm sağlamayı amaçlayan bir madde seçim yöntemidir (Geerlings vd., 2011). Her soru, yetenek seviyesindeki varyansı en fazla azaltacak şekilde belirlenir ve böylece tahmin doğruluğu adım adım artırılır. Özellikle testin başında düşük güvenilirlik seviyelerinin olduğu durumlarda avantajlıdır, çünkü başlangıçtaki belirsiz tahminler hızla iyileştirilebilir (Penfield, 2006). Her yeni madde, önceki yanıtlar doğrultusunda belirsizliği minimize edecek şekilde

seçildiğinden, testin güvenilirliğini artırır (Veldkamp & Verschoor, 2019). Ancak, yalnızca varyans azaltımına odaklanması nedeniyle içerik dengesi gibi diğer psikometrik özellikleri göz ardı edebilir. Bu da testin sonunda belirli konuların aşırı ağırlıklandırılmasına yol açabilir (Ueno & Sonmuang, 2010). Ayrıca, sürekli varyans hesaplamaları yapması nedeniyle hesaplama maliyetleri yükselebilir.

Progressive- Aşamalı

Progressive yöntemi, testin başında daha kolay maddeler seçilerek katılımcının teste adaptasyonunu sağlamayı ve yetenek seviyesi arttıkça daha zor maddelere geçiş yapmayı hedefleyen bir stratejidir (Butterfield, 2016). Bu yöntem, özellikle katılımcının öğrenme veya performans düzeyindeki değişimi zaman içinde izlemek için tasarlanan testlerde kullanılır (Geerlings vd., 2011). Progressive yöntemi, katılımcının başlangıçta kolay sorularla karşılaşmasını sağlayarak, testin daha az stresli bir ortamda başlamasına ve bireyin teste alışmasına yardımcı olur. Bu yöntemin en önemli avantajlarından biri, katılımcının test boyunca gelişimini izleme olanağı sağlamasıdır. Testin sonunda, bireyin performansına ilişkin daha doğru ve güvenilir sonuçlar elde edilmesine katkıda bulunur. Özellikle öğrenci odaklı testlerde, bireyin zamanla gelişen bilgi birikimini ölçmek için etkili bir yaklaşımdır. Progressive yönteminin bu özelliği, eğitim ortamlarında ve bireysel öğrenme süreçlerini izleyen uygulamalarda büyük bir fayda sağlar. Bununla birlikte, Progressive yöntemi bazı sınırlamalara sahiptir. Testin başında daha kolay maddelerin seçilmesi, varyansı düşürebilir ve bireyin yetenek seviyesinin hassas bir şekilde tahmin edilmesini geciktirebilir. Ayrıca, başlangıç aşamalarında seçilen maddeler, daha üst düzey yetenek seviyelerine sahip bireyler için yeterince ayırt edici olmayabilir. Bu da testin doğruluğunu etkileyebilir (Barrada vd., 2008).

Proportional-Orantılı

Proportional yöntemi, katılımcının yetenek seviyesine orantılı olarak madde seçimini sağlayan bir stratejidir (Choi & Swartz, 2009). Her yeni madde, önceki yanıtlar doğrultusunda belirli bir orana göre seçilerek testin dinamik şekilde uyarlanmasını mümkün kılar (Lord, 1980). Bireyin yetenek seviyesine uygun sorular sunarak ölçme doğruluğunu artırır ve büyük veri setlerinde bireyselleştirilmiş test deneyimi sunmada etkilidir (Veldkamp & Verschoor, 2019). Bu sayede test doğruluğu ve geçerliliği artarken, bireysel farklılıklar hassas şekilde ölçülebilir. Ancak, hesaplama sürecinin karmaşıklığı büyük ölçekli testlerde işlem süresini uzatabilir ve yüksek işlem gücü gerektirebilir. Ayrıca, madde havuzunun doğru kalibre edilmesi ve testin içerik dengesinin korunması gereklidir (Ueno & Sonmuang, 2010).

Kullback-Leibler Divergence (KL)

KL, bilgi teorisi bağlamında kullanılan bir yöntem olup, bireye uyarlanmış testlerde katılımcının mevcut yetenek seviyesini daha doğru bir şekilde tahmin etmek için iki olasılık dağılımı arasındaki farkları minimize etmeyi amaçlar (Belov & Armstrong, 2010). Bu yöntem, bilgi kaybını en aza indirerek, daha doğru ve güvenilir sonuçların elde edilmesine olanak tanır. Özellikle yüksek doğruluk gerektiren testlerde, katılımcının yetenek düzeyine uygun maddelerin seçilmesini sağlayarak ölçme doğruluğunu artırır. KL, bireye uyarlanmış testlerde iki ana amaç doğrultusunda kullanılır: katılımcının yetenek düzeyine en uygun maddeleri seçmek ve test güvenliğini sağlamak için madde maruziyetini kontrol etmek (Belov & Armstrong, 2010). Yöntem, her bir maddenin sağladığı bilginin, katılımcının mevcut yetenek tahminine göre ağırlıklandırılmasını mümkün kılar. Bununla birlikte, KL, hesaplama açısından karmaşık bir yapıya sahip olduğundan, büyük ölçekli testlerde uygulama zorluklarına yol açabilir. Bu nedenle, madde havuzunun dikkatli bir şekilde kalibre edilmesi ve algoritma gerekliliklerinin test tasarımında dikkate alınması gerekmektedir. Bu özellikleriyle KL, bilgi kaybını minimize ederek ve doğruluğu artırarak, bireye uyarlanmış testlerde etkili bir yöntem olarak öne çıkar (Kaplan vd., 2015).

2.4.5. Yetenek Kestirim Yöntemleri

Yetenek kestirimi, bireye uyarlanmış testlerde katılımcıların yetenek düzeylerinin doğru bir şekilde tahmin edilmesinde kullanılan temel süreçtir. Bu süreç, her yeni soruyla birlikte katılımcının performansını değerlendirir ve böylece doğru sonuçlar elde etmek için testin her aşamasında yetenek seviyesini günceller. Bu yöntem, özellikle MTK modeline dayalı testlerde yaygın olarak kullanılmaktadır. Yetenek kestirimi, genellikle testin başında belirsiz olan yetenek seviyelerini doğru bir şekilde tahmin etmeye yönelik bir yaklaşım sağlar. Bu süreç, doğru bir şekilde yapılmazsa, testin güvenilirliği ve geçerliliği de düşebilir. Bu nedenle, doğru kestirim yöntemlerinin seçilmesi ve uygulanması, BOBUT testlerinin etkinliğini büyük ölçüde artırır (Lord, 1980; Embretson & Reise, 2000).

Yetenek kestirimi, genellikle üç ana yöntemle yapılır: *Maximum Likelihood (ML)*, *Expected A Posteriori (EAP)* ve *Maximum A Posteriori (MAP)*. Her bir yöntem, katılımcının yetenek seviyesini tahmin etmek için farklı istatistiksel yaklaşımlar kullanır. Bu yöntemler, MTK modeline dayalı testlerde madde seçim stratejilerinin etkinliğini artırmaya yardımcı olur. ML, her bir test cevabına dayalı olarak en olası yetenek seviyesini tahmin ederken, EAP ve MAP yöntemleri, Bayesci yaklaşımı kullanarak daha doğru sonuçlar elde etmeyi amaçlar

(Embretson & Reise, 2000; Segall, 2005). Bu yöntemlerin her biri, testin genel doğruluğunu artırırken, testin zaman verimliliği ve doğru sonuç üretme kapasitesini de artırır (Hambleton & Swaminathan, 1985; De Ayala, 2018).

Yetenek kestirimi sürecinde kullanılan bu üç yöntem, özellikle testin doğruluğu ve geçerliliği açısından kritik öneme sahiptir. Bununla birlikte, her bir yöntemin avantajları ve sınırlamaları da bulunmaktadır. Bu nedenle, bu yöntemlerin doğru bir şekilde seçilmesi, her bir testin amacına uygun şekilde optimize edilmesi gereklidir. BOBUT uygulamalarının etkinliğini artırmak için bu yöntemlerin doğru bir biçimde entegre edilmesi önemlidir (van der Linden, 2017; Weiss, 1985). Madde yetenek kestirin yöntemlerine ait matematiksel modelleme ve genel özellikler aşağıda verilmiştir.

Maximum Likelihood- Maksimum Olabilirlik (ML) Yöntemi

ML, yetenek kestirimi yöntemlerinin en temel ve yaygın kullanılan yöntemlerinden biridir. ML, testin her aşamasında, katılımcının verdiği yanıtlarla birlikte, en olası yetenek seviyesini tahmin etmeye çalışır (Bandolos, 2014). Bu yöntem, her bir test cevabının, katılımcının yetenek seviyesinin bir fonksiyonu olduğunu varsayar ve verilen yanıtların en olası olduğu yetenek seviyesini belirler. Matematiksel olarak, ML yöntemi şu şekilde ifade edilebilir:

$$\hat{\theta}_{ML} = \arg \max_{\theta} L(\theta | \mathbf{x})$$

Burada, θ_{ML} katılımcının tahmin edilen yetenek seviyesi (theta), $L(\theta|x)$ ise gözlemlenen veriler (x) altında, θ olasılık fonksiyonunu temsil eder. Bu fonksiyon, verilen test cevaplarının gözlemlenen veriler ile en uyumlu olan parametreyi (yetenek seviyesi) seçer. Yani, ML yöntemi, gözlemlenen yanıtların olasılığını maksimize etmeyi amaçlar.

ML yönteminin avantajı, hesaplamalarının nispeten basit ve hızlı olmasıdır (Brummel & Drasgow, 2010). Ayrıca, veriler yeterince büyükse ve doğru bir modelleme yapılmışsa, ML yöntemi doğru sonuçlar verebilir (Gorin vd., 2005). Ancak, bu yöntem bazı durumlarda sınırlamalar gösterebilir. Örneğin, testin başlangıcında katılımcının yetenek seviyesi belirsiz olduğu için, testin başındaki maddeler, başlangıçta daha fazla belirsizlik yaratabilir ve kestirim daha az güvenilir olabilir (Gorin vd., 2005). Bu nedenle, ML yöntemi genellikle daha fazla veri ile güçlendirilmiş testlerde daha doğru sonuçlar üretir. Bu metodun, daha büyük ve homojen veri setlerinde en iyi performansı sergilediği görülmektedir (Lord, 1980).

Expected A Posteriori (EAP)- Beklenen Artık Değer Yöntemi

EAP yöntemi, Bayesci bir yaklaşımı kullanarak yetenek kestirimi yapar ve bir katılımcının yetenek seviyesi hakkında en iyi tahmini, testin önceki sonuçlarına ve prior (ön) dağılıma dayalı olarak yapar (Chapman, 2022). EAP, her bir katılımcı için, gözlemlenen yanıtların olasılıklarını dikkate alarak, posterior dağılımını (testin olasılıklarıyla güncellenmiş dağılım) hesaplar ve bunun beklenen değerini kullanarak yetenek seviyesini tahmin eder. Matematiksel olarak, EAP şu şekilde ifade edilebilir:

$$\hat{\theta}_{EAP} = \int \theta f(\theta | \mathbf{x}) d\theta$$

Burada, θ_{EAP} , katılımcının tahmin edilen yetenek seviyesi, $f(\theta|x)$ ise, gözlemler ve prior bilgilere dayalı olarak güncellenmiş posterior dağılımını temsil eder. Bu yaklaşımda, testin başında katılımcının yetenek seviyesi hakkında bir prior bilgi (veya dağılım) kabul edilir ve bu bilgi, her yeni soruyla birlikte güncellenir. EAP yöntemi, daha fazla veri toplandıkça daha doğru sonuçlar verir çünkü katılımcının prior dağılımı ve gözlemler doğrultusunda daha güvenilir bir tahmin yapılır.

EAP'nin avantajı, özellikle az veri ile yapılan testlerde daha doğru sonuçlar elde edilmesini sağlamasıdır (Chen, 1996). Bayesci bir yaklaşım kullanarak, daha önceki test cevaplarından elde edilen bilgilerle katılımcının yetenek seviyesi hakkında güvenilir tahminler yapılabilir (De Ayala, 2018)

Ayrıca, EAP yöntemi, katılımcıların yetenek seviyelerini modellemek için daha esnek bir yol sunar. Ancak, EAP'nin hesaplamaları, özellikle büyük veri setlerinde ve karmaşık modellerde zaman alıcı olabilir. Bu nedenle, EAP yöntemi genellikle daha küçük testlerde ve daha az veri ile daha etkili sonuçlar verir (Hambleton & Swaminathan, 1985).

Maximum A Posteriori (MAP)- Maksimum Artık Değer Yöntemi

Maximum A Posteriori (MAP), EAP'nin bir varyasyonu olup, Bayesci yaklaşımın daha güçlü bir formudur. MAP, EAP gibi gözlemlenen verileri ve prior dağılımı kullanarak yetenek kestirimi yapar, ancak MAP'in farkı, katılımcının yetenek seviyesini tahmin ederken en yüksek posterior olasılığı maksimize etmeye çalışmasıdır (Wang, Hanson & Lau, 1999). MAP, daha güvenilir bir tahmin elde etmek amacıyla, posterior dağılımdan en olasılıklı değeri seçer. Matematiksel olarak, MAP şu şekilde ifade edilebilir:

$$\hat{\theta}_{MAP} = \arg \max_{\theta} [p(\mathbf{x} | \theta)p(\theta)]$$

Burada, $p(\mathbf{x}|\theta)$ gözlemlenen verilerin, belirli bir yetenek seviyesi (θ) dayanarak elde edilme olasılığıdır ve $p(\theta)$ prior dağılımdır. MAP, posterior dağılımdan, yani verilerle güncellenmiş prior dağılımdan en olasılıklı yetenek seviyesini tahmin eder. Bu yöntem, EAP'ye göre daha belirgin ve keskin bir tahmin sağlar çünkü maksimum olasılığı tercih eder.

MAP'in avantajı, daha güçlü bir prior bilgisine sahip olduğunda ve daha az belirsizlik olduğunda, testin sonunda daha doğru sonuçlar elde edilmesidir (Tsao vd., 2010). MAP yöntemi, özellikle daha fazla prior bilgiye dayalı kararların alındığı ve daha homojen veri setleri ile çalışıldığı durumlarda etkili bir şekilde kullanılır (Alotaibi & Selouani, 2010). Ancak, MAP'in zorlukları, prior dağılımın doğru seçilmesinin zor olabilmesi ve bazı durumlarda daha karmaşık hesaplamalar gerektirmesidir. Bu nedenle, MAP genellikle daha güçlü prior bilgilerine sahip testlerde ve daha büyük veri setlerinde kullanılmaktadır (De Ayala, 2018).

2.4.6. Testi Sonlandırma Kuralları

Testi sonlandırma kuralının belirlenmesi bireye uyarlanmış testlerde önemli bir aşamadır. Bu kurallar, testin sonlanacağı noktayı belirleyerek, gereksiz soruları engeller ve katılımcının yetenek seviyesini daha kullanışlı bir şekilde ölçmeye odaklanır. Testi sonlandırmak için kullanılan kurallar, testin başında ve sonunda verilerin ne kadar güvenilir olduğunu değerlendirmek için tasarlanır (Segall, 2005). Bu süreç, testin verimliliğini artırırken, aynı zamanda katılımcının performansını doğru bir şekilde ölçmek için minimum sayıda sorunun kullanılmasını sağlar. Testin sonlandırılması hem zaman yönetimi hem de kaynakların verimli kullanılması açısından kritik bir rol oynar. Bu kurallar, farklı test uygulama durumlarına ve testin amacına göre değişiklik gösterebilir. Testin sonlandırılmasında kullanılan kriterler arasında belirli bir güven seviyesine ulaşma, belirli bir soru sayısına ulaşma ya da zaman sınırlaması gibi farklı yaklaşımlar yer alır (Weiss, 1985).

Yaygın olarak kullanılan sonlandırma kuralından bir tanesi **standart hata** kullanılarak yapılan sonlandırmadır. Standart hata, testin katılımcının yetenek seviyesini ne kadar doğru tahmin ettiğini ölçen bir parametredir (Segall, 2005). Yetenek kestirimi süreci devam ettikçe, sistem katılımcının yetenek düzeyini daha hassas bir şekilde tahmin eder. Testin sonlandırılmasına karar verilmeden önce, katılımcının yetenek seviyesindeki belirsizlik, standart hata ile ölçülür. Eğer sistem, belirli bir standart hata seviyesine ulaşmışsa, yani

katılımcının yetenek seviyesi yeterince doğru tahmin edilmişse, test sonlandırılır. Matematiksel olarak, standart hata, testin her aşamasında Fisher bilgisi kullanılarak hesaplanabilir. Fisher bilgisi, her bir maddeyi yanıtlama sırasında katılımcının yetenek düzeyindeki belirsizliği ne kadar azalttığını gösterir. Standart hata, bu bilgiyi kullanarak, tahminin doğruluğunun ne kadar olduğunu belirler. Yani, standart hata ne kadar küçükse, kestirim o kadar güvenilirirdir (Lord, 1980). Standart hata temelli sonlandırma, testin kullanışlı ve doğru bir şekilde sonlanmasını sağlar çünkü gereksiz soruların sorulmasının önüne geçer (Weiss & Kingsbury, 1984).

Standart hata temelli sonlandırma kuralı, bireye uyarlanmış testlerin etkinliğini artıran kritik bir faktördür. Bu yaklaşımda, testin sonlandırılması, katılımcının yetenek seviyesinin ne kadar doğru bir şekilde kestirildiğine bağlıdır. Standart hata, testin her aşamasında, katılımcının yetenek seviyesindeki belirsizliği ölçen bir parametredir (Embretson & Reise, 2000). Yani, testin sonunda, yetenek kestiriminin ne kadar güvenilir olduğunu anlamak için standart hata kullanılır. Standart hata ne kadar düşükse, kestirim o kadar doğru kabul edilir. Bu yüzden testin sonlandırılması için belirli bir standart hata seviyesi kullanılır, bu seviyeler genellikle 0,30 veya 0,50 gibi değerler olabilir. Bu tür bir sonlandırma, katılımcının gerçek yetenek seviyesinin doğru bir şekilde yansıtılmasını sağlamak amacıyla, gereksiz soruları ortadan kaldırarak testin verimliliğini artırır (Weiss, 1984).

0,30'luk bir standart hata, genellikle oldukça yüksek doğruluk seviyelerini ifade eder. Bu durumda, katılımcının yetenek seviyesi, yeterince doğru bir şekilde kestirilmektedir ve daha fazla soru sormaya gerek yoktur. 0,30'luk bir standart hata seviyesi, testin sonunda, katılımcının performansına dayalı olarak yapılan tahminlerin %95 güven aralığında doğru olduğu anlamına gelir. Yani, bu noktada yapılan kestirim, katılımcının yetenek seviyesi hakkında oldukça güvenilir bir tahmin sunar. Bu seviyeye ulaşıldığında, testin sonlandırılması, gereksiz verilerin toplanmasının önüne geçer ve testin verimliliğini artırır. Özellikle, daha kısa testlerde, 0,30 gibi düşük bir standart hata seviyesi, katılımcının yetenek düzeyini doğrulukla ölçmek için yeterlidir. Ancak, bu tür bir kuralın uygulanabilmesi için, başlangıçta kullanılan madde havuzunun ve kestirim stratejilerinin doğru bir şekilde tasarlanmış olması gerekir. 0,30'luk bir standart hata, genellikle daha kısa testlerde ve daha net cevaplar veren katılımcılar için uygun olabilir (Lord, 1980).

0,50'lik bir standart hata, yine kullanılan ölçütlerden bir tanesidir. Bu durumda, katılımcının yetenek seviyesi hala tahmin edilebilir, ancak tahminin doğruluğu 0,30 seviyesine göre daha düşük olacaktır. 0,50'lik bir standart hata seviyesi, genellikle katılımcının performansının henüz tam olarak netleşmediği, daha fazla bilgi edinilmesi gereken bir durumu

yansıtır. Bu durumda, daha fazla soruya ihtiyaç duyulabilir. Yani, 0,50'lik bir standart hata, testin henüz yeterince kesin sonuçlar vermediğini ve daha fazla veri toplamayı gerektirdiğini gösterir. Bu tür bir sonlandırma kuralı, genellikle başlangıç seviyesindeki katılımcılar veya testin başında belirli bir başarı düzeyine ulaşamayan bireyler için daha uygun olabilir. Bu durumda, katılımcının yetenek seviyesini daha hassas bir şekilde kestirmek için daha fazla madde sorulması gerekebilir. 0,50'lik standart hata, daha karmaşık testler ve katılımcıların yetenek düzeylerinin net olmadığı durumlar için yaygın olarak tercih edilir (Hambleton & Swaminathan, 1985). Her iki sonlandırma türü de testin katılımcıların yetenek seviyelerini doğru bir şekilde ölçmesini sağlar, ancak hangi seviyenin kullanılacağı, testin amacına ve katılımcıların performansına göre değişir. Bu sonlandırma türlerinin doğru şekilde seçilmesi, testin verimliliğini artırırken, aynı zamanda gereksiz sorulardan kaçınılmasını sağlar (Weiss, 1984).

Sabit madde sonlandırma kuralı ise, belirli bir soru sayısına ulaşılmasıyla testin tamamlanmasını hedefler. Bu yöntem, genellikle testin süresini sınırlı tutmak veya testin belirli bir sayıda sorudan sonra sonlanmasını sağlamak amacıyla kullanılır. Sabit madde sonlandırma, testin başlangıcında, belirli bir madde sayısına göre sonlandırma yapılmasını önerir. Bu strateji, özellikle testin süresini veya maddelerin sayısını sınırlamak isteyen durumlar için uygundur. Sabit madde sonlandırma, özellikle her katılımcıya aynı sayıda soru sunulması gerektiğinde veya her testin aynı uzunlukta olması gerektiğinde kullanılabilir. Ancak, sabit madde sonlandırma, testin kesinliğini artırmak adına her katılımcı için yeterli sayıda sorunun sağlanmaması riski taşıyabilir. Bu nedenle, bu yöntem, testin sonucunun doğruluğunu sağlamak açısından, testin içerik ve madde havuzunun yeterliliğine bağlı olarak dikkatlice kullanılmalıdır (van der Linden & Glas, 2010).

Sabit madde sonlandırma kurallarının bir avantajı, belirli bir zaman diliminde testin tamamlanmasını sağlamasıdır (Babcock & Weiss, 2009). Bu durum, özellikle zaman sınırlı sınavlar için önemlidir. Ayrıca, sabit madde sonlandırma, daha önceden belirlenmiş bir sınav süresi veya madde sayısı olan testlerde uygulanması kolay bir stratejidir. Ancak, sabit madde sayısına dayalı sonlandırmanın dezavantajı, katılımcıların yetenek seviyelerini tam olarak belirlemek için yeterli veri toplanmadan testin sona ermesidir. Bu, bazı testlerde, özellikle daha fazla madde ve veri gerektiren durumlarda, istenilen doğruluk seviyesini sağlamayabilir.

2.4.7. Bilgisayar Ortamında Bireye Uyarlanmış Testlerin Avantaj ve Dezavantajları

BOBUT eğitimden sağlık sektörüne kadar birçok alanda giderek daha fazla tercih edilen bir değerlendirme yöntemidir. Bu testlerin temel prensibi, bireyin yetenek seviyesini hızla ve hassas bir şekilde tahmin ederek her bir test sorusunu bu seviyeye uygun hale getirmektir. Bu yaklaşım, geleneksel test yöntemlerine kıyasla hem ölçme doğruluğunu artırmakta hem de test süresini kısaltmaktadır. Ancak, bu uygulamaların avantajları kadar dikkat edilmesi gereken bazı dezavantajları da bulunmaktadır.

BOBUT'un en önemli avantajlarından biri, test süresinin kısaltılmasıdır. Weiss (1985), BOBUT sistemlerinin geleneksel testlere kıyasla yaklaşık %50 daha az soru ile bireyin yetenek düzeyini aynı doğrulukla tahmin edebildiğini belirtmiştir. Bu hem test uygulayıcıları hem de katılımcılar için büyük bir zaman tasarrufu sağlar. Ayrıca, bu testlerde bireyin yetenek düzeyine uygun sorular sunulması, katılımcının motivasyonunu artırabilir. Wainer vd. (2000), bu uyarlanabilirliğin test anksiyetesini azalttığını ve katılımcıların test sürecine daha olumlu bir yaklaşım sergilediğini vurgulamıştır.

Bir diğer avantaj, bu testlerin ölçme doğruluğunu artırmasıdır. Thissen ve Steinberg (1986), BOBUT'un bireyin yetenek düzeyine uygun sorular sunarak tahmin hatasını minimize ettiğini ortaya koymuştur. Bu, özellikle geniş bir yetenek dağılımına sahip gruplarda önemlidir. Geleneksel testlerde her birey aynı soruları yanıtlarken, BOBUT her birey için farklı bir test oluşturur. Bu da testin daha adil ve bireyselleştirilmiş bir değerlendirme sağlamasına olanak tanır.

BOBUT, ayrıca test içeriğinin güvenliğini artırır. Geleneksel kâğıt kalem testlerinde, soruların önceden paylaşılması veya kopya çekilmesi riski daha yüksektir. Kingsbury ve Houser (1988), BOBUT her bireye farklı sorular yöneltildiği için bu tür güvenlik sorunlarının önemli ölçüde azaldığını ifade etmiştir. Ayrıca BOBUT, testlerin dijital ortamda saklanması ve yönetilmesi sayesinde, lojistik açıdan da önemli kolaylıklar sunar.

Ancak BOBUT uygulamalarının dezavantajları da göz ardı edilmemelidir. İlk olarak, bu sistemlerin geliştirilmesi ve uygulanması oldukça maliyetlidir. Thompson ve Weiss (2019), bir BOBUT sisteminin yazılım ve donanım gereksinimlerinin yanı sıra, soru havuzunun oluşturulmasının da ciddi bir maliyet ve uzmanlık gerektirdiğini belirtmiştir. Özellikle küçük ölçekli organizasyonlar için bu maliyetler, BOBUT uygulamasının önünde bir engel teşkil edebilir.

Bir diğer dezavantaj, BOBUT'un teknik altyapıya bağımlı olmasıdır. Bu sistemler, genellikle internet bağlantısı ve güçlü bir donanım gerektirir. Özellikle kırsal bölgelerde veya teknolojik altyapının sınırlı olduğu yerlerde, BOBUT'un uygulanması zorlaşabilir. Bu durum, dijital uçurumun derinleşmesine ve eşit erişim sorunlarına yol açabilir (Lunz & Stahl, 1998).

BOBUT'un karşılaştığı bir başka sorun ise, soru havuzunun yeterliliği ve kalitesidir. Bir BOBUT uygulamasının başarılı olabilmesi için geniş ve iyi kalibre edilmiş bir soru havuzuna ihtiyaç vardır. Chang ve Ying (1999), yetersiz veya dengesiz bir soru havuzunun, testin ölçme doğruluğunu ve adaletini olumsuz etkileyebileceğini vurgulamıştır. Ayrıca, soru havuzunun sürekli güncellenmesi ve geçerlik analizlerinin yapılması gerekmektedir.

Son olarak, BOBUT kullanıcı deneyimi açısından bazı zorluklar yaratabilir. Özellikle teknolojiye alışık olmayan bireyler, bu tür testlerde kendilerini rahat hissetmeyebilir. Karışık ara yüzler veya teknik aksaklıklar, test sürecini olumsuz etkileyebilir. Ayrıca, test katılımcılarının bireysel farklılıkları (örneğin, teknoloji kaygısı veya okuryazarlık düzeyi), uygulamaların başarısını etkileyebilir. Bu nedenle, BOBUT sistemlerinin tasarımında kullanıcı dostu ara yüzlerin geliştirilmesi kritik öneme sahiptir (van der Linden & Glas, 2000).

Sonuç olarak, BOBUT, ölçme ve değerlendirme süreçlerinde devrim niteliğinde avantajlar sunmaktadır. Ancak, bu sistemlerin etkin bir şekilde uygulanabilmesi için maliyet, altyapı ve soru havuzu gibi faktörlerin dikkatle ele alınması gerekmektedir. BOBUT uygulamaları, doğru bir şekilde tasarlandığında ve uygulandığında, eğitim ve diğer alanlarda daha adil, hassas ve verimli bir değerlendirme yöntemi sunabilir.

2.5. İlgili Araştırmalar

2.5.1. Bilgisayar ortamında bireye uyarlanmış ve geleneksel kalem-kâğıt testlerinin karşılaştırılmasına ait çalışmalar

BOBUT, eğitim ve değerlendirme alanında son yıllarda önemli bir yenilik olarak karşımıza çıkmaktadır. Geleneksel kâğıt kalem testlerinin sınırlılıklarını aşmak amacıyla geliştirilen bu testler, bireysel yetenek ve ihtiyaçlara göre özelleştirilmiş bir ölçme ve değerlendirme süreci sunmaktadır. Özellikle testlerin geçerlik, güvenirlik ve uygulama süreleri gibi ölçütlerde daha etkili sonuçlar sağladığına dair birçok araştırma bulunmaktadır. Bu kapsamda yapılan çalışmalar şu şekildedir:

Weiss ve Betz (1973) çalışmalarında kâğıt kalem testinin bilgisayar ortamında kişiselleştirilerek uygulanmasının avantaj ve dezavantajlarını değerlendirmişlerdir. Araştırmada elde edilen sonuçlara göre, BOBUT uygulamasının geçerlik ve güvenirlik açısından daha yüksek olduğu ve uygulama süresinin daha kısa olduğu gözlemlenmiştir. Yine benzer bir çalışmada ise McBride ve Martin (1983), BOBUT ve geleneksel kâğıt kalem testlerinin karşılaştırmasına odaklanmış, bireye özelleştirilmiş testlerin kâğıt kalem testlerine kıyasla daha az soruyla daha yüksek güvenirlik sağladığını ortaya çıkarmışlardır. Aynı şekilde, geniş örneklem gruplarında, bireye özelleştirilmiş testlerin geçerliliğinin daha yüksek olduğu belirlenmiştir. BOBUT uygulamasında kullanılan sözel yetenek kestirimleri, aynı sayıda soruya sahip kâğıt kalem testlerine göre daha güvenilir sonuçlar vermiştir.

Kingsbury ve Houser (1988), ilkökul test sistemi bağlamında BOBUT ve kâğıt kalem testlerinin başarı düzeyi kestirimlerini karşılaştırmışlar ve BOBUT ve kâğıt kalem testlerinden benzer puanlar elde edildiği ve bu iki test arasındaki ilişkinin kâğıt kalem testleri arasındaki ilişkiyle benzer olduğu ortaya konulmuştur. Yetenek ölçümlerine üzerine yapılan bir diğer karşılaştırmaya sahip olan çalışmada ise Kaptan (1993), bireysel olarak bilgisayar destekli uyarlanabilir test yöntemi ile kâğıt kalem testi karşılaştırarak yetenek ölçümü yapmıştır. Rasch modeli kullanarak çok aşamalı uygulama seçen araştırmacı, BOBUT uygulamasının sınav süresi ve kullanılan madde sayısı açısından avantajlı olduğunu belirten bulgular elde etmiştir.

Lunz ve Bergstrom (1994), BOBUT'un uygulama koşullarını inceleyen bir ampirik çalışma gerçekleştirmiştir. Bu çalışmada, BOBUT'un doğru yanıt olasılıklarını etkili bir şekilde hedefleyebildiği ve öğrenci performansını önemli ölçüde etkilemeden daha uzun testler oluşturulabileceği bulunmuştur. Bu durum, BOBUT ölçüm bütünlüğünü korurken test uzunluğunu optimize edebildiğini ve kâğıt kalem testlerine kıyasla daha kişiselleştirilmiş bir değerlendirme deneyimi sunduğunu ortaya koymaktadır.

Weiss (2004), bilgisayar ortamında uyarlanmış testlerin (eğitim ve rehberlik alanlarında etkili ve verimli ölçüm sağlama potansiyelini ele almaktadır. Çalışmada, BOBUT'un geleneksel testlere kıyasla daha kısa sürede daha doğru sonuçlar üretebildiği ve bireyselleştirilmiş bir değerlendirme sunduğu vurgulanmaktadır. Uyarlanabilir yapısı sayesinde test edilen bireyin yetenek seviyesine uygun sorular seçilerek ölçme hassasiyeti artırılmakta ve test süreci daha kullanışlı hale gelmektedir.

Kingsbury ve Hauser (2004) bir başka çalışmalarında ise BOBUT'un yararlarını incelemek amacıyla dört ve sekizinci sınıf öğrencileri üzerinde gerçekleştirilen bir çalışmada, okuma ve matematik testlerinin geleneksel test ve BOBUT olarak uygulanmış ve test bilgileri

karşılaştırılmıştır. Araştırmanın sonucuna göre, BOBUT uygulaması, yetenek dağılımının uçlarında yer alan öğrenciler için sağlanan bilgi miktarını geleneksel uygulamaya göre üç kat artırmıştır.

Cömert (2008), geleneksel yöntemle BOBUT'u madde sayısı ve yeterlilik puanları açısından karşılaştırılmasını yapmıştır. Araştırmanın sonucunda, BOBUT'ta kâğıt kalem testleriyle yapılan kıyasla daha az madde kullanıldığı, BOBUT'ta sorulan soru sayısı, özellikle çoklu seçenekli uygulama seçildiğinde, diğer tekniklere göre daha yüksek olduğu ve BOBUT'un değişken yöntemiyle elde edilen yeterlilik puanlarının, geleneksel yöntemle elde edilen puanlara daha yakın olduğu bulunmuştur. BOBUT formunun kâğıt kalem formuyla korelasyonunu incelendiği bir diğer çalışmada ise BOBUT formunun kâğıt kalem formuyla oldukça yüksek korelasyon gösterdiğini ortaya koymuştur (Oswald vd., 2014).

2.5.2. Eğitimde BOBUT kullanımına ilişkin görüş ve deneyimleri içeren çalışmalar

Son yıllarda yapılan araştırmalar, BOBUT'un öğrencilerin sınav deneyimleri, stres düzeyleri, motivasyonları ve katılımları üzerindeki etkilerini anlamaya yönelik önemli bulgular ortaya koymuştur. Hem öğrenciler hem de öğretmenler üzerinde yapılan çalışmalar, BOBUT'un geleneksel kâğıt kalem testlerine kıyasla çeşitli avantajlar sunduğunu ve eğitimde daha etkili bir ölçme-değerlendirme yöntemi olarak kabul edilebileceğini göstermektedir. Bu bölümdeki, BOBUT'un eğitimde kullanımına ilişkin bireylerin görüş ve deneyimlerini ele alan çalışmalar incelenmiş, bu testlerin sunduğu faydalar ve karşılaşılan zorluklar çeşitli araştırmacılar tarafından dile getirilmiştir. Bu kapsamda yapılan çalışmalar şu şekildedir:

Bringsjord (2001), BOBUT ve geleneksel kâğıt kalem testlerinin bireysel sınav katılımcılarının deneyimlerindeki etkilerini, özellikle bilişsel etkileri incelemek için 40 üniversite öğrencisiyle gerçekleştirilen bir veri toplama süreci kullanmıştır. Araştırma, nicel verilerin yanı sıra gözlem ve görüşmelere dayalı nitel veriler de toplanmıştır. Sonuçlar, BOBUT uygulamasını alan öğrencilerin, kâğıt kalem uygulamasını alan öğrencilere kıyasla analitik düşünme sorularına yanıt verirken daha uzun süreler harcadıklarını, soruları şematik olarak çizmek için daha fazla kâğıt ve farklı problem çözme stratejileri kullandıklarını ortaya koymaktadır.

Bir diğer çalışmada ise Miller (2003), BOBUT testlerinin yetenek öngörüsündeki güvenilirliği ve öğrencilerin BOBUT ve kalem-kâğıt testlerine karşı tutumlarını bölgesel düzeyde incelemiştir. Araştırma sonuçları, kâğıt kalem testi ile BOBUT arasında yüksek bir

ilişki olduğunu gösterirken, kız öğrencilerin erkek öğrencilere göre BOBUT puanlarının daha yüksek olduğunu göstermektedir. Bazı öğrencilerin BOBUT uygulamasını, bazılarının ise kâğıt kalem testini gönüllü olarak seçtiğini belirten araştırmacı, özellikle kâğıt kalem testini seçenlerin, tüm soruları görmek ve verdikleri yanıtları yeniden gözden geçirmek istedikleri gibi nedenlerden dolayı kâğıt kalem testini tercih ettikleri belirlenmiştir.

Tercan (2010), öğretmen adaylarının sınıf içi ve uzaktan eğitimde kullanabilecekleri KTK temelli testlerin ve bilgisayar destekli adaptif testlerin kullanımına yönelik görüşlerini incelemiştir. Araştırmacılar, KTK ve bilgisayar destekli testler için bir araç geliştirerek öğretmen adaylarının görüşlerine başvurmuşlardır. Araştırma sonuçları, öğretmen adaylarının BOBUT kavramıyla tanışmadıklarını, önceden hazırlanmış soru havuzları mevcut olduğunda bilgisayar destekli testlerin sınıf içi etkinliklerde kullanımına istekli olduklarını ve uzaktan eğitimde bilgisayar destekli testlerin etkili bir ölçme süreci sağlayacağına inandıklarını göstermiştir.

2.5.3. Farklı test stratejilerinin simülasyonla karşılaştırılmasına ait çalışmalar

BOBUT, test uygulamalarında kullanılan stratejiler ve yöntemlerle ölçme hassasiyeti, test verimliliği ve uyarılma doğruluğu gibi kritik unsurlar üzerinde önemli etkiler yaratmaktadır. Son yıllarda yapılan çalışmalar, farklı BOBUT stratejilerinin karşılaştırılması ve bu stratejilerin test sonuçlarına olan etkilerini anlamaya yönelik kapsamlı bulgular sunmaktadır. Özellikle madde seçim algoritmaları, yetenek kestirim yöntemleri, test sonlandırma kuralları ve adaptasyon stratejileri gibi değişkenler, bireyselleştirilmiş testlerin etkinliğini artırmada önemli bir rol oynamaktadır. Bu bölümde, literatürde yer alan çeşitli çalışmalar ışığında, farklı BOBUT stratejilerinin karşılaştırılması ve bu stratejilerin ölçme-değerlendirme süreçlerindeki avantaj ve dezavantajları ele alınmaktadır. Bu kapsamda yapılan çalışmalar şu şekildedir:

Iseri (2002) çalışmasında matematik puanlarını karşılaştırmak amacıyla farklı bireyselleştirilmiş test stratejileri kullanmıştır. Cevaplayıcıların kestirilen yanıtlayıcı yeteneklerini, yeteneğin kestirilme yöntemi, madde sunum sıklığı, test sonlandırma yöntemi ve cevapları kontrol edebilme değişkenlerine göre karşılaştırmıştır. 3 PL modeline göre ölçeklenen maddelerle çalışmasını yapmış ve Bayes yöntemi ile yeteneğin maksimum olabilirlik yöntemine göre daha etkili olduğunu ve cevapları kontrol edebildikleri stratejide başarılı öğrencilerin yeteneklerinde düşüş olduğunu belirtmiştir. Bir diğer çalışmada ise (Scullard, 2007), Strong ilgi envanterinin bilgisayar ortamında bireye uyarlanabilir formunu geliştirmeyi amaçlamıştır. Araştırmada, üç farklı test sonlandırma kuralı kullanılarak simülasyonlar

gerçekleştirilmiştir. Çalışma sonucunda, uygulanan madde sayısında %60 oranında azalma olduğu ve MTK'ye göre kestirilen θ düzeyleri ile elde edilen puanlar arasındaki korelasyon katsayılarının 0,91-0,98 arasında değiştiği tespit edilmiştir.

Sulak (2013), kullanılan madde seçme yöntemleri, yetenek kestirim yöntemleri ve test durdurma kuralları karşılaştırmalı olarak incelenmiştir. Araştırma kapsamında, 250 maddelik bir madde havuzu ve 2000 simülatif katılımcı oluşturulmuştur. Çalışmada, madde seçme yöntemleri, yetenek kestirim yöntemleri ve test durdurma gibi değişkenlere dayalı 30 farklı koşul analiz edilmiştir. Bulgular, sabit test uzunluğuna dayalı durdurma kuralında en çok olabilirlik tahmini yönteminin, beklenen sonsal dağılım yöntemine göre daha yüksek standart hata değerleri sağladığını göstermiştir. A-tabakalama yöntemi, madde havuzu kullanımında en iyi sonuçları verirken, farklı test uzunluklarında ve durdurma kurallarında da yüksek ölçme kesinliği sağlamıştır. Genel olarak EAP yönteminin test uzunluğunu kısalttığı, ML yönteminin ise daha fazla madde kullanımına yol açtığı görülmüştür. Araştırma, madde seçimi ve yetenek kestirim yöntemlerinin etkili bir şekilde belirlenmesinin testlerin etkinliğini artırdığı sonucuna ulaşmıştır.

Han (2018) çalışmasında, BOBUT için simülasyon çalışmaları yapmayı kolaylaştıran SimulCAT adlı yazılımın prensiplerini tanıtmaktadır. SimulCAT, BOBUT simülasyon çalışmaları için geliştirilmiş olup, madde seçimi, madde maruziyet kontrolü ve içerik dengeleme gibi bileşenleri desteklemektedir. Ayrıca, test uzunluğu belirleme ve puan tahmin algoritmaları gibi yöntemleri de içermektedir. Çalışma, SimulCAT'in kullanımına dair adım adım bir öğretici örnek sunarak, uygulayıcılara ve test geliştiricilere rehberlik etmeyi amaçlamaktadır.

Tat (2019), açıklayıcı madde tepki modellerinin BOBUT madde havuzu geliştirme sürecine, uygulanan madde sayısına ve yetenek parametrelerinin kestiriminin kesinliğine etkisini incelemiştir. Araştırma kapsamında, her birinde 100 madde bulunan 10 madde havuzu ve 1440 katılımcıya ait cevap örüntüleri simülasyonla üretilmiştir. Madde havuzları Rasch model, örtük regresyon, doğrusal lojistik test modeli ve örtük regresyon doğrusal lojistik test modeli ile kalibre edilmiştir. Post-hoc simülasyonlar gerçekleştirilmiş ve tüm modellerin benzer yetenek puan kestirimleri yaptığı, ancak alt gruplar dikkate alındığında modeller arasında anlamlı farklılıklar bulunduğu tespit edilmiştir. Parametre kestirim modelinin madde kullanım sıklığı, test uzunluğu ve test örtüşme oranı üzerinde etkisinin olmadığı görülmüştür. Çalışma, madde havuzunda birey ve madde eş değişkenlerinin etkin olduğu durumlarda açıklayıcı madde tepki modellerinin kullanılmasını önermektedir.

Gür ve Gülleroğlu (2020), 0-1 şeklinde puanlanan maddelerde farklı ölçme koşullarında madde kullanım sıklığı kontrol yöntemlerinin ölçme duyarlılığına ve test güvenliğine etkisi incelenmiştir. Monte Carlo simülasyon yöntemiyle gerçekleştirilmiştir. Simülasyon çalışması, 48 farklı koşulu içermekte olup, madde kullanım sıklığı yöntemlerinin duyarlılık ve güvenilirlik indeksleri hesaplanmış ve birbiriyle karşılaştırılmıştır. Analizler sonucunda, farklı madde kullanım sıklığı kontrolü koşullarında genel olarak ölçme duyarlılıklarında büyük farklılıklar tespit edilmemiştir.

2.5.4. Bilgisayar ortamında bireye uyarlanmış testlerin farklı alanlardaki uygulamaları üzerine çalışmalar

BOBUT, sağlık, dil yeterliliği, bilişsel yetenekler ve eğitim gibi çeşitli alanlarda ölçme ve değerlendirme süreçlerini daha etkili hale getirmek amacıyla yaygın bir şekilde kullanılmaktadır. Son yıllarda yapılan araştırmalar, BOBUT'un geçerlik, güvenilirlik, ölçüm hassasiyeti ve değerlendirme yanlılığı gibi kritik ölçütler üzerindeki olumlu etkilerini ortaya koymaktadır. Simülasyonlar ve canlı uygulamalar aracılığıyla elde edilen bulgular, BOBUT'un yalnızca ölçme süreçlerini kolaylaştırmakla kalmayıp aynı zamanda ölçüm sonuçlarının doğruluğunu artırdığını göstermektedir. Bu bölümde, BOBUT'un farklı alanlardaki uygulamalarına yönelik gerçekleştirilen araştırmalar ele alınarak, bu testlerin etkililiği, avantajları ve sınırlılıkları incelenmektedir. Çeşitli disiplinlerde yapılan çalışmaların bulguları aşağıda sunulmaktadır:

Gibbons vd. (2008), 800 hasta üzerinde duygu durum ve anksiyete ölçme araçlarının (DASS) BOBUT uygulamasının etkililiğini incelemiştir. İlk uygulama simülasyon amacıyla gerçekleştirilirken, ikinci uygulama canlı BOBUT uygulaması için yapılmıştır. Her iki uygulama sonrasında, madde sayısının %95 oranında azaldığı tespit edilmiştir. Ayrıca, canlı BOBUT uygulamasından elde edilen θ düzeyleri ile 616 maddenin tamamının uygulanması sonucunda elde edilen puanlar arasındaki korelasyon katsayısı yüksek bulunmuştur. Bir diğer çalışmada (Kocalevent vd., 2009), stres algısını ölçmek için 104 maddelik bir ölçme aracının BOBUT formunun geliştirilmesi üzerine yapılmıştır. Simülasyon ve canlı BOBUT uygulamalarında özelliğin kestirimi için 0,32 standart hata değeri kullanılmıştır. Kestirilen θ düzeyleri arasındaki korelasyon katsayısı tüm boyutlar için yüksek bulunmuştur. Bu sonuçlar, BOBUT formunun stres algısını ölçmek için güvenilir bir ölçme aracı olabileceğini göstermektedir.

Şenel (2017) çalışmasında, görme engelli öğrencilerin BOBUT ile okuyucu-kodlayıcı yardımıyla aldıkları kâğıt kalem testlerindeki dinlediğini anlama becerileri karşılaştırılmış ve bu uygulamalara ilişkin görüşleri incelenmiştir. Sesli okuma düzenlemesi ile tasarlanan BOBUT uygulaması için çevrimiçi bir platform geliştirilmiş ve testler, 7. ve 8. sınıfa devam eden 51 görme engelli öğrenciye uygulanmıştır. Araştırma sonuçlarında, BOBUT uygulamasında elde edilen yetenek puanlarının kâğıt kalem testine göre anlamlı şekilde düşük olduğu ancak puanlar arasında yüksek düzeyde pozitif bir ilişki bulunduğu görülmüştür. Ayrıca BOBUT'un daha az maddeyle daha kısa sürede tamamlandığı ve daha güvenilir kestirimler yaptığı belirtilmiştir. Öğrenci görüşlerinden elde edilen içerik analizinde, teknik özellikler, test özellikleri ve psikolojik etkiler olmak üzere üç tema belirlenmiş ve öğrencilerin BOBUT'a genel olarak olumlu yaklaştığı ancak farklı sayıda madde verilmesinin testin geçerliliğiyle ilgili soru işaretleri oluşturduğu tespit edilmiştir.

2.5.5. Hem simülasyon hem de canlı BOBUT uygulaması içeren çalışmalar

BOBUT, eğitimden sağlığa kadar birçok alanda geleneksel kâğıt kalem testlerine alternatif olarak, ölçme süreçlerini daha verimli ve etkili hale getirmek amacıyla kullanılmaktadır. Türkiye'de ve uluslararası düzeyde yapılan araştırmalar, BOBUT'un uygulanabilirliğini test etmek için hem post-hoc simülasyonlar hem de canlı uygulamalar gerçekleştirilmiştir. Bu çalışmalar, BOBUT'un ölçme hassasiyeti, kullanışlılığı, madde tasarrufu ve eşdeğerlik açısından sağladığı avantajları kapsamlı bir şekilde ortaya koymaktadır. Bu bölümde, farklı alanlardaki BOBUT uygulamalarının kâğıt kalem testleriyle karşılaştırıldığı ve bu uygulamalardan elde edilen sonuçların detaylı bir şekilde incelendiği araştırmalar ele alınmaktadır. Bu kapsamda yapılan çalışmalar şu şekildedir:

Evans (2010) çalışmasında simüle edilen verilerin gerçek sınav verilerini ne kadar doğru yansıttığını incelemiştir. Araştırmada, gerçek ve simüle edilmiş veriler hem BOBUT hem de doğrusal test formatlarında analiz edilmiştir. Ayrıca, üç farklı test uzunluğunun (25, 35 ve 50 madde) yetenek kestirimi üzerindeki etkisi incelenmiştir. Sonuçlar, doğrusal testlerde simüle verilerle yapılan yetenek kestirimlerinin değişkenlik gösterdiğini ve madde seçimi, test uzunluğu ve katılımcı yetenek seviyelerine bağlı olarak doğruluğun farklılaştığını göstermektedir. Özellikle, uç madde parametre değerlerine sahip testlerin kararsızlık yarattığı ve daha düşük doğrulukta kestirim sağladığı bulunmuştur.

Kalender (2011), yükseköğretime öğrenci seçme sınavının yerine geçebilecek BOBUT uygulamasının uygulanabilirliğini araştırmıştır. Çalışmada gerçek verilere ve simülasyonlara

yer verilmiş, bilgisayar ortamında bireyselleştirilmiş formatın öğrenci seçme sınavının geleneksel kâğıt kalem formatıyla karşılaştırıldığında öğrencilere sorulan soru sayısında azalmalar sağlandığı, yetenek kestirimlerinin yüksek güvenilirliğe sahip olduğu, öğrenci seçme sınavlarından elde edilen yetenek kestirimleri ile BOBUT simülasyonlarından gelen kestirimler arasındaki korelasyonların yüksek olduğu belirtilmiştir.

Bulut ve Kan (2012), BOBUT yöntemini, Akademik Personel ve Lisansüstü Eğitime Giriş Sınavı (ALES) için uygunluğu açısından araştırmışlardır. 2008 yılında uygulanan sınavda yer alan 10.000 kişi rastgele seçilmiş ve bu kişilerin yanıtlarından MTK'nın üç parametrelili lojistik modeline göre madde ve yetenek parametreleri kestirilmiştir. Mevcut sorulardan oluşturulan bir havuzla BOBUT uygulaması simülatif olarak gerçekleştirilmiştir. Post-hoc simülasyonlar sonucunda, BOBUT uygulamasının EAP yöntemiyle 0,25, 0,30 ve 0,40 hata eşik değerleriyle gerçekleştirilebileceği tespit edilmiştir. BOBUT uygulamaları sayesinde soru sayısında %70'e kadar azalma sağlanmış, geleneksel kâğıt kalem testiyle BOBUT uygulamaları arasında yüksek korelasyonlar elde edilmiştir. Araştırma sonuçları, uygun koşullar sağlandığında ALES'in BOBUT yöntemiyle uygulanabileceğini göstermektedir.

Kosan-Aytuğ (2013), Ankara Üniversitesi Tıp Fakültesi'nde yürütülen "Gelişim Sınavı" için bir soru havuzu hazırlanmasını ve bu havuzun kullanımıyla BOBUT yönteminin uygulanabilirliğini test etmeyi hedeflemiştir. Soru havuzunun hazırlanmasında Rasch modeli kullanılarak iki kategorili cevaplar elde edilmiştir. Araştırma sonuçları, kâğıt kalem testi uygulamalarıyla BOBUT yöntemi ile elde edilen yetenek kestirimleri arasında yüksek korelasyon olduğunu göstermektedir. Bu araştırma, Rasch yöntemi için uygun bir soru havuzu hazırlanabileceğini, bu soru havuzunun BOBUT yöntemiyle uygulanarak güvenilir sonuçlar elde edilebileceğini ortaya koymaktadır. BOBUT uygulamasının kullanımıyla sınava girenlerin klasik sınavlarda elde edilenlerden çok daha verimli sonuçlar alabilecekleri düşünülmektedir.

Kezer ve Koç (2014), İngilizce kelime testi üzerinden bir BOBUT uygulaması yapmışlardır. Kâğıt kalem testi ile BOBUT arasında karşılaştırmalı bir analiz gerçekleştirilmiş ve bu karşılaştırma 67 kişi üzerinde yapılmıştır. 67 kişiye her iki uygulama arasındaki yetenek testlerinin değerleri karşılaştırılmıştır. Aynı zamanda yaklaşık 1000 kişiden oluşan bir grup farklı başlatma, sonlandırma ve yetenek kestirme yöntemleri kullanılarak yapay veri üretilmiş ve yetenek kestirimleri arasındaki korelasyon incelenmiştir. BOBUT uygulamasında daha az sayıda madde kullanıldığı ve kalem-kâğıt uygulamasından elde edilen yetenek ölçümlerinin tutarlı olduğu bulunmuştur.

Özbaşı ve Demirtaşlı (2016), bu araştırmada lisans birinci sınıf öğrencilerine uygulanan Bilgi ve İletişim Teknolojileri dersi muafiyet sınavında BOBUT testinin kullanılabilirliğini incelemiştir. Bu çalışmada, test sonlandırma koşullarının ve sabit madde kullanımının etkisiyle farklı BOBUT stratejileri oluşturulmuştur. BOBUT uygulaması, kâğıt kalem testine göre daha az madde kullanarak öğrencilerin yeterlik ölçülerini kestirebilir ve güvenilir sonuçlar verebilir. Canlı BOBUT uygulaması, kâğıt kalem testine göre önemli ölçüde madde tasarrufu sağlamaktadır.

Aybek (2018), Kendini Değerlendirme Envanteri (KDE)'nin BOBUT olarak uygulanabilirliğini incelemek amacıyla yapılmıştır. Araştırmada, 1144 lise öğrencisinin verileri kullanılarak post-hoc simülasyonlar gerçekleştirilmiş ve ayrıca canlı bir BOBUT uygulaması oluşturulmuş ve 25 öğrenci üzerinde uygulanmıştır. Öğrencilerin kâğıt kalem formundan elde ettikleri sonuçlar ile BOBUT uygulamasının sonuçları karşılaştırılmış ve KDE'nin BOBUT olarak başarılı bir şekilde uygulanabileceği gösterilmiştir.

Şimşek (2017), Becerilere Güven Mesleki İlgi Envanteri'nin BOBUT ile uygulanabilirliğini incelemiştir. Bu amaçla, kademeli tepki modeli aracılığıyla parametre kestirimleri gerçekleştirilmiş post-hoc simülasyon çalışmaları yapılmıştır. Simülasyon sonuçları, canlı BOBUT uygulaması için EAP, WMFI madde seçme yöntemi ve $SH < 0,50$ test durdurma kuralının seçilmesinin uygun olduğunu göstermiştir. Ayrıca, BOBUT uygulamasının kâğıt kalem testine göre %50 oranında daha az madde kullanıldığı ve süre açısından da %77 oranında tasarruf sağlandığı görülmüştür. Çalışma sonucunda, BOBUT uygulaması ile kâğıt kalem testi arasında yüksek düzeyde korelasyon bulunmuş ve BOBUT uygulamasının, kâğıt kalem testi ile eşdeğerlik sağladığı bulunmuştur.

Şahin (2017) Anadolu Üniversitesi Yabancı Diller Yüksekokulu'nda (YDYO) öğrenim gören hazırlık sınıfı öğrencilerinin gramer, kelime bilgisi ve genel yetenek becerilerini ölçen çok boyutlu bir bireyselleştirilmiş bilgisayarlı testin geliştirilmesini amaçlanmıştır. Araştırmada, geçmiş yıllarda kullanılan 200 maddelik dört farklı veri setinden bir madde havuzu oluşturulmuş, MTK analizleri sonucunda üç boyutlu bifaktör modelin en uygun model olduğu belirlenmiştir. Gerçek zamanlı BOBUT uygulaması için bir hibrit simülasyon yapılmış, farklı yetenek kestirim yöntemleri, madde seçim yöntemleri ve sonlandırma kurallarının test edildiği 36 farklı koşul analiz edilmiştir. Bulgulara göre, D-rule madde seçimi ve MAP yetenek kestirimi yöntemleri ile standart hataya dayalı sonlandırma kuralı en iyi istatistikleri vermiştir. Gerçek zamanlı çok boyutlu BOBUT uygulaması 99 öğrenci üzerinde gerçekleştirilmiş ve test sonunda ortalama 12.3 madde ile öğrencilerin yetenek düzeylerinin güvenilir bir şekilde

kestirildiği görülmüştür. Ayrıca, 200 maddelik havuzun %30'unun kullanıldığı, katılımcıların %75'inin 10-12 madde yanıtladığı tespit edilmiştir. Kâğıt kalem testiyle gerçek zamanlı uygulama arasında yüksek korelasyon bulunmuş ve BOBUT etkili bir ölçme aracı olduğu sonucuna ulaşılmıştır.

Demir (2019), Mesleki Olgunluk Ölçeği'ni daha kısa sürede ve daha az madde ile ölçebilmek amacıyla BOBUT uygulamasını geliştirmiştir. İlk olarak, 667 bireyin verileri kullanılarak Mesleki Olgunluk Ölçeği'nin madde parametreleri belirlenmiş ve farklı madde seçim yöntemleri (MFI, OWMİ, PWMI, EMI, EPMV, PWEPMI) ile farklı sonlandırma kuralları ($SH < 0,30$, $SH < 0,50$, $SH < 0,70$, $MS = 10$ madde, $MS = 20$ madde) altında simülasyon analizleri gerçekleştirilmiştir. Analizler sonucunda, madde seçim yöntemleri arasında ölçme doğruluğu açısından önemli bir fark bulunmamakla birlikte, ortalama madde sayısı bakımından Maksimum Fisher Bilgi yöntemi ve $SH < 0,30$ sonlandırma kuralı en uygun BOBUT koşulları olarak belirlenmiştir. Çalışmanın ikinci aşamasında, Concerto platformu kullanılarak BOBUT uygulaması geliştirilmiş ve 33 birey üzerinde uygulanmıştır. Sonuçlar, BOBUT ile klasik testten elde edilen mesleki olgunluk düzeyleri arasında orta düzeyde, pozitif yönlü anlamlı bir ilişki olduğunu göstermiştir.

Gökcan (2023) İngilizce kelime bilgisini daha kısa sürede ve daha az madde ile ölçebilmek amacıyla Vocabulary Size Test'in (VST) BOBUT versiyonu (BBT-VST) geliştirmiştir. İlk olarak, 140 maddelik VST'nin kâğıt kalem formu 1622 lisans ve lisansüstü öğrenciye uygulanmış ve testin geçerlik kanıtları araştırılmıştır. Geçerli bir ölçme aracı olduğu belirlendikten sonra, aynı verilerle post-hoc simülasyonlar gerçekleştirilmiştir. Simülasyon bulgularına göre en uygun BOBUT kuralları; EAP yetenek kestirimi, MFI madde seçimi, " $SH \leq 0,225$ " sonlandırma kuralı ve "randomesque = 5" madde kullanım kontrol yöntemi olarak belirlenmiştir. Bu kurallara dayalı olarak, Concerto platformunda BOBUT-VST geliştirilmiş ve 60 lisans ve lisansüstü öğrenci üzerinde hem kâğıt kalem hem de BOBUT versiyonları uygulanmıştır. Sonuçlar, BOBUT ve kâğıt kalem testinden elde edilen yetenek puanları arasında yüksek bir ilişki olduğunu göstermiştir. Ayrıca, BOBUT uygulamasında ortalama test uzunluğu 12 madde olarak bulunmuş ve %91 oranında daha az madde kullanılarak İngilizce kelime bilgisinin güvenilir şekilde kestirildiği tespit edilmiştir.

BÖLÜM III

YÖNTEM

Bu bölümde, araştırma sorularını yanıtlamak için kullanılan araştırma modeli, çalışma grubu, veri toplama aracının geliştirilme süreci, çevrimiçi ortam ve post-hoc simülasyon çalışmaları açıklanmış ve elde edilen verilere dayalı olarak kullanılan istatistiksel yöntemler hakkında bilgi verilmiştir.

3.1. Araştırmanın Modeli

Bu çalışmada bilgisayar ortamında bireye uyarlanmış test uygulamasının farklı yetenek tahmin yöntemleriyle öğrencilere uygulanmasına, uygulamalardan elde edilen verilerin, post-hoc simülasyon ve doğrusal test sonuçlarıyla karşılaştırmasına odaklanmaktadır. Bu kapsamda öncelikle 155 maddelik bir madde havuzu oluşturulmuştur. Bu madde havuzundaki maddeler kâğıt-kalem testi aracılığıyla öğrencilere uygulanmış ve elde edilen veriler üzerinde maddeler kalibre edilmiştir. Kalibre edilen maddeler üzerinden farklı testi başlatma, sonlandırma ve madde seçim yöntemleri kullanılarak farklı yetenek kestirim yöntemleri (ML, EAP ve MAP) için post-hoc simülasyon çalışması yapılmıştır. Bir sonraki aşamada ise testi başlatma, sonlandırma ve madde seçim yöntemleri post-hoc simülasyon sonuçlarından hareketle sabit tutularak farklı yetenek kestirim yöntemleri olan ML, EAP ve MAP yöntemleri kullanılarak ayrı uygulamalar gerçekleştirilmiştir. Bir sonraki aşamada ise 25 maddelik doğrusal bir test oluşturulmuş ve uygulanmıştır. Nihai aşamada üç farklı canlı BOBUT ve doğrusal test uygulamaları farklı strateji ve uygulama koşulları açısından karşılaştırılmıştır. Söz konusu tüm bu uygulamalar BOBUT'un farklı yetenek kestirim yöntemleriyle gerçek-simülasyon koşullarının test edilmesini sağlaması bakımından temel araştırma modelinde tasarlanmıştır.

Temel araştırmalar, kuramlara dayanarak varsayımlar geliştirme, bu varsayımları test etme, elde edilen sonuçları bilimsel olarak yorumlama ve yeni bilgilerin ortaya çıkmasını sağlama amacını taşır (Karasar, 2010). Creswell (2014), temel araştırmaların kuramsal bilgi üretimine odaklandığını ve bu sayede bilimsel disiplinlerin gelişimine katkı sağladığını belirtmektedir. Benzer şekilde Fraenkel ve Wallen (2008), temel araştırmaların mevcut bilgiye yenilerini ekleyerek genel prensipler ve teoriler geliştirmeyi amaçladığını ifade etmektedir. Bu araştırma modeli, bilimsel yöntemleri takip ederek, nesnel verilerin toplanması, analizi ve

yorumlanması üzerine odaklanır. Araştırmacılar, deneyler, gözlem, anketler veya diğer veri toplama tekniklerini kullanarak verileri elde ederler. Daha sonra, bu verileri istatistiksel analizlerle değerlendirirler ve sonuçları çıkarımlarda bulunurlar. Temel araştırmalar, bilimsel bilginin gelişimine katkıda bulunmak, teorileri test etmek ve genel anlamda bir konuyu daha iyi anlamak için önemli bir araçtır (Fraenkel vd., 2019).

3.2. Çalışma Grubu

Araştırmada ön çalışma grubu ve iki farklı çalışma grubu yer almaktadır. Ön çalışma grubu madde yazım süreci sonrası maddelerin uygulama grubunda kontrol edilmesi amaçlı ön çalışma grubudur. Bu çalışma grubu 60 tane ortaokul 7. ve 8.sınıf öğrencisinden oluşmaktadır. Ön çalışma grubuna madde havuzunda bulunan 203 madde uygulanmış ve maddeleri çözme süreleri, davranışları gözlenmiş ve gruptan anlaşılmayan ve yanlış anlaşılan kısımlar konularında görüşleri alınmıştır. Birinci çalışma grubu yine ortaokul 7. ve 8. sınıf öğrencilerinden olmak üzere 1502 kişiden oluşmaktadır. Söz konusu bu çalışma grubundan elde edilen verilerle testin kalibrasyon çalışmaları gerçekleştirilmiştir. Ayrıca post-hoc simülasyon çalışmaları bu grup verileriyle gerçekleştirilmiştir. İkinci çalışma grubu ise yine ortaokul 7. ve 8.sınıf öğrencilerinden oluşan 263 kişiyi içermektedir. Bu çalışma grubuna üç farklı canlı BOBUT ile bir doğrusal test olmak üzere dört kez uygulama yapılmıştır. Söz konusu birinci ve ikinci çalışma grubundan elde edilen verilerle araştırmanın alt problemlerine cevap aranmıştır. Çalışma gruplarına ait özellikler Tablo 3.1’de verilmiştir.

Tablo 3.1. *Çalışma Gruplarına İlişkin Sınıf ve Cinsiyet Dağılımları*

Çalışma Grubu	Okul	Sınıf		Cinsiyet		Toplam
		7.sınıf n (%)	8.sınıf n (%)	Kız n (%)	Erkek n (%)	
Ön çalışma grubu	A	25 (41,67)	35 (58,33)	28 (46,67)	32 (53,33)	60
	B	236 (15,71)	223 (14,85)	229 (15,25)	230 (15,31)	1502
	C	118 (7,86)	129 (8,59)	115 (7,66)	132 (8,79)	
	D	171 (11,38)	217 (14,45)	176 (11,72)	212 (14,11)	
	E	194 (12,92)	214 (14,25)	181 (12,05)	227 (15,11)	
	Toplam	719 (47,87)	783 (52,13)	701 (46,67)	801 (53,33)	
2. çalışma grubu	F	134 (50,95)	129 (49,05)	125 (47,53)	138 (52,47)	263

Tablo 3.1 incelendiğinde ön çalışma grubunda bir okuldaki toplam 60 öğrenci bulunduğu görülmektedir. Bu grubun %41,67’si (n = 25) 7. sınıfta, %58,33’ü (n = 35) ise 8. sınıftadır.

Cinsiyet dağılımına göre grubun %46,67'si (n = 28) kız, %53,33'ü (n = 32) erkektir. Birinci çalışma grubuna ise dört farklı okuldan olmak toplam 1502 öğrenci dâhildir. Öğrencilerin %47,87'si (n = 719) 7. sınıfta, %52,13'ü (n = 783) ise 8. sınıftadır. Cinsiyet dağılımına göre grubun %46,67 (n = 701) kız öğrenciler, %53,33 (n = 801) ise erkek öğrencilerden oluşmaktadır. Son olarak ikinci çalışma grubunda ise bir okuldan toplam 263 öğrenci yer almaktadır. Bu grubun %50,95'i (n = 134) 7. sınıfta, %49,05'i (n = 129) ise 8. sınıftadır. Cinsiyet açısından grupta kız öğrenciler %47,53 (n =125), erkek öğrenciler ise %52,47 (n=138) oranındadır.

3.3. Veri Toplama Aracı

3.3.1. Test Geliştirme Süreci

Araştırmada veri toplama aracı olarak kullanılmak üzere Türkçe okuduğu anlama testi geliştirilmiştir. Test geliştirme kapsamında çeşitli görüşler alınarak alan yazında adı geçen test geliştirme adımlarına uyulmuştur (Anastasi & Urbina, 1997; Croker & Algina, 1986; DeVellis, 2017; Haladyna, 2004; Messick, 1989; Nunnally & Bernstein, 1994). Araştırma kapsamında veri toplama aracının geliştirilmesine ait adımlar şu şekildedir:

Amacın Belirlenmesi

Bu testin temel amacı, 7. ve 8. sınıf Türkçe müfredatında yer alan okuduğunu anlama kazanımlarına dayalı bir test geliştirmektir. Bu kapsamda, Millî Eğitim Bakanlığı (MEB) (2018) tarafından yayımlanan müfredattan toplamda 30 farklı okuduğunu anlama kazanımı belirlenmiş ve bu kazanımlar testin temelini oluşturmuştur. Belirlenen kazanımlar, öğrencilerden beklenen bilişsel süreçleri ve müfredat hedeflerini yansıtmakta olup, testin içeriği bu kazanımlar doğrultusunda oluşturulmuştur. Kazanımlar EK-1'de gösterilmiştir.

Madde Yazarlarının ve Ölçme ve Değerlendirme Ekibinin Oluşturulması

Madde yazım ekibi, Türkçe eğitimi konusunda uzman 6 kişiden oluşturulmuştur. Ekibin 3 üyesi en az yüksek lisans düzeyinde akademik unvana sahip akademisyen, 2 üyesi öğretmen, 1 üyesi ise deneyimli bir madde yazarıdır. Ayrıca, sürecin ölçme ve değerlendirme boyutunda 2 uzman yer almış, soruların kazanımlara, taksonomilere ve müfredat gerekliliklerine uygunluğu değerlendirilmiştir. İki uzman arasında uyumsuzluk durumunda, son bir ölçme ve değerlendirme uzmanına başvurulmuş ve böylece tüm test maddelerinin tutarlılığı sağlanmıştır.

Maddelerin Yazımı

Maddelerin yazımında, bireysel soru yazımı, istasyon tekniği (her yazılan sorunun diğer soru yazarı tarafından denetlenmesi), hazır soru kullanımı ve yapay zekâ destekli soru yazımı gibi çeşitli yöntemler bir arada kullanılmıştır. Bu çok yönlü yaklaşım, soruların müfredata uygunluğunu, öğrencilerin bilişsel düzeylerine hitap etmesini ve farklı okuduğunu anlama becerilerini ölçmesini sağlamayı amaçlamıştır. Ayrıca test maddeleri yenilenmiş Bloom taksonomisi (Anderson & Krathwohl, 2001), Webb'in bilgi derinliği (Webb, 2005), PISA (OECD, 2019) ve PIRLS (Mullis vd., 2015) bilişsel taksonomilerine göre de değerlendirilmiştir. Her soru ilgili taksonomi kapsamında incelenmiş ve ait olduğu bilişsel düzeylere ayrılmıştır.

Ön Uygulama ve Madde Seçimi

Testin geçerliliğini sağlamak amacıyla, 60 kişilik bir grup üzerinde ön uygulama yapılmıştır. Bu aşamada, 203 soru öğrencilere uygulanmış ve soruların geçerliliği, öğrenciler üzerindeki etkisi, anlamlılığı ve netliği değerlendirilmeye alınmıştır. Ön uygulama sonuçlarına dayanarak 155 soru ana uygulama için seçilmiştir. Bu sorular, kazanımlar, taksonomiler ve genel uygunluk açısından detaylı şekilde gözden geçirilmiştir. Sorular, testin amacına uygunlukları göz önünde bulundurularak titizlikle seçilmiştir. Madde havuzuna ait madde, kazanım ve bilişsel alan örnekleri EK-2'de verilmiştir.

Ana Uygulama

Seçilen 155 soru, 6 farklı forma bölünerek birinci çalışma grubundaki tüm katılımcılara uygulanmıştır. Soruların formlara bölme işlemi kazanım, taksonomi uygunluğu ve rastgelelik ilkelerine göre yapılmıştır. Bu kapsamda formların ilgili özellikler açısından eş değer olmasına gayret gösterilmiştir. Toplamda her katılımcı için 6 uygulama tasarlanmış ve her uygulama benzer zamanlarda gerçekleştirilmiştir. Bazı formlar tamamlanamamışsa, eksik formların tamamlanabilmesi için ek bir uygulama yapılmıştır. Bu şekilde, tüm öğrencilerin testi tamamlayabilmesi, verilerin eksiksiz ve güvenilir bir şekilde toplanması amaçlanmıştır.

3.3.2. Verilerin düzenlenmesi

Uygulama sonrası elde edilen veriler, kalitenin artırması ve analizlerin sağlıklı bir şekilde yapılabilmesi için dikkatli bir şekilde düzenlenmiştir. Verilerin düzenlenmesi, özellikle eksik veri ve yanlış cevaplar gibi durumların yönetilmesi açısından kritik bir aşamadır. İlk adımda edilen verilerin doğruluğunu sağlamak adına veri 6 farklı gruba ayrılmıştır. Bu gruplar,

öğrencilerin yanıt verdiği maddelerde rastgele tahmin ve boş bırakılan madde sayısına göre belirlenmiştir. Rastgele tahmin bazı öğrencilerin sorulara bilinçli bir şekilde yanıt vermek yerine formları rastgele işaretlemesi durumudur. Rastgele tahmin, öğrencinin sorulara sistematik bir düşünme süreci olmadan, tamamen rastgele şıklar seçerek yanıt vermesidir. Rastgele tahmin süreci öğrenci cevaplarının belirli bir desen içermesi, test yanıt süresinin kısa olması ve testin zorluk düzeyi ile uyumun olmaması durumlarıyla tanımlanmıştır. Bu kapsamda oluşturulan gruplar şu şekildedir:

- **Grup 0:** Tam veri, rastgele tahmin yok, boş yok (792 kişi)
- **Grup 1:** Rastgele tahmin yok, boş sayısı 1-25 (343 kişi)
- **Grup 2:** Rastgele tahmin yok, boş sayısı 26-50 (169 kişi)
- **Grup 3:** Rastgele tahmin 1 form, boş sayısı 26-50 (110 kişi)
- **Grup 4:** Rastgele tahmin yok, boş sayısı 51-75 (88 kişi)
- **Grup 5:** Rastgele tahmin 1 form, boş sayısı 51-75 (20 kişi)
- **Grup 6:** 2 ve üstü rastgele tahmin formu, boş sayısı 75 üstü (162 kişi)

Grup 5 ve Grup 6, aşırı hatalı veri barındırdığı için çalışma dışı bırakılmış ve sadece diğer gruplar dikkate alınmıştır. Bir sonraki aşamada ise cevap örüntüleri arasında tutarlılık durumlarına belirlenmiştir. Öğrencilerin bu tutarsızlıkları üzerinden yeni gruplar oluşturulmuştur. Grup dağılımı şu şekilde belirlenmiştir:

- **Grup 0:** Tam veri, rastgele tahmin yok, boş yok (549 kişi)
- **Grup 1:** Rastgele tahmin yok, boş sayısı 1-25 (362 kişi)
- **Grup 2:** Rastgele tahmin yok, boş sayısı 26-50 (285 kişi)
- **Grup 3:** Rastgele tahmin 1 form, boş sayısı 26-50 (72 kişi)
- **Grup 4:** Rastgele tahmin yok, boş sayısı 51-75 (151 kişi)
- **Grup 5:** Rastgele tahmin 1 form, boş sayısı 51-75 (24 kişi)
- **Grup 6:** 2 ve üstü rastgele tahmin formu, boş sayısı 75 üstü (59 kişi)

Grup 5 ve Grup 6 tekrar çalışma dışı bırakıldıktan sonra, diğer veri gruplarıyla devam edilmiştir. Veri düzenlemesi sonrasında, grupların sayıları belirli bir düzene konmuş bazı veri grupları oluşturulmuştur.

- **Veri-1:** İlk veri grubuna göre, gruplar 0-1-2-3-4; toplamda 1502 kişi.
- **Veri-2:** İlk veri grubuna göre, gruplar 0-1-2; toplamda 1304 kişi.
- **Veri-3:** İkinci veri grubuna göre, gruplar 0-1-2-3-4; toplamda 1419 kişi.
- **Veri-4:** İkinci veri grubuna göre, gruplar 0-1-2; toplamda 1196 kişi.
- **Veri-5:** İlk veri grubuna göre, grup 0; toplamda 792 kişi.
- **Veri-6:** İkinci veri grubuna göre, grup 0; toplamda 549 kişi.

Kayıp Veri Atama ve Uç Değer Analizi

Veri düzenlemesinin bir parçası olarak eksik veriler için model tabanlı kayıp veri atama yöntemi kullanılmıştır. Bu yöntem, regresyon ve Bayesci yöntemler gibi tekniklerle kayıp verilerin tahmin edilmesini sağlar (Rubin, 2004). Özellikle MTK bağlamında, bu tür yöntemlerin genellikle daha doğru sonuçlar verdiği üzerine alan yazında görüşler bulunmaktadır (Endersi, 2022; Schafer, 1997). Kayıp veriler R programında, "mice" paketi (Van Buuren & Groothuis-Oudshoorn, 2011; Zhang, 2016) aracılığıyla lojistik regresyon yöntemi kullanılarak tahmin edilmiştir.

Ayrıca uç değerlerin belirlenmesi için tek değişkenli ve çok değişkenli analizler yapılmış, tek değişkenli uç değerleri belirlemek amacıyla cevaplara ait Z değerleri incelenmiş ve tüm değerlerin -3 ile +3 arasında olduğu belirlenmiştir (Akbaş ve Koğar, 2020; Barnett & Lewiss, 1994). Çok değişkenli uç değer kapsamında ise Mahalanobis uzaklıkları incelenmiş ve tüm değerlere ait p değerlerinin 0,01'in üzerinde olduğu belirlenmiştir (Tabachnick & Fidell, 2013).

Veri Türlerinin İncelenmesi

Veri düzenlemesinin ardından farklı veri grupları arasında çeşitli psikometrik nitelikler açısından farklılık olup olmadığını değerlendirmek üzere bazı istatistiksel işlemler gerçekleştirilmiştir. İlk olarak farklı veri gruplarında maddelerin yanlılık gösterip göstermediğini belirlemek üzere değişen madde fonksiyonu (DMF) analizleri için "difR" paketi (Magis, Beland, Raiche ve Magis, 2015) kullanılarak Mantel-Haenszel testi uygulanmıştır. Bu test, farklı gruplar arasındaki maddelerin nasıl işlediğini incelemek için kullanılır ve maddelerin geçerliğini artırmak amacıyla önemli bir adımdır (Osterlind & Everson, 2009; Zumbo, 1999). Veri grupları düşük örneklemden daha büyük örnekleme göre karşılaştırılmış nihai aşamada en

düşük örneklem sayısına sahip Veri-1 (549 kişi) ile Veri-6 (1502 kişi) karşılaştırılarak sonuçlar değerlendirilmiştir. Yapılan analiz sonucu elde edilen değerler EK-3'te verilmiştir.

EK-3'te yer alan Mantel-Haenszel yöntemi ile gerçekleştirilen DMF analizi sonucunda, testin farklı veri grupları arasında maddelerin adil ve dengeli bir ölçüm sağlayıp sağlamadığı değerlendirilmiştir. DMF analizinde, her bir madde için odak grup (Veri-1) ve referans grup (Veri-6) arasında istatistiksel olarak anlamlı farklılıkların olup olmadığı incelenmiştir. Testte yer alan toplam 155 maddeden 3'ünün (M44, M140, M148) istatistiksel olarak anlamlı DIF gösterdiği tespit edilmiştir. Bu maddelerin Mantel-Haenszel ki-kare istatistikleri sırasıyla 8,41 ($p < 0,01$), 6,32 ($p < 0,05$) ve 5,43 ($p < 0,05$) olarak hesaplanmıştır. Bu sonuçlar, ilgili maddelerin farklı gruplar için farklı şekilde işlev gösterdiğini göstermektedir. Etki büyüklüğü açısından değerlendirildiğinde, tüm maddeler ETS Delta ölçeğinde "A" kategorisinde sınıflandırılmıştır. Bu kategori, DIF etkisinin ihmal edilebilir düzeyde olduğunu ifade etmektedir (Zumbo, 1999). Ayrıca maddelerin psikometrik niteliklerinin belirlenmesi safhasında bu maddelerin çalışma dışında kaldığı belirlenmiştir. Bunun dışında DIF tespit edilmeyen diğer 152 madde ise gruplar arasında benzer bir şekilde işlev göstermektedir. Sonuç olarak tüm veri türleri arasında maddelerin aynı işlevi gösterdiği belirlendiğinde en büyük örnekleme sahip Veri-6 ile madde parametrelerinin kestirilmesine karar verilmiştir.

Veri türleri arasında madde parametrelerinin farklılaşıp farklılaşmadığını belirlemek üzere Veri-1 ve Veri-6 için 2PL modele göre a ve b parametreleri hesaplanmış sonrasında bu parametreler için bağımsız gruplar t testi analizi yapılmış ve Pearson momentler çarpımı korelasyon katsayıları hesaplanmıştır. Elde edilen sonuçlar Tablo 3.2'de verilmiştir.

Tablo 3.2. Veri1-Veri6 Madde Parametrelerinin Karşılaştırılması

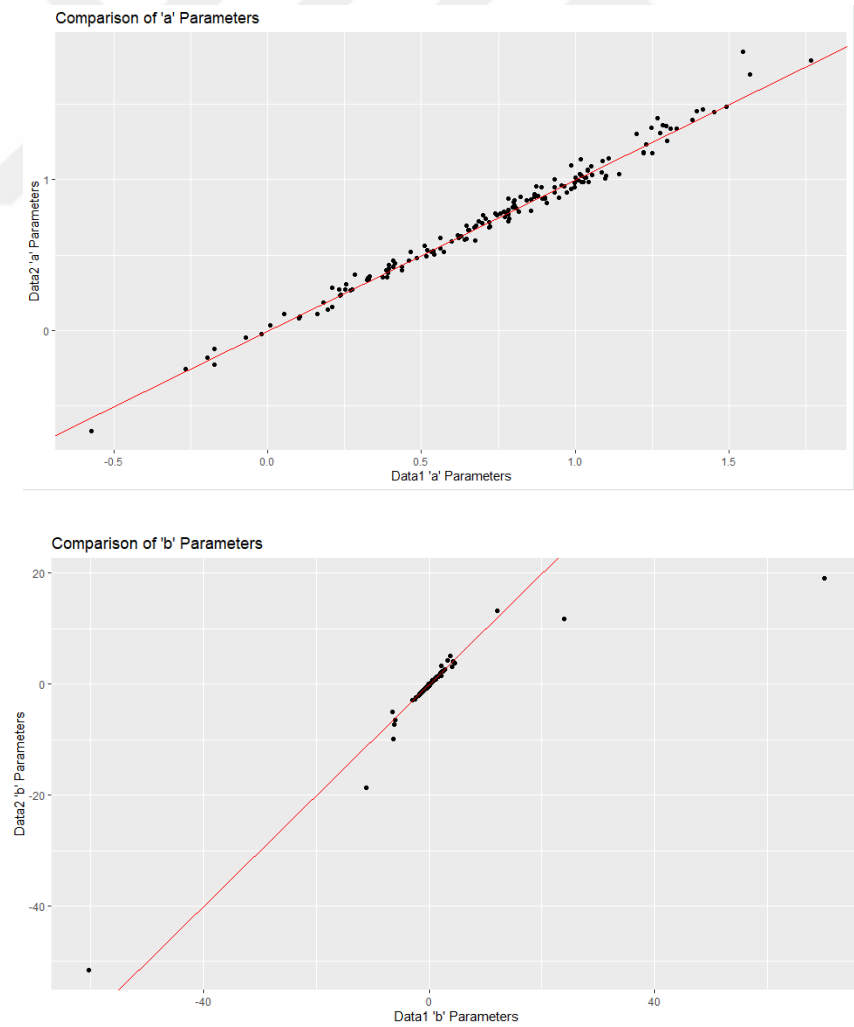
Parametre	t	sd	p	Cohen's d	r
a parametresi	-1,99	154	0,049	0,16	0,98
b parametresi	1,59	154	0,113	0,13	0,87

2PL modeline göre elde edilen a ve b parametreleri, bağımsız gruplar t -testi ile analiz edilmiştir. a parametreleri için yapılan test sonucunda, iki grup arasında istatistiksel olarak anlamlı bir fark bulunsa da, $t(154) = 1,99$, $p = 0,049$, bu farkın küçük bir etki büyüklüğü (Cohen's $d = 0,16$) sebebiyle pratikte anlamlı olmayacağına karar verilmiştir. Öte yandan, b parametreleri için yapılan analizde iki grup arasında anlamlı bir fark bulunmamıştır [$t(154) =$

1,59, $p = 0,113$]. Bu durum, grupların madde güçlük parametreleri açısından benzer olduğunu göstermektedir. Genel olarak analiz sonuçları, test maddelerinin ayırt edicilik parametrelerinde küçük çaplı farklılıklar bulunduğunu ancak güçlük düzeylerinin gruplar arasında tutarlı olduğunu ortaya koymaktadır.

2PL modeli kapsamında elde edilen a ve b parametreleri için yapılan korelasyon analizinde, a parametreleri arasında çok yüksek bir korelasyon ($r = 0,98$) tespit edilmiştir. Bu sonuç, ayırt edicilik parametrelerinin iki veri grubu arasında oldukça tutarlı bir şekilde ölçüldüğünü göstermektedir. b parametreleri arasında ise yüksek düzeyde bir korelasyon ($r = 0,87$) bulunmuştur, bu da zorluk parametrelerinin iki grup arasında büyük ölçüde benzer işlev gördüğünü ifade etmektedir. Genel olarak her iki parametre için de yüksek düzeyde korelasyon tespit edilmesi, testin ayırt edicilik ve zorluk parametrelerinin gruplar arasında tutarlı ve güvenilir olduğunu göstermektedir.

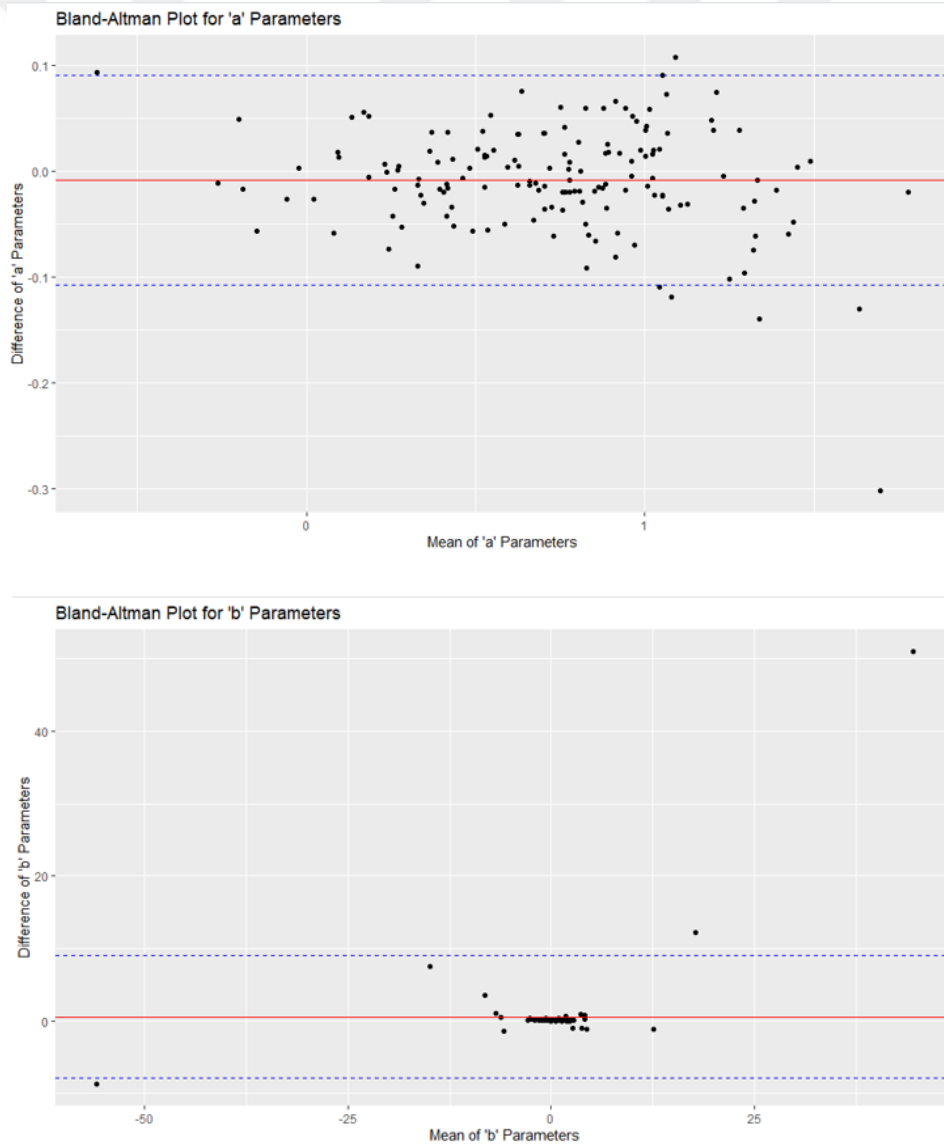
a ve b parametrelerine ait dağılım grafikleri Şekil 3.1’de verilmiştir.



Şekil 3.1. a ve b Parametrelerine Ait Dağılım Grafikleri

2PL modeline ait a ve b parametreleri için yapılan karşılaştırma grafiklerinde, iki veri grubunun parametre değerleri arasındaki ilişki incelenmiştir. a parametreleri grafiğinde verilerin regresyon doğrusuna yakın bir şekilde dağılması ve korelasyonun çok yüksek olması, ayırt edicilik parametrelerinin iki grup arasında tutarlı bir şekilde ölçüldüğünü göstermektedir. b parametreleri için ise yüksek bir korelasyon değeri tespit edilmesine rağmen regresyon doğrusundan uzaklaşan birkaç uç değer dikkat çekmektedir. Bu durum, zorluk parametrelerinin genel olarak tutarlı olmakla birlikte, belirli maddeler için gruplar arasında farklılık gösterebileceğini işaret etmektedir. Ancak söz konusu bu maddelerin psikometrik niteliklerin incelenmesi safhasında madde havuzunun dışında kalan maddeler olduğu belirlenmiştir. Genel olarak testin ayırt edicilik ve zorluk parametreleri açısından güvenilir olduğu söylenebilir.

a ve b parametrelerine ait Bland Altman dağılım grafikleri Şekil 3.2’de verilmiştir.



Şekil 3.2. a ve b parametrelerine ait Bland-Altman grafikleri

Bland-Altman analizine göre *a* parametreleri için iki grup arasındaki farkların büyük çoğunluğu %95 güven aralığı içinde kalmış ve ortalama farkın etrafında simetrik bir dağılım göstermiştir. Bu durum, ayırt edicilik parametrelerinin iki grup arasında yüksek bir tutarlılığa sahip olduğunu göstermektedir. *b* parametreleri için ise farkların çoğunluğu güven aralığı içinde kalmakla birlikte birkaç uç değer gözlemlenmiştir. Bu sonuçlar, zorluk parametrelerinin genel olarak tutarlı olmasına rağmen bazı maddelerde farklılık gösterebileceğini işaret etmektedir. Ancak sonraki incelemelerde söz konusu bu maddelerin çalışma dışında kaldığı belirlenmiştir. Genel olarak Bland-Altman analizi, testin parametrelerinin iki veri grubu arasında güvenilir bir şekilde ölçüldüğünü desteklemektedir.

Sonuç olarak tüm veri türleri arasında maddelerin aynı işlevi gösterdiği belirlendiğinde en büyük örnekleme sahip Veri-6 ile madde parametrelerinin kestirilmesine karar verilmiştir.

Okuduğunu Anlama Testine ilişkin Betimsel Değerler

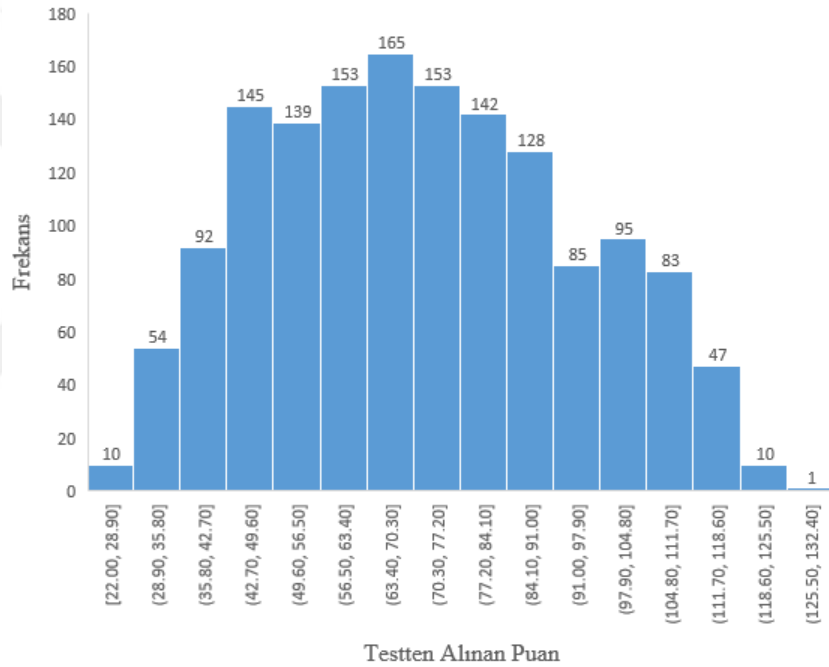
Okuduğunu anlama testi 0-1 şeklinde ikili puanlanan bir başarı testi özelliği göstermesi sebebiyle 0-130 arasında puan alınabilen bir ölçme aracıdır. 1502 öğrenciden elde edilen betimsel değerler Tablo 3.3'te gösterilmiştir:

Tablo 3.3. *Okuduğunu Anlama Testine Ait Test İstatistikleri*

Test İstatistikleri	Değer
Minimum	22,00
Maksimum	127,00
Ortalama	71,41
Sapma	22,55
Varyans	508,60
Basıklık	-0,852 (0,126)
Çarpıklık	0,158 (0,063)
Çok değişkenli basıklık	33,72 (p<0,001)
Çok değişkenli çarpıklık	41111,65 (p<0,001)
KR-20	0,96
Bileşik güvenirlik	0,91

Okuduğunu anlama testinden elde edilen betimsel istatistiklere göre öğrenciler 0-130 arasında puan alabilecekleri testten 22 ile 127 arasında puan almıştır. Bu puanların ortalaması 71,41 olup standart sapması 22,55 ve varyansı 508,60'tır. Çarpıklık değeri 0,158 ve basıklık

değeri -0,852 olup bu değerler test puanlarının normal dağılıma yakın olduğunu göstermektedir. Ancak çok değişkenli çarpıklık (411111,65) ve çok değişkenli basıklık (33,72) değerleri $p < 0,001$ düzeyinde anlamlıdır, bu durum testin çok değişkenli normal dağılım varsayımını karşılamadığını işaret etmektedir (Mardia, 1970). Test puanlarına ait KR-20 güvenirlik katsayısı 0,96 ve bileşik güvenirlik değeri 0,91'dir. Bu değerler da testin oldukça yüksek bir iç tutarlılığa sahip olduğunu göstermektedir (Fornell & Larcker, 1981; Kuder & Richardson, 1937). Bu bulgular, testin öğrencilerin okuduğunu anlama becerilerini güvenilir bir şekilde ölçtüğünü ortaya koymaktadır. Öğrencilerin testten aldıkları puanlara ait dağılımları Şekil 3.3'te gösterilmiştir.



Şekil 3.3. Okuduğunu Anlama Testi Puan Dağılımları

Şekil 3.3'te, okuduğunu anlama testinden alınan puanların frekans dağılımı görülmektedir. Histogram incelendiğinde, puanların yaklaşık olarak normal dağıldığı söylenebilir. Puanlar genellikle orta düzeyde yoğunlaşmış olup özellikle 63,40 ile 70,30 aralığında maksimum frekansa (165 öğrenci) ulaşmıştır. Daha düşük ve daha yüksek puan aralıklarında frekanslar kademeli olarak azalmaktadır. En düşük frekans, 22,00-28,90 ve 125,50-132,40 puan aralıklarında sadece 10 öğrencide gözlemlenmiştir. Bu durum, öğrenci puanlarının büyük çoğunluğunun orta düzeyde toplandığını ve uç değerlerde az yoğunlukta olduğunu göstermektedir. Bu dağılım, testin geniş bir öğrenci kitlesini yeterli şekilde ayırt ettiğini ve puanların genellikle normal dağılıma yakın olduğunu ortaya koymaktadır.

3.3.3. MTK Varsayımlarının İncelenmesi

Tek Boyutluluk

Açımlayıcı faktör analizi (AFA), okuduğunu anlama testinin yapısal olarak tek boyutluluğunu değerlendirmek için kullanılmıştır. Bu analizde, madde puanlarının ikili yapısı dikkate alınarak tetrakorik korelasyon matrisi hesaplanmış ve analizler bu matrise dayanarak yürütülmüştür (Embretson & Reise, 2000). Analiz sonuçlarının değerlendirilmesinde şu kriterler dikkate alınmıştır:

1. Faktör Öz Değerleri: Faktör öz değerlerinin, özellikle ilk faktörün diğer faktörlere göre belirgin şekilde büyük olup olmadığı incelenmiştir. Alan yazında AFA sonucunda başat bir faktör yer almasının tek boyutluluğun göstergesi olarak kabul edilmektedir (Crocker & Algina, 1986; Embretson & Reise, 2000).
2. Faktör Yükleri: Maddelerin tek faktör altında yeterince yüksek yükler (0,30 ve üzeri) alıp almadığı değerlendirilmiştir (Çokluk vd., 2012).
3. Yamaç Birikinti Grafiği (Scree Plot): Faktör öz değerlerinin grafikteki eğim noktası ile belirgin bir şekilde azaldığı yer, belirlenen faktör sayısına ilişkin bir gösterge olarak kullanılmıştır (Cattell, 1966).
4. Paralel Analiz: Rastgele oluşturulan veri setlerinden elde edilen öz değerlerle gerçek veriden elde edilen öz değerlerin karşılaştırılması yapılmıştır (Horn, 1965). Tek boyutluluğun sağlanıp sağlanmadığı bu analiz sonuçlarına göre yorumlanmıştır.

155 maddelik AFA sonucu elde edilen faktör yükleri EK-4'te verilmiştir. Faktör yükleri incelendiğinde 23 maddenin faktör yükünün $<0,30$ altı olması sebebiyle çalışma dışında bırakılmıştır. Sonraki süreçte ise yerel bağımsızlık testinde problemlili Q3 değerine sahip iki madde de çalışma dışında bırakılarak madde havuzunun 130 maddeden oluşmasına karar verilmiştir. Çıkarılan maddeler sonrası kalan 130 maddelik madde havuzuna ait öz değerlere Tablo 3.4'te, faktör yükleri Tablo 3.5'te, yamaç birikinti ve paralel analiz grafikleri Şekil 3.4'te verilmiştir.

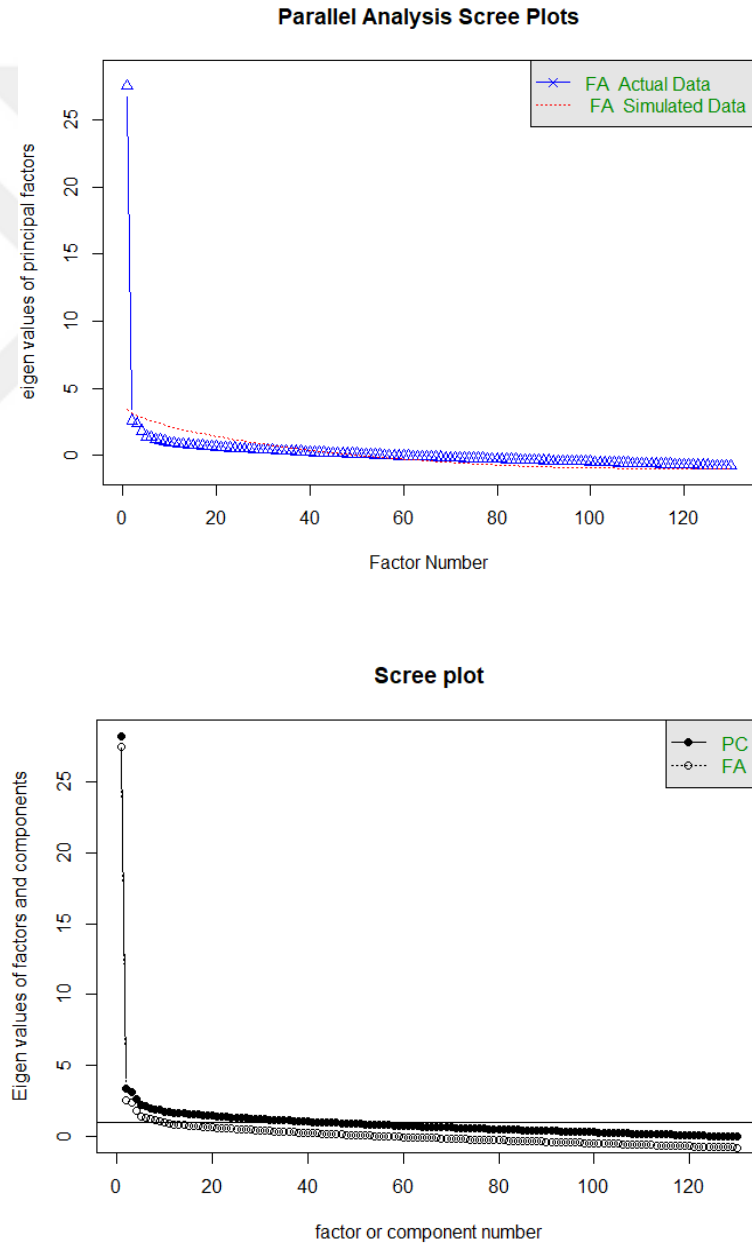
Tablo 3.4. *Okuduğunu Anlama Testinin Bileşenlerine Ait Öz Değerler*

Bileşenler	Öz değer	Kümülatif
1	27,51	27,51
2	2,58	30,09
3	2,37	32,48
4	1,81	34,29
5	1,39	35,68

Tablo 3.5. *Okuduğunu Anlama Madde Havuzuna Ait Faktör Yükleri*

Madde No	Faktör yükü	Madde No	Faktör yükü	Madde No	Faktör yükü	Madde No	Faktör yükü	Madde No	Faktör yükü
M1	0,37	M31	0,26	M61	0,55	M91	0,39	M121	0,58
M2	0,38	M32	0,32	M62	0,52	M92	0,57	M122	0,32
M3	0,50	M33	0,48	M63	0,59	M93	0,42	M123	0,33
M4	0,55	M34	0,35	M64	0,33	M94	0,52	M124	0,53
M5	0,43	M35	0,31	M65	0,52	M95	0,32	M125	0,47
M6	0,30	M36	0,48	M66	0,60	M96	0,33	M126	0,41
M7	0,40	M37	0,46	M67	0,52	M97	0,59	M127	0,42
M8	0,59	M38	0,38	M68	0,32	M98	0,55	M128	0,55
M9	0,31	M39	0,46	M69	0,53	M99	0,47	M129	0,56
M10	0,51	M40	0,40	M70	0,57	M100	0,52	M130	0,49
M11	0,63	M41	0,42	M71	0,56	M101	0,54		
M12	0,51	M42	0,52	M72	0,42	M102	0,33		
M13	0,58	M43	0,32	M73	0,61	M103	0,43		
M14	0,45	M44	0,31	M74	0,50	M104	0,40		
M15	0,52	M45	0,51	M75	0,31	M105	0,47		
M16	0,33	M46	0,36	M76	0,46	M106	0,30		
M17	0,50	M47	0,63	M77	0,29	M107	0,55		
M18	0,37	M48	0,50	M78	0,45	M108	0,36		
M19	0,45	M49	0,33	M79	0,61	M109	0,45		
M20	0,36	M50	0,58	M80	0,63	M110	0,58		
M21	0,45	M51	0,50	M81	0,44	M111	0,59		
M22	0,31	M52	0,30	M82	0,40	M112	0,52		
M23	0,47	M53	0,52	M83	0,49	M113	0,44		
M24	0,30	M54	0,52	M84	0,51	M114	0,31		
M25	0,42	M55	0,45	M85	0,59	M115	0,31		
M26	0,43	M56	0,30	M86	0,47	M116	0,32		
M27	0,37	M57	0,47	M87	0,42	M117	0,63		
M28	0,58	M58	0,46	M88	0,31	M118	0,44		
M29	0,52	M59	0,48	M89	0,53	M119	0,37		
M30	0,47	M60	0,66	M90	0,46	M120	0,52		

Öz değer ve faktör yükleri incelendiğinde ilk bileşenin, toplam varyansın %27,51'ini açıkladığı, test maddelerinin önemli bir kısmının aynı temel boyuta yüklendiğini görülmektedir. Diğer bileşenlerin öz değerleri (2,58, 2,37, 1,81, ve 1,39) ise birinci bileşenle kıyaslandığında oldukça düşüktür. Kümülatif açıklanan varyans %35,68 gibi düşük bir seviyede kalmasına rağmen, ilk bileşenin açık ara baskın olması, testin büyük ölçüde tek bir faktör etrafında yapılandığını göstermektedir. Faktör yükleri genel olarak 0,30 ile 0,63 arasında değişmektedir. Bu, maddelerin büyük bir kısmının ortak bir faktöre anlamlı şekilde yüklendiğini ve bu faktörün testin ana yapısını temsil ettiğini göstermektedir. AFA sonucu elde edilen yamaç birikinti ve paralel analiz grafiği Şekil 3.4'te verilmiştir.



Şekil 3.4. Paralel Analiz ve Yamaç Birikinti Grafiği

Paralel analiz sonuçlarına göre verideki öz değerlerin ilk faktör için anlamlı derecede büyük olduğu ve tek faktörün diğerlerinden belirgin şekilde ayrıldığı görülmüştür. Yamaç birikinti grafiğinde ise öz değerlerin birinci faktörden sonra hızlı bir düşüş gösterdiği ve ikinci faktörden itibaren eğrinin yataylaştığı gözlemlenmiştir. Bu durum, verinin baskın bir faktör tarafından açıklandığını ve tek boyutluluk varsayımını desteklediğini göstermektedir.

Yerel Bağımsızlık

MTK modellerinin temel varsayımlarından biri olan yerel bağımsızlık, maddelerin yanıtlarının yalnızca ölçülen temel faktöre bağlı olması ve bu faktör kontrol edildiğinde maddeler arasında sistematik bir ilişkinin bulunmaması gerektiğini ifade eder. Yerel bağımsızlık, genellikle Q3 artıklık korelasyonları ve model uyum indeksleri ile değerlendirilir (Yen, 1984). Bununla birlikte yerel bağımsızlık, bir testin tek boyutluluğu sağladığı varsayımına dayalıdır (Embretson & Reise, 2000, Hambleton & Swaminathan, 1985). Bu nedenle, bu çalışmada öncelikle testin tek boyutluluğu değerlendirilmiş ve ardından yerel bağımsızlık test edilmiştir. Test aşamasında maddelerin Q3 eşik değeri aşıp aşmadığını aşma durumunda ise bunu meydana getiren problemli madde çiftleri tespit edilmiştir. Bunun sonucunda 2 maddenin (Eski madde no: 104-126) problemli madde çiftleri olması sebebiyle çalışma dışında bırakılmıştır. Nihai aşamada 130 maddelik havuz üzerinden gerçekleştirilen Q3 artık değerler ve uyum değerleri Tablo 3.6’da gösterilmiştir.

Tablo 3.6. Q3 İstatistik Sonuçları ve Uyum Değerleri

Değer	Sonuç
Q3 Minimum	-0,14
Q3 Maksimum	0,20
Q3 Medyan	-0,009
Q3 Ortalama	-0,007
RMSEA	0,030
SRMR	0,037
CFI	0,948
TLI	0,947

Sonuçlar incelendiğinde Q3 değerlerinin minimum -0,14, maksimum 0,20, medyan-0,009 ve ortalama -0,007 değerleri elde edilmiştir. Q3 değerlerinin 0,20’nin altında kalması ve artık korelasyonlarının düşük seviyelerde seyretmesi, yerel bağımsızlık varsayımının

sağlandığını göstermektedir (Yen, 1984). Ayrıca, RMSEA (0,030), SRMR (0,037), CFI (0,948) ve TLI (0,947) gibi uyum indekslerinin sonuçları, modelin tek boyutlu bir yapıyla uyumlu olduğunu ortaya koymaktadır (Hu & Bentler, 1999). Bu bağlamda, testin hem tek boyutluluk hem de yerel bağımsızlık varsayımlarını karşılayarak MTK analizlerine uygun bir yapı sunduğu sonucuna ulaşılmıştır.

Model Veri Uyumu

MTK'nın temelinde yer alan modeller, farklı parametre yapılarına dayanır ve bu parametreler üzerinden maddelerin zorluk, ayırt edicilik ve şans gibi özelliklerini tahmin eder. Bu bağlamda, 1PL (Rasch), 2PL ve 3PL gibi modeller, testlerin yapısına ve amaçlarına uygun olarak tercih edilir. Ancak bu modeller arasında doğru seçim yapmak yalnızca teorik uygunlukla değil, aynı zamanda modelin ampirik veriye sağladığı uyumla da ilişkilidir. Bu çalışmada, AIC, BIC, SABIC gibi bilgi kriterleri ve log-likelihood değerleri kullanılarak farklı MTK modellerinin uyumu değerlendirilmiş, nihai olarak en uygun model belirlenmeye çalışılmıştır. Bu kapsamda elde edilen uyum değerleri Tablo 3.7'de gösterilmiştir

Tablo 3.7. Model-Veri Uyum Değerleri

Model	AIC	BIC	SABIC	HQ	Log-			
					Likelihood	X ²	sd	p
1PL	236446,1	237142,3	236726,2	236705,4	-118092			
2PL	234239,5	235621,3	234795,3	234754,2	-116860	2464,62	129	<0,001
3PL	233798,3	235871,0	234632,1	234570,4	-116509	701,14	130	<0,001

Tablo 3.7'deki model karşılaştırma sonuçları, farklı parametre düzeylerine sahip MTK modellerinin (1PL, 2PL ve 3PL) veri uyumunu değerlendirmektedir. 1PL modeli, en yüksek AIC (236446,1) ve BIC (237142,3) değerleriyle veriye en az uyum sağlayan modeldir. Ayrıca Log-Likelihood değeri (-118092) en düşük olup bu modelin tüm maddeler için ayırt ediciliği sabit kabul etmesinin veriyi yetersiz temsil ettiğini göstermektedir (Hambleton & Swaminathan, 1985). 2PL modeline geçildiğinde, AIC (234239,5) ve BIC (235621,3) değerlerinin düşmesi, daha iyi bir uyuma işaret etmektedir. Log-Likelihood'un iyileşmesi (-116860) ve X² testinin anlamlı olması, 2PL modelinin 1PL'ye kıyasla açık bir üstünlüğe sahip olduğunu göstermektedir. 3PL modeline bakıldığında ise en düşük AIC (233798,3) ve BIC (235871,0) değerleri ve en yüksek Log-Likelihood (-116509) ile teorik olarak en iyi uyum gözlense de 3PL'nin sağladığı ek uyumun sınırlı olduğuna ve modelin karmaşıklığının artmasının dezavantaj yaratabileceğini işaret eder (Baker, 2001; Embretson & Reise, 2000).

Özellikle şans parametresinin 3PL modeline dâhil edilmesi sapmalara yol açabilir ve hesaplama yükünü artırabilir (Lord, 1980). Ayrıca ilgili uyum değerleri arasında 2PL ve 3PL model arasında çok fark olmadığı ki kare sonucunun ise yüksek örneklemelerde anlamlı çıkmaya elverişli olduğu bilinmektedir (Bentler & Bonett, 1980). Tüm sonuçlar göz önünde bulundurulduğunda 2PL modeli, daha az karmaşıklıkla veriye güçlü bir uyum sağlayarak pratik ve teorik açıdan en uygun seçenek olarak öne çıkmaktadır. Bu model, maddelerin ayırt ediciliklerini ayrı ayrı tahmin ederek testlerin geçerliğini ve güvenilirliğini artırırken gereksiz parametre ekleme riskinden de kaçınır (van der Linden & Pashley, 2010). Nihayetinde hem teorik prensipler hem de ampirik bulgular ışığında ve araştırmanın genel amacı doğrultusunda 2PL modeli tercih edilmiştir.

Madde Parametrelerinin Değişmezliği

Madde parametrelerinin değişmezliğini göstermek adına 1502 kişilik veri rastgele olarak iki gruba ayrılmış ve bu gruplar üzerinden çeşitli analizler gerçekleştirilmiştir. İlk olarak söz konusu maddelerin her iki grupta yapısal olarak aynı olup olmadığını belirlemek üzere DMF analizi gerçekleştirilmiştir. Mantel- Haenzel analizi sonucu maddelere ait değerler EK-5'te verilmiştir. Sonuçlar incelendiğinde Mantel-Haenzel yöntemiyle yapılan DMF analizi sonuçlarına göre her bir madde gruplar arasında farklılık gösterip göstermediği açısından incelenmiştir. Analizde kullanılan χ^2 eşik değeri 3,84 ($p < 0,05$) olarak belirlenmiş ve bu eşiği aşan 3 madde (M23, M46, M111) anlamlı DMF göstermiştir. M23 ($\chi^2 = 4,27, p = 0,038$), M46 ($\chi^2 = 3,86, p = 0,049$) ve M111 ($\chi^2 = 6,58, p = 0,010$) küçük DMF etkisine sahip olarak ancak bu da "A" kategorisinde, yani ihmal edilebilir düzeydedir. Bu sonuçlar testin genel adilliğinin büyük ölçüde korunduğunu göstermektedir.

Bir sonraki aşamada her iki grup için 2PL model kapsamında madde parametreleri elde edilmiştir. Her iki gruba ait madde parametreleri EK-6'da her iki gruba ait madde karakteristik eğrileri EK-7'de verilmiştir. Madde parametreleri elde edilen iki arasında ortalama fark ve korelasyonlar incelenmiştir. Her iki grup için yapılan analiz sonucu elde edilen sonuçlar Tablo 3.8'de verilmiştir.

Tablo 3.8. *Madde Parametrelerinin Değişmezliği için Madde Parametrelerinin Karşılaştırılması*

Parametre	t	sd	p	r
a parametresi	0,17	129	0,831	0,90
b parametresi	0,09	129	0,928	0,96

2PL modeline göre elde edilen a ve b parametreleri, bağımsız gruplar t -testi ile analiz edilmiştir. a parametreleri için yapılan test sonucunda, iki grup arasında istatistiksel olarak anlamlı bir fark olmadığı belirlenmiştir, $t(129)= 0,17$, $p = 0,831$. Öte yandan b parametreleri için yapılan analizde, iki grup arasında anlamlı bir fark bulunmamıştır, $t(129)=0,09$, $p = 0,928$. Bu durum, grupların hem ayırt edicilik hem de güçlük parametreleri açısından benzer olduğunu göstermektedir.

2PL modeli kapsamında elde edilen a ve b parametreleri için yapılan korelasyon analizinde, a parametreleri arasında çok yüksek bir korelasyon ($r = 0,90$) tespit edilmiştir. Bu sonuç, ayırt edicilik parametrelerinin iki veri grubu arasında oldukça tutarlı bir şekilde ölçüldüğünü göstermektedir. b parametreleri arasında ise çok yüksek düzeyde bir korelasyon ($r = 0,96$) bulunmuştur, bu da zorluk parametrelerinin iki grup arasında büyük ölçüde benzer işlev gördüğünü ifade etmektedir. Genel olarak her iki parametre için de yüksek düzeyde korelasyon tespit edilmesi, testin ayırt edicilik ve zorluk parametrelerinin gruplar arasında tutarlı ve güvenilir olduğunu göstermektedir. Dolayısıyla ilgili iki alt grup için a ve b parametrelerinin benzer düzeyde olduğu söylenebilir

Yetenek Parametrelerinin Değişmezliği

Yetenek parametrelerinin değişmezliğini varsayımını test etmek üzere 130 soruluk madde havuzu rastgele olacak şekilde ikiye bölünmüştür. Bu kapsamda bölünme sonrası oluşan iki alt test şu şekilde oluşturulmuştur.

Tablo 3.9. *Yetenek Parametrelerinin Değişmezliği için Oluşturulan Alt Testler*

Testler	Madde Numaraları
Alt Test 1	M3, M6, M9, M10, M12, M17, M18, M20, M22, M23, M25, M29, M37, M38, M41, M42, M43, M44, M46, M47, M50, M51, M54, M56, M57, M58, M60, M61, M63, M64, M66, M68, M71, M72, M76, M77, M79, M80, M84, M85, M87, M89, M91, M92, M94, M95, M97, M99, M101, M102, M104, M106, M108, M111, M112, M114, M118, M120, M122, M123, M124, M126, M127, M128, M130
Alt Test2	M1, M2, M4, M5, M7, M8, M11, M13, M14, M15, M16, M19, M21, M24, M26, M27, M28, M30, M31, M32, M33, M34, M35, M36, M39, M40, M45, M48, M49, M52, M53, M55, M59, M62, M65, M67, M69, M70, M73, M74, M75, M78, M81, M82, M83, M86, M88, M90, M93, M96, M98, M100, M103, M105, M107, M109, M110, M113, M115, M116, M117, M119, M121, M125, M129

Her iki alt test için 2PL model kapsamında madde parametreleri elde edilerek grupta yer alan 1502 kişinin yetenek tahminleri yapılmıştır. Söz konusu bu yetenek tahminleri arasında korelasyon değerleri incelenmiştir. Analiz sonucu alt test ve genel testler arasındaki korelasyon katsayıları Tablo 3.10’da verilmiştir.

Tablo 3.10. *Yetenek Parametrelerinin Değişmezliği için Oluşturulan Alt Testler Arasında Korelasyon Katsayıları*

Testler	Alt Test 1	Alt Test 2	Genel Test
Alt Test 1	1,00		
Alt Test 2	0,94*	1,00	
Genel Test	0,93*	0,92*	1,00

* $p < 0,001$

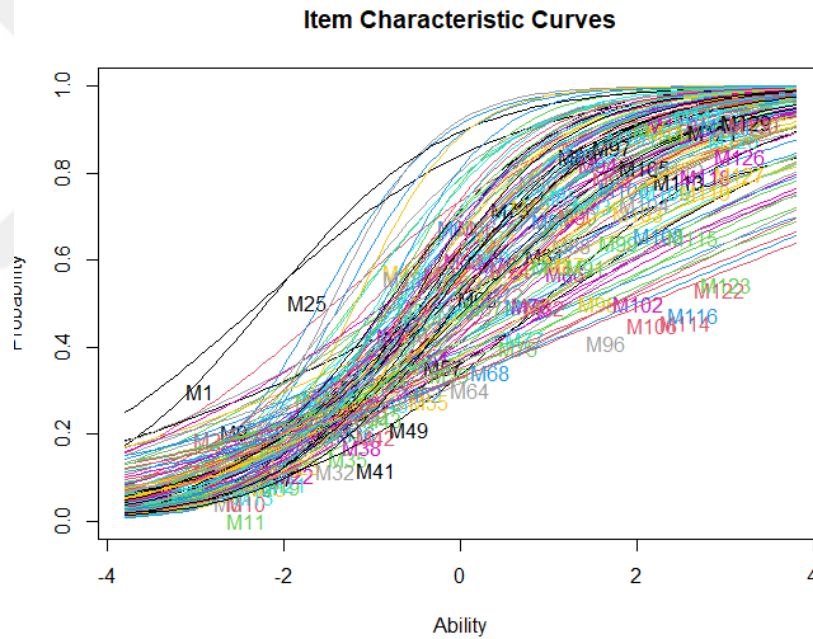
Tablo 3.10’a göre, Alt Test 1 ve Alt Test 2 arasında 0,94 ($p < 0,001$) düzeyinde pozitif ve güçlü bir korelasyon bulunmuştur. Ayrıca, Alt Test 1 ve Genel Test arasındaki korelasyon 0,93 ($p < 0,001$) düzeyindedir. Benzer şekilde, Alt Test 2 ile Genel Test arasındaki korelasyon 0,92 ($p < 0,001$) olarak hesaplanmıştır. Korelasyon katsayılarının yüksek ve anlamlı olması, bireylere ait yetenek tahminlerinin madde setlerinden bağımsız olduğunu ve yetenek parametrelerinin değişmezliği varsayımını desteklediğini göstermektedir. Bu sonuçlar, madde setlerinin rastgele bölünmesi durumunda bile yetenek parametrelerinin değişmezliğinin sağlandığını ve bireylerin yetenek tahminlerinin kullanılan madde setinden etkilenmediğini göstermektedir. Bu durum, kullanılan 2PL modelinin yetenek tahminleri açısından tutarlılığını ve madde bağımsızlığını doğrulamaktadır. Bu bağlamda, test uygulamalarında yetenek parametrelerinin güvenilir bir şekilde tahmin edilmesine yönelik kanıtlar sunmaktadır.

3.3.4. Madde ve Test İstatistikleri

130 soruluk okuduğunu anlama testinin MTK varsayımlarını karşıladığı belirlendikten sonra 2PL model kapsamında madde parametreleri kestirilmiştir. 130 maddelik okuduğunu anlama havuzuna ait a ve b parametreleri ve bunlara ait standart hatalara ilaveten madde bilgi miktarları ve madde uyum değerleri EK-8’de verilmiştir.

Madde parametrelerine ait değerler incelendiğinde maddelerin ayırt edicilik parametrelerinin 0,355-1,842 arasında değişen değerler aldığı belirlenmiştir. Madde ayırt edicilik parametresinin 0,200 üstü değerlerin kabul edilebilir değerler olduğu belirtilmektedir (Baker, 2001; De Ayala, 2018; Embretson & Reise, 2000). Buradan hareketle madde havuzundaki tüm maddelerin ilgili yetenek düzeyi için ayırt edici maddeler olduğu belirmiştir.

Maddelerin güçlük parametrelerinin ise -2,208 ile 1,916 arasında değerler aldığı belirlenmiştir. Genel olarak testin orta yetenek düzeyine sahip kişiler için daha fazla madde barındırdığı ancak uç yetenek düzeyindeki katılımcılar için de maddelerin bulunduğu belirlenmiştir. Madde bilgi miktarlarının ise 0,160 ile 1,782 arasında değişen değerler aldığı belirlenmiştir. Madde bilgi miktarının yüksek olduğu bölgeler, testin bireylerin yetenek düzeylerini daha iyi ölçtüğü alanları temsil etmektedir (Embretson & Reise, 2000). Madde bilgi miktarına ilişkin literatürde net bir kesme değeri olmamakla birlikte bilgi miktarının artması istenen bir durumdur. Madde havuzunda genellikle madde bilgi düzeyi 0,400'ün üzerinde maddeler olduğu ancak bazı maddelerin daha düşük bilgi miktarlarına sahip olmasına rağmen özellikle uç yetenek aralığındaki maddeler için olması sebebiyle testte önemli yerleri olduğu düşünülmektedir. Madde havuzunda yer alan maddelerin bireysel madde karakteristik eğrileri EK-9'da tüm maddelere ait madde karakteristik eğrileri Şekil 3.5'te verilmiştir.

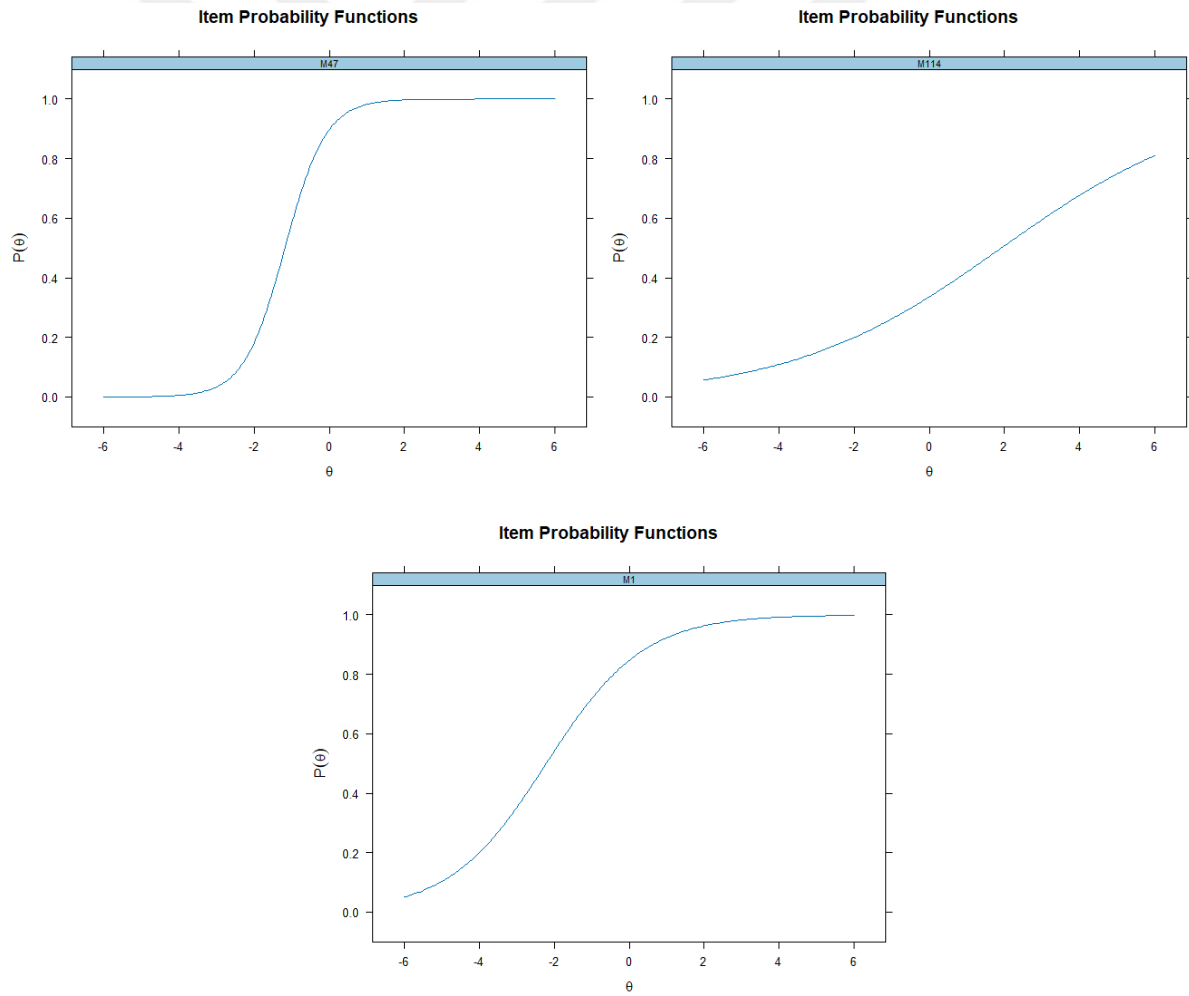


Şekil 3.5. *Tüm Maddelere ait Madde Karakteristik Özellikleri*

MKE grafiği, bireylerin yetenek düzeylerine (θ) bağlı olarak maddelerin doğru cevaplanma olasılıklarını göstermektedir. Şekil 3.5'teki her eğri, MTK kapsamında ilgili maddenin ayırt edicilik ve zorluk parametrelerini temsil etmektedir. Örneğin, M25 gibi eğrisi daha dik olan maddeler, yüksek ayırt ediciliğe sahip olup, belirli yetenek düzeylerinde bireyler arasında güçlü bir ayırım yaparken; M1 gibi daha yatay eğrili maddeler, düşük ayırt ediciliğe sahip ve testin ölçüm gücüne daha az katkıda bulunmaktadır. Eğrilerin yatay eksenindeki konumu, maddelerin zorluk düzeyini yansıtır; M96 gibi eğriler, daha yüksek yetenek düzeylerine hitap ederken, düşük zorluk düzeyindeki maddeler daha geniş bir popülasyonu

hedef alır. Bu grafik, test maddelerinin yetenek düzeylerine göre ne kadar etkili olduğunu ve testin genel ölçüm doğruluğunu değerlendirmek için önemli bilgiler sunmaktadır.

Bu madde havuzu, geniş bir yetenek aralığında ölçüm yapma kapasitesine sahip, çeşitli ayırt edicilik ve zorluk düzeyleri sunan bir yapıya sahiptir. Şekil 3.6'dan testte hem düşük hem de yüksek yetenek düzeylerinde etkili olan maddeler bulunduğunu ve bu maddelerin testin genel bilgi fonksiyonuna önemli katkılar sağladığını göstermektedir. Yüksek ayırt ediciliğe sahip maddeler, belirli yetenek düzeylerinde bireyler arasında hassas bir ayırım yaparak testin ölçüm doğruluğunu artırmaktadır. Ayrıca zorluk düzeylerinin dağılımı, testin farklı yetenek gruplarına hitap etmesini sağlamaktadır. Bu çeşitlilik, testin genel güvenilirliğini ve geçerliğini destekleyen önemli bir avantajdır. Madde havuzu kapsamında ayırt ediciliği en yüksek (M47), hem ayırt ediciliği en düşük hem de güçlük düzeyi en yüksek (M114) ve güçlük düzeyi en düşük (M1) maddelere ait madde karakteristik eğrileri Şekil 3.6'da verilmiştir.

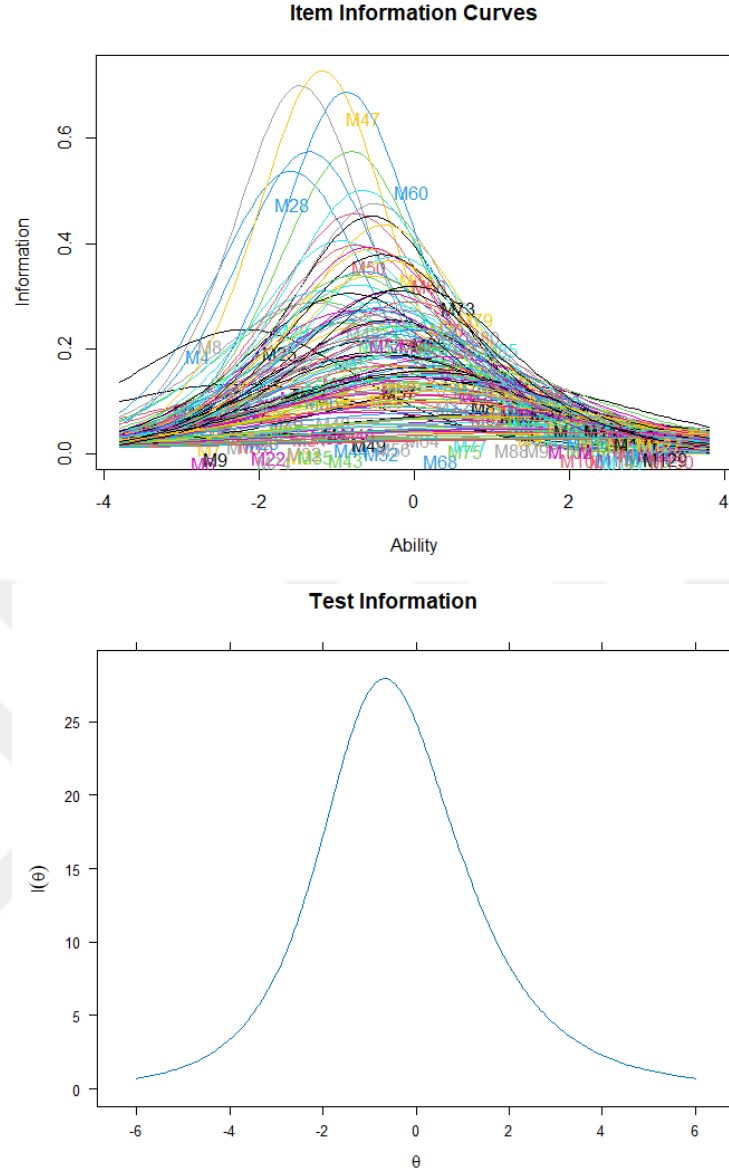


Şekil 3.6. M47, M114 ve M1'e ait madde karakteristik eğrileri

M47 maddesi, testin en ayırt edici maddesi olup ($a = 1,842$), yüksek ayırt edicilik parametresi sayesinde bireylerin yetenek düzeylerinde hassas bir ayırım yapabilmektedir. Şekil, maddenin doğru cevaplanma olasılığının ($P(\theta)$) $-1,00$ civarında hızlı bir şekilde yükseldiğini ve bu yetenek seviyesindeki bireyler için ($b = -1,178$) maksimum bilgi sağladığını göstermektedir. Eğrinin dikliği, maddenin ölçüm doğruluğunu ve testin güvenilirliğine olan katkısını artırmaktadır. Orta zorluk seviyesinde yer alan M47 hem düşük hem de yüksek yetenek düzeylerini kapsayarak dengeli bir ölçüm sağlamaktadır. Bu özellikleriyle M47, testin genel performansını optimize eden kritik bir madde olarak değerlendirilmektedir.

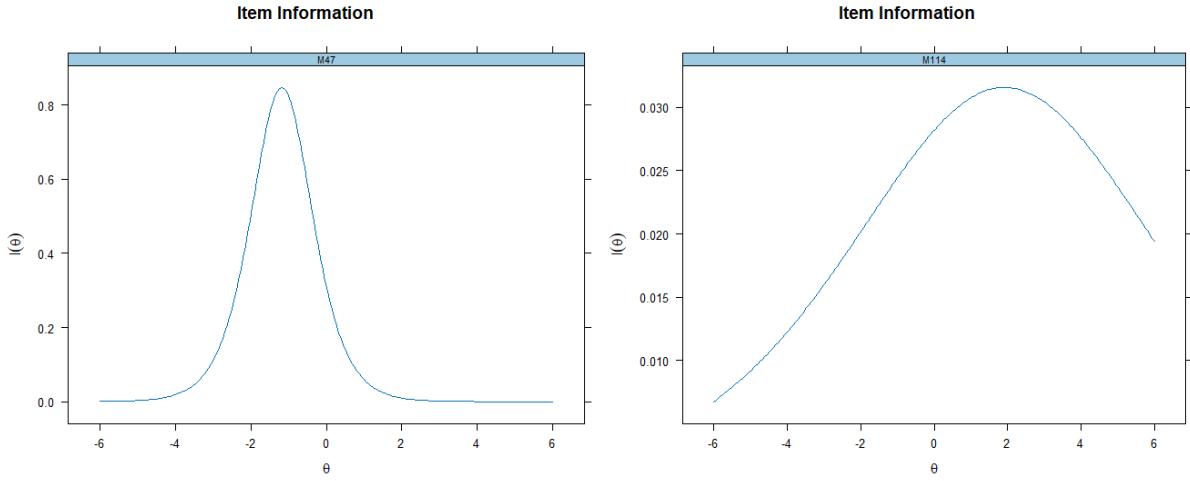
M114 maddesi, madde havuzundaki en az ayırt edici ($a = 0,355$) ve güçlük parametresi en yüksek ($b = 1,916$) madde olarak dikkat çekmektedir. Grafik, doğru cevaplanma olasılığının ($P(\theta)$) yetenek düzeyine (θ) bağlı olarak kademeli ve yavaş bir şekilde arttığını göstermektedir. Eğrinin yatay eksen boyunca çok yavaş dikleşmesi, maddenin ayırt edicilik parametresinin düşük olduğunu ve yetenek düzeyleri arasındaki farklılıkları yeterince güçlü şekilde ayırt edemediğini göstermektedir. Ayrıca eğrinin yükselmesi için daha yüksek yetenek düzeylerine ihtiyaç duyulması, M114'ün yüksek zorluk seviyesinde yer aldığını ve yalnızca yüksek yetenek düzeyindeki bireyler için etkili bir ölçüm sağlayabileceğini ifade etmektedir. Bu özellikleriyle M114 yüksek yetenek düzeyindeki bireyler için spesifik ölçüm sağlayabilen bir madde olarak değerlendirilebilir.

M1 maddesi, madde havuzundaki güçlük düzeyi en düşük ($b = -2,208$) madde olarak öne çıkmaktadır. Grafik, doğru cevaplanma olasılığının ($P(\theta)$) düşük yetenek seviyelerinde ($\theta < -2$) hızla artmaya başladığını ve bu durumun maddenin kolay bir madde olduğunu gösterdiğini ortaya koymaktadır. Eğrinin θ ekseninde erken yükselmesi, düşük yetenek düzeyindeki bireylerin bile bu maddeyi doğru cevaplama olasılığının yüksek olduğunu ifade etmektedir. Eğrinin dikliği ise ayırt edicilik seviyesinin ($a = 0,775$) orta düzeyde olduğunu göstermektedir. M1, düşük yetenek seviyesindeki bireyler için etkili bir ölçüm sağlamak ve testin bu grubun ölçüm doğruluğunu artırmasına katkıda bulunmaktadır. Madde havuzunda yer alan maddelerin bireysel madde bilgi fonksiyonları EK-10'da tüm maddelere ait madde bilgi fonksiyonları ve test bilgi fonksiyonu Şekil 3.7'de verilmiştir.



Şekil 3.7. *Teste ve Maddelere ait Bilgi Fonksiyonları*

Şekil 3.7 testin farklı yetenek seviyelerinde ne kadar bilgi sağladığını ortaya koymaktadır. Eğrilerin yoğunlaştığı -2 ile $+2$ arasındaki yetenek aralığında testin en yüksek ölçüm doğruluğuna ulaştığı, özellikle M47, M28 ve M60 gibi maddelerin bu bölgedeki katkılarının güçlü olduğu görülmektedir. Ortalama yetenek düzeyinde ($\theta = 0$) test oldukça etkili bilgi sağlarken, uç yetenek seviyelerinde (-3 ve $+3$) bilgi sağlayan maddelerin sayısının sınırlı olduğu görülmektedir. Genel olarak test, hedef popülasyonun büyük kısmına uygun şekilde optimize edilmiş olup yüksek bilgi sağlayan maddeler ölçüm doğruluğuna önemli katkıda bulunmaktadır. Madde havuzu kapsamında madde bilgi düzeyi en yüksek (M47) ve madde bilgi düzeyi en düşük (M114) maddelere ait madde bilgi fonksiyonları Şekil 3.8’de verilmiştir.



Şekil 3.8. Madde47 ve M114'e ait Madde Bilgi Fonksiyonları

M47 maddesi, madde havuzundaki bilgi düzeyi en yüksek madde olarak dikkat çekmektedir. Şekil 3.8, maddenin sağladığı bilgi miktarının ($I(\theta)$) yetenek düzeyine (θ) göre değişimini göstermektedir. Eğri, $\theta = 0$ civarında maksimum seviyeye ulaşarak, bu maddenin ortalama yetenek düzeyindeki bireyler için en fazla bilgi sağladığını ifade etmektedir. Eğrinin dar yapısı, maddenin özellikle belirli bir yetenek aralığında (-1 ile $+1$) yüksek ölçüm doğruluğuna sahip olduğunu, ancak bu aralığın dışında katkısının sınırlı olduğunu göstermektedir. Yüksek bilgi miktarı, bu maddenin testin genel güvenilirliğine ve doğruluğuna önemli bir katkı sağladığını ortaya koymaktadır. Bu özellikleriyle M47, hedef popülasyonun ortalama yetenek düzeyine odaklanan testlerde kritik bir madde olarak değerlendirilebilir.

M114, madde bilgi düzeyi en düşük madde olarak dikkat çekmektedir. Şekil, bu maddenin sağladığı bilgi miktarının ($I(\theta)$) yetenek düzeyine (θ) bağlı olarak oldukça sınırlı olduğunu göstermektedir. Eğri, maksimum bilgi düzeyine $\theta = 2$ civarında ulaşmaktadır. Bu durum, M114'ün yalnızca yüksek yetenek düzeylerindeki bireyler için sınırlı ölçüm doğruluğu sunduğunu ve diğer yetenek düzeylerinde bilgi miktarlarının düşük olduğunu göstermektedir.

3.4. Post-hoc simülasyon süreci

Genel özellikler

BOBUT sistemlerinde simülasyon çalışmaları, test tasarımının etkinliğini değerlendirmek ve optimize etmek için kritik bir role sahiptir. Post-hoc simülasyonlar, genellikle mevcut bir veri seti üzerinde BOBUT'un performansını analiz etmek için kullanılan bir yöntemdir (van der Linden & Pashley, 2010). Bu tür simülasyonlar, önceden toplanmış yanıt verileri üzerinden testin birey yetenek seviyelerini ne kadar doğru tahmin ettiğini ve testin bilgi fonksiyonunun nasıl optimize edildiğini incelemek için kullanılır (Luecht & Nungester, 1998). Post-hoc simülasyonlarda, farklı başlangıç yetenek seviyeleri, madde seçim algoritmaları ve test sonlandırma kriterleri gibi koşullar analiz edilebilir (Chang & Ying, 1999). Bu süreçte BOBUT yazılım paketleri, testin farklı bileşenlerini analiz etmek için kapsamlı bir analiz ortamı sunar (Eggen & Verschoor, 2006).

R programlama dili için geliştirilmiş catR paketi, bireye uyarlanmış testler için kapsamlı bir araç seti sunar (Magis & Raiche, 2012). catR, MTK modeli altında bireye uyarlanmış testlerin simülasyonunu destekler ve farklı madde seçim algoritmaları, başlangıç stratejileri ve sonlandırma kuralları dâhil olmak üzere çeşitli parametrelerin ayarlanmasına olanak tanır. Farklı madde seçim algoritmaları kullanılarak test hassasiyeti artırılabilir (Magis & Raiche, 2012). Ayrıca catR kullanılarak belirli bir yetenek seviyesi aralığında testin bilgi fonksiyonunu analiz etmek mümkündür. Post-hoc simülasyon çalışmaları için catR, belirli bir veri setinden BOBUT simülasyonları yaparak testin performansını analiz etmek için etkili bir araç sağlar. Kullanıcılar, başlangıç yetenek seviyelerini (θ) tanımlayabilir, farklı test durdurma kriterlerini uygulayabilir ve çeşitli yetenek tahmin yöntemlerini karşılaştırabilir (Choi vd., 2010). catR, 0-1 veri yapısı ile test simülasyonu yapmak için özellikle uygundur ve büyük ölçekli simülasyon çalışmalarını kolaylaştırır. Bu nedenle catR, hem akademik araştırmalarda hem de uygulamalı test tasarımı projelerinde yaygın bir şekilde kullanılmaktadır.

Bu çalışmada, bireye uyarlanmış testlerin etkinliğini değerlendirmek amacıyla, 1502 kişiden elde edilen 0-1 yapısında bir veri seti üzerinde post-hoc simülasyon çalışması gerçekleştirilmiştir. Post-hoc simülasyonlar, mevcut yanıt verilerinden hareketle BOBUT sistemlerinin farklı parametreler altında nasıl performans gösterdiğini incelemek için etkili bir yöntemdir (Magis & Raiche, 2012; van der Linden & Glas, 2010). Çalışmanın temel amacı; farklı başlangıç yetenek düzeyleri, madde seçim algoritmaları, yetenek kestirim yöntemleri ve test sonlandırma kurallarının testin ölçme doğruluğu ve etkinliği üzerindeki etkilerini

karşılaştırmaktır. Bu bağlamda, BOBUT uygulamasında başlangıç yetenek düzeyi olarak iki farklı ölçüt kullanılmıştır: ***theta = 0*** (nötr başlangıç) ve ***theta = -0,5:0,5*** (geniş bir başlangıç aralığı) (Lord, 1980; Wainer, 2000). Başlangıç yetenek düzeyinin doğru belirlenmesi, özellikle testin ilk aşamalarında doğru madde seçimlerini sağlayarak testin genel performansını artırmaktadır (Chang & Ying, 1999).

Madde seçim algoritması, BOBUT uygulamalarında kritik bir rol oynamaktadır çünkü her bir bireyin yetenek düzeyine en uygun maddeyi seçmek testin bilgi fonksiyonunu optimize eder (van der Linden & Pashley, 2010). Bu çalışmada, madde seçim yöntemi olarak ***MFI***, ***KL*** ve ***MEI*** yöntemleri kullanılmıştır. MFI, bireylerin yetenek seviyesine en fazla bilgi sağlayan maddeyi seçerken KL algoritması, yetenek tahmini doğruluğunu artırmayı hedefler (Eggen & Verschoor, 2006). MEI ise test boyunca beklenen bilgi miktarını maksimize ederek test etkinliğini artırır (Choi vd., 2010).

Yetenek kestirim yöntemi olarak, literatürde yaygın olarak kullanılan üç farklı yöntem değerlendirilmiştir: ***ML***, ***EAP*** ve ***MAP*** (van der Linden & Glas, 2010). ML yöntemi, gözlemlenen yanıt verilerine dayalı olarak bireylerin yetenek seviyesini tahmin eder ve yeterli madde yanıtı olduğunda etkili sonuçlar verir. EAP, yetenek tahmini için bireyin yanıt verileri ile ön bilgi dağılımını da dikkate alır, bu da özellikle az sayıda madde sunulan testlerde avantaj sağlar (Lord, 1980). MAP yöntemi ise bireyin yanıt verilerini ve ön bilgi dağılımını birleştirerek tahmin doğruluğunu artırmayı hedefler (Magis & Raiche, 2012).

Test sonlandırma kuralları, BOBUT'un uzunluğunu ve hassasiyetini belirlemek için kritik bir parametredir (Chang & Ying, 1999). Bu çalışmada, beş farklı sonlandırma kuralı uygulanmıştır: ***standart hata*** tabanlı sonlandırma kuralları olarak ***0,30*** ve ***0,50***, ***sabit madde sayısı*** tabanlı sonlandırma kuralı olarak ***25 madde*** ve kombine kurallar olarak ***0,30 ile 25 madde birlikte*** ve ***0,50 ile 25 madde birlikte***. Standart hata tabanlı sonlandırma, testin ölçme doğruluğunu kontrol ederken, sabit madde sayısı tabanlı sonlandırma, test uzunluğunu belirlemek için kullanılmıştır (Eggen & Verschoor, 2006). Bu kuralların kombinasyonu, BOBUT'un hem ölçme doğruluğunu hem de uygulama verimliliğini optimize etmeyi amaçlar. Çalışmada bu farklı başlangıç, madde seçim, yetenek kestirim ve sonlandırma kuralları kombinasyonlarıyla toplamda ***90 farklı koşul*** test edilmiştir. Bu tasarım, BOBUT uygulamalarının farklı parametreler altında nasıl performans gösterdiğini detaylı bir şekilde analiz etme fırsatı sunar.

Bu çalışmada kullanılan bilgisayar, AMD Ryzen 5 1600 Six-Core Processor (3.2 GHz) işlemciye, 3200 MHz işlemci hızına ve 16 GB RAM kapasitesine sahiptir. Bu donanım, post-hoc simülasyon süreçlerinde kullanılmıştır.

Karşılaştırma Ölçütleri

Simülasyon süresi

Simülasyon süresi, BOBUT uygulamalarında algoritmaların etkinliğini değerlendirmek için kritik bir ölçüttür ve testin pratik uygulanabilirliği, işlem yükü ve maliyet etkinliği ile ilişkilidir (van der Linden & Glas, 2000; Weiss, 2013). Bu çalışmada, başlama kuralları, madde seçim yöntemleri, yetenek kestirim algoritmaları ve sonlandırma kriterlerinin test süreçleri üzerindeki etkileri analiz edilmiştir. Örneğin, standart hata kriteri düşük tutulduğunda (0,30), kestirim doğruluğu için daha fazla madde sunulması gerekebilir, bu da süreci uzatır (Kingsbury & Weiss, 2004). Ayrıca, karmaşık madde seçim algoritmaları (KL yöntemi) daha fazla işlem gücü gerektirirken, MFI gibi basit yöntemler genellikle daha hızlıdır (Chang & Ying, 1996). Simülasyon süresi, büyük veri setlerinde operasyonel maliyetleri etkiler, bu nedenle verimli algoritmaların geliştirilmesi önemlidir (Cohen & Wollack, 2006). Uzun süren testler ise katılımcı yorgunluğuna yol açarak testin geçerliğini azaltabilir (Pitkin & Vispoel, 2001). Kısa simülasyon süreleri, operasyonel avantajlar sağlamanın yanı sıra testin uygulanabilirliğini ve kullanıcı memnuniyetini artırır (van der Linden, 2005).

Ortalama Test Uzunluğu

Ortalama test uzunluğu, BOBUT uygulamalarında, her bir katılımcıya sunulan toplam madde sayısını ifade eder ve testin verimliliğini değerlendiren önemli bir ölçüttür (van der Linden & Glas, 2000). Test uzunluğu, kullanılan başlangıç kuralları, madde seçim yöntemleri, yetenek kestirim algoritmaları ve sonlandırma kriterlerinden doğrudan etkilenir. Örneğin, standart hata temelli sonlandırma kriterleri daha yüksek doğruluk sağlarken testin süresini ve dolayısıyla ortalama madde sayısını artırabilir (Weiss, 2004). Sabit madde sayısına dayalı sonlandırma yöntemleri ise test uzunluğunu öngörülebilir kılarken bireyselleştirme kapasitesini sınırlayabilir (Kingsbury & Zara, 1989). Ortalama test uzunluğu, özellikle geniş ölçekli uygulamalarda önemli bir operasyonel faktördür; kısa testler, kullanıcı memnuniyetini artırabilir ve işlem maliyetlerini düşürebilir (Pitkin & Vispoel, 2001). Bununla birlikte test uzunluğunun çok kısa olması, yetenek kestiriminin doğruluğunu ve güvenirliliğini olumsuz etkileyebilir (Chang & Ying, 1996). Dolayısıyla test uzunluğu ile doğruluk arasında optimal bir denge sağlanması gereklidir.

Gerçek Yetenek ile Kestirilen Yetenek Korelasyonu

Gerçek yetenek ile kestirilen yetenek arasındaki korelasyon, bir BOBUT uygulamasının doğruluk ve geçerliğini değerlendiren temel bir ölçüttür (van der Linden & Glas, 2000). Bu korelasyon, testin bireyin gerçek yetenek seviyesini ne kadar doğru yansıttığını gösterir ve genellikle kullanılan yetenek kestirim algoritması (örneğin, ML, EAP veya MAP) ile yakından ilişkilidir (Weiss, 2004). EAP gibi Bayesian yöntemler, özellikle kısa testlerde yetenek kestirimi için daha istikrarlı sonuçlar verirken ML gibi maksimum olabilirlik yöntemleri daha uzun testlerde yüksek doğruluk sağlar (Chang & Ying, 1996). Korelasyon değeri 1'e ne kadar yakınsa testin bireyselleştirme kapasitesi ve kestirim doğruluğu o kadar yüksektir (Cohen & Wollack, 2006).

RMSE (Root Mean Square Error)

RMSE, bir BOBUT uygulamasının yetenek kestirim hatalarını değerlendiren önemli bir ölçüttür ve kestirilen yetenek değerleri ile gerçek yetenek değerleri arasındaki ortalama farkın karekökünü ifade eder (van der Linden & Glas, 2000). Bu ölçüt, testin genel doğruluğunu ve yetenek kestiriminin güvenilirliğini belirlemede kritik bir role sahiptir. RMSE'nin düşük olması, testin ölçme doğruluğunun yüksek olduğunu ve yetenek kestiriminde sistematik hataların minimum seviyede gerçekleştiğini gösterir (Weiss, 2004). BOBUT uygulamalarında RMSE, test süreçlerini optimize etmek ve test süresi ile kestirim doğruluğunu dengelemek için önemli bir parametre olarak kullanılır. Örneğin, daha düşük standart hata kriterleri (örneğin, 0.30), yetenek kestiriminde daha az hata yapılmasını sağlayarak RMSE'yi azaltabilir; ancak bu durumda testin tamamlanması için daha fazla madde sunulması gerekebilir, bu da test süresini uzatır (Chang & Ying, 1996). Bununla birlikte çok kısa testler kullanıldığında RMSE genellikle artar çünkü yeterli veri toplanmadığı için yetenek kestirimi daha belirsiz hâle gelir.

Bias (Yanlılık)

Bias, bir BOBUT uygulamasında sistematik hatayı ölçen önemli bir parametredir ve yetenek kestirimlerinin gerçek yetenek değerlerine kıyasla sistematik bir sapma içerip içermediğini gösterir (Weiss, 2004). Bias'ın sıfıra yakın olması, yetenek kestirimlerinin tarafsız olduğunu ve herhangi bir belirli yetenek grubuna karşı avantaj veya dezavantaj oluşturmadığını ifade eder. Bu durum, testin adil ve dengeli ölçme yaptığını gösterir. Bias'ın değerlendirilmesi, özellikle farklı yetenek düzeylerine sahip test katılımcıları için önemlidir. Bazı madde seçim yöntemleri veya kestirim algoritmaları belirli yetenek aralıklarında sistematik olarak düşük veya yüksek kestirimler yapabilir. Örneğin, düşük yetenek düzeylerindeki bireyler için

kestirimler sistematik olarak yüksek çıkıyorsa, testin bu grubu olduğundan daha yetenekli değerlendirdiği söylenebilir (van der Linden & Hambleton, 1997). Benzer şekilde, yüksek yetenek düzeylerindeki bireyler için kestirimler sürekli düşükse bu durum testin bu kişilerin yeteneklerini yeterince yansıtamadığını gösterir.

Madde Örtüşme Oranı

Madde örtüşme oranı, farklı test oturumlarında aynı maddelerin ne sıklıkla kullanıldığını ölçen önemli bir parametredir ve BOBUT'un bireyselleştirme kapasitesini değerlendirmek için kritik bir ölçüt olarak kabul edilir (Cohen & Wollack, 2006). Bu oran, bir testin farklı bireyler için ne kadar benzersiz bir madde seçimi sunduğunu belirlemek açısından önemlidir. Özellikle geniş bir madde havuzuna sahip testlerde, düşük madde örtüşme oranı testin her katılımcıya özgü olmasını sağlayarak bireye uyarlanmış testlerin temel avantajlarından biri olan kişiselleştirme ilkesini destekler (Wainer vd., 2000). Madde örtüşme oranının düşük olması, test güvenliğini artırmanın yanı sıra, madde ezberleme veya paylaşma yoluyla haksız avantaj elde edilmesini önler. Geleneksel sabit testlerde her katılımcıya aynı sorular yöneltildiğinden, belirli test maddelerinin önceden bilinmesi testin geçerliğini olumsuz etkileyebilir. Ancak BOBUT uygulamalarında düşük madde örtüşme oranı, her bireyin farklı sorularla karşılaşmasını sağlayarak testin güvenliğini artırır ve ölçme doğruluğunu korur (Davey & Parshall, 1995).

Ortalama Standart Hata

Ortalama standart hata, bilgisayarlı uyarlamalı testlerde bireylerin yetenek düzeylerinin tahmin edilmesinde kullanılan bir güvenilirlik ölçütüdür. Weiss (2004) tarafından önerilen bu ölçüt, testin ne kadar hassas ve tutarlı bir şekilde yetenek kestirimi yaptığını değerlendirmek için kullanılır. Ortalama standart hata, testin her bir aday için hesaplanan yetenek puanlarının ne kadar güvenilir olduğunu gösterir. Düşük bir ortalama standart hata değeri, testin yetenek kestirimlerinin daha doğru ve tutarlı olduğunu ifade eder.

Ortalama Madde Bilgisi

Ortalama madde bilgisi, BOBUT'ta kullanılan maddelerin, bireylerin yetenek seviyelerini ölçme kapasitesini ifade eden bir kavramdır. van der Linden & Glas (2000) tarafından ele alınan bu kavram, test sırasında seçilen maddelerin ne kadar bilgi sağladığını ve bu bilginin yetenek kestiriminin doğruluğunu nasıl etkilediğini açıklar. Madde bilgisi, özellikle bireyin yetenek düzeyine yakın olan maddelerde daha yüksek olma eğilimindedir. Yüksek bilgi

değerine sahip maddeler, bireyin yetenek seviyesini daha hassas bir şekilde ölçer ve bu da yetenek kestiriminin doğruluğunu artırır.

3.5. Canlı BOBUT ve Doğrusal Test Uygulama Süreci

Canlı BOBUT Uygulama Süreci

Canlı BOBUT uygulaması kapsamında, üç farklı yetenek kestirme yöntemlerinin performansını karşılaştırmak amacıyla 6 farklı uygulama biçimi tasarlanmıştır. Her bir uygulama biçimi rastgele olarak 7. ve 8. sınıfta öğrenim gören toplam 263 ortaokul öğrencisine uygulanmıştır. Tüm öğrenciler; ML, EAP ve MAP yöntemlerini içeren 3 farklı testi tamamlamış, ancak bu testlere girme sıraları farklı olacak şekilde rastgele atanmıştır. Sıralamalar, ML-EAP-MAP, ML-MAP-EAP, EAP-ML-MAP, EAP-MAP-ML, MAP-EAP-ML ve MAP-ML-EAP şeklinde belirlenmiş ve her sıraya bir sınıf atanmıştır. Bu tasarım, teste girme sırasının sonuçlara olan etkisini kontrol altına almayı sağlamış ve yöntemler arasındaki farkların adil bir şekilde analiz edilmesine olanak tanımıştır (Schneider & Shiffrin, 1977). Tüm uygulamalarda yetenek başlangıç düzeyi, madde seçme yöntemi ve sonlandırma kuralı sabitlenmiş, bu sayede uygulamalar arasında yalnızca yetenek kestirme yöntemlerinin etkisi incelenmiştir. Ayrıca test sırasında süre kısıtlaması uygulanmamış, öğrencilerin rahat bir şekilde testleri tamamlamasına olanak tanınmıştır.

Doğrusal Testin Oluşturulması ve Uygulama Süreci

Üç farklı BOBUT uygulaması dışında çalışma grubunda yer alan öğrencilere Concerto platformu üzerinden 25 maddelik doğrusal bir test uygulaması gerçekleştirilmiştir. 25 maddenin oluşturulmasında 30 kazanımın içerisinde 25 kazanım kritik kazanım olarak belirlenmiştir. EK-1’de yer alan koyu renkli kazanımlar kritik kazanımları oluşturmuştur. Kritik kazanım kapsamında yazılan sorulardan normal dağılım eğrisi dikkate alınarak her yetenek aralığına düşmesi gereken madde sayısı belirlenerek o aralığı en iyi temsil edecek a ve b parametrelerine sahip maddeler seçilerek tasarlanmıştır. İlgili test kapsamında 130 maddelik soru havuzu içerisinde 4, 8, 13, 19, 22, 25, 28, 32, 35, 49, 50, 61, 64, 66, 71, 89, 91, 92, 95, 106, 110, 115, 119, 122, 124 numaralı toplam 25 madde uygulamaya dâhil edilmiştir.

3.6. Çevrimiçi BOBUT Uygulama Platformu-CONCERTO

Araştırmada ikinci çalışma grubuna uygulanan Canlı BOBUT uygulamaları Concerto platformu üzerinden gerçekleştirilmiştir. Concerto, interaktif testlerin, anketlerin, sınavların ve diğer ölçme araçlarının oluşturulması, dağıtılması ve yönetilmesi için kullanılan bulut tabanlı bir test oluşturma platformudur. Kullanıcı dostu ara yüzü ve esnek özellikleri sayesinde eğitim, işe alım, yetenek değerlendirmesi, araştırma ve daha birçok alanda yaygın olarak kullanılmaktadır. Concerto platformunun bazı özellikleri şu şekildedir (Aybek, 2016; University of Cambridge Psychometrics Centre, 2024):

- **Test Oluşturma Aracı:** Concerto, çeşitli test türlerini oluşturmanıza olanak tanır. Çoktan seçmeli, doğru/yanlış, açık uçlu sorular ve görevler gibi farklı soru tiplerini destekler. Ayrıca soruları kategorilere ayırabilir, puanlamalar ekleyebilir ve testin düzenini özelleştirebilir.
- **Esnek Dağıtım Seçenekleri:** Oluşturduğunuz testleri çevrimiçi olarak paylaşılabilir. Testler Concerto üzerinden doğrudan öğrencilere, çalışanlara veya katılımcılara gönderilebilir.
- **Veri Analizi ve Raporlama:** Concerto, test sonuçlarını otomatik olarak analiz eder ve ayrıntılı raporlar oluşturur. Bu raporlar, katılımcıların performansını ve güçlü/eksik yönlerinin anlaşılmasına yardımcı olur. Ayrıca istatistiksel analizler, grafikler ve özelleştirilebilir filtreleme seçenekleri sunar.
- **Güvenlik ve Gizlilik:** Concerto, kullanıcı verilerini güvende tutmak için gelişmiş güvenlik önlemleri kullanır. Verilerin gizliliği ve güvenliği konusunda endişe duymadan testler oluşturulabilir ve yönetebilir.
- **Entegrasyon ve Özelleştirme:** Concerto, diğer sistemlerle entegre olabilen bir yapıya sahiptir. API'leri aracılığıyla veri senkronizasyonu yapabilir ve mevcut sistemlerle Concerto'yu entegre edilebilir.

Concerto platformu, özelleştirilebilir bir şekilde BOBUT geliştirmek için kullanılabilir. Concerto, BOBUT için gereken temel bileşenleri sağlayarak bu süreci kolaylaştırır. Ancak Concerto kendi başına bu uygulamayı gerçekleştiremez. Farklı BOBUT algoritmaların çalışıldığı uygulama ve dosyalara aracılık eder. Çalışma kapsamında anlık hata ve tahmin kestirimleri R üzerinden catR paketi aracılığıyla yapılmıştır. Canlı BOBUT uygulamasına ilişkin Concerto platformunun uyarlanma adımları şu şekildedir:

1. Soru bankası oluřturma: Arařtırma kapsamında madde parametreleri belirlenmiř okuduđunu anlama testine ait 130 madde ilgili platforma aktarılmıřtır.
2. Bařlangıç seviyesini belirleme: Post-hoc simölasyon sonuçları kapsamında farklı yetenek kestirim yöntemleri için yapılan üç canlı BOBUT kapsamında bařlangıç seviyesi $\theta = -0,5:0,5$ aralıđı sabitlenmiřtir
3. Kiřilerin yeteneklerinin kestirimi: Arařtırma kapsamında ele alınan üç farklı yetenek kestirim yöntemi her uygulamada birisi olacak řekilde belirlenmiřtir.
4. Madde seřim yöntemi: Post-hoc simölasyon sonuçları kapsamında farklı yetenek kestirim yöntemleri için yapılan üç canlı BOBUT kapsamında madde seřim yöntemi MFI olarak sabitlenmiřtir.
5. Testi durdurma kuralı: Post-hoc simölasyon sonuçları kapsamında farklı yetenek kestirim yöntemleri için yapılan üç canlı BOBUT kapsamında sonlandırma kuralı olarak hem standart hatanın (0,30) hem de sabit durdurma kuralının (25 madde) beraber kombinasyonu seřilerek sabitlenmiřtir. Platforma yönelik ekran görüntüleri řekil 3.9’da verilmiřtir.

HOŞGELDİNİZ-1

Sevgili Öğrenci,

Bu testte **Türkçe okuduğunu anlama** soruları yer almaktadır. Sorular, her biri **4 seçenekten oluşan** çoktan seçmeli formattadır.

- Test, her öğrenciye özel olarak hazırlanmıştır ve sizin performansınıza göre uyarlanacaktır.
- Sınava başlamadan önce ad soyad, okul numarası ve sınıf bilgilerinizi doğru girdiğinizden emin olunuz.
- **Tüm soruları yanıtlamanız gerekmektedir.**
- **Sınav kurallarına uymanız zorunludur.**
 - Sınav başladıktan sonra sorularla ilgili herhangi bir soru sormak yasaktır.
 - Sınav boyunca sessiz bir şekilde çalışmanız beklenmektedir.

Sorularınızı dikkatlice okuyunuz ve size en doğru görünen cevabı işaretleyiniz. Başarılar dileriz!

Teste Başla

Soru:

Penguen ailesinde on yedi farklı tür vardır. Ama hepsinin ortak özelliği, kuş olmalarına rağmen uçamayıp yüzcü olmalarıdır. Bu yüzden omuz yerine kürek gibi güçlü bacakları ve kısa kanatları vardır. Güney Kutbu, Yeni Zelanda, Avustralya, Güney Amerika ve Güney Afrika'da yaşarlar. Bir penguen ne kadar soğuk ve buzlu bir yerde yaşarsa o kadar uzun ve tombul olur. Sıcak yerlere gittikçe penguenlerin boylarının kısa olduğu görülür. Penguenler yüzdüğü için tüyleri, kuş tüylerinden farklıdır. Sık, kısa ve vücut sıcaklığını istenilen şekilde tutabilen tüyleri vardır. Tüylerin hemen altında kalın yağ tabakası bulunur. Böylece eski kürk derece soğukta bile üşümezler.

Bu metinde penguenlerle ilgili aşağıdaki bilgilerden hangisine verilmemiştir?

A. Kaç türü olduğuna

B. Bölgesel özelliklerine

C. Tüylerinin özellikleri ve işlevine

D. Beslenme şekillerine

İleri

Test Bitti.

Katılımınız için teşekkür ederiz.

Lütfen sınav tamamen sona erene kadar yerinizde sessizce bekleyiniz

Next

Şekil 3.9. Concerto Başlangıç, Soru ve Sonlandırma Ekran Görüntüleri

3.7. Verilerin Analizi

Araştırmanın ilk aşamasında post-hoc simülasyon sonuçları genel olarak incelenmiştir. Simülasyon süresi, test uzunluğu, gerçek ve kestirilen yetenek arasındaki korelasyon, RMSE, bias, standart hata ve madde bilgisi ortalama değerleri başlama kuralları, madde seçim yöntemleri, yetenek kestirim yöntemleri ve sonlandırma kuralları açısından incelenmiştir.

Araştırmanın ikinci aşamasında post-hoc simülasyon sonuçlarından elde edilen değerler başlama kuralları, madde seçim yöntemleri, yetenek kestirim yöntemleri ve sonlandırma kurallarının kombinasyonları kapsamında incelenmiştir. İncelemeler grafikler üzerinden yapılarak en iyi ve en yetersiz performansı gösteren durumlar belirlenmiştir. Ayrıca canlı BOBUT uygulaması için kullanılan başlama kuralı, madde seçim yöntemi ve sonlandırma kuralları bu kapsamda belirlenmiştir.

Araştırmanın üçüncü aşamasında canlı BOBUT ve doğrusal test uygulamaları sonuçları kendi içerisinde ve birbirleriyle karşılaştırılmıştır. Uygulamalar çeşitli dağılım ve saçılım grafiklerine ek olarak korelasyonel ve ortalama karşılaştırma testleri aracılığıyla incelenmiştir. Araştırmada aynı öğrenci grubundan farklı ölçümler alındığı için tekrarlı ölçümler varyans analizi, ilişki katsayılarının belirlenmesinde Pearson korelasyon katsayısı kullanılmıştır.

BÖLÜM IV

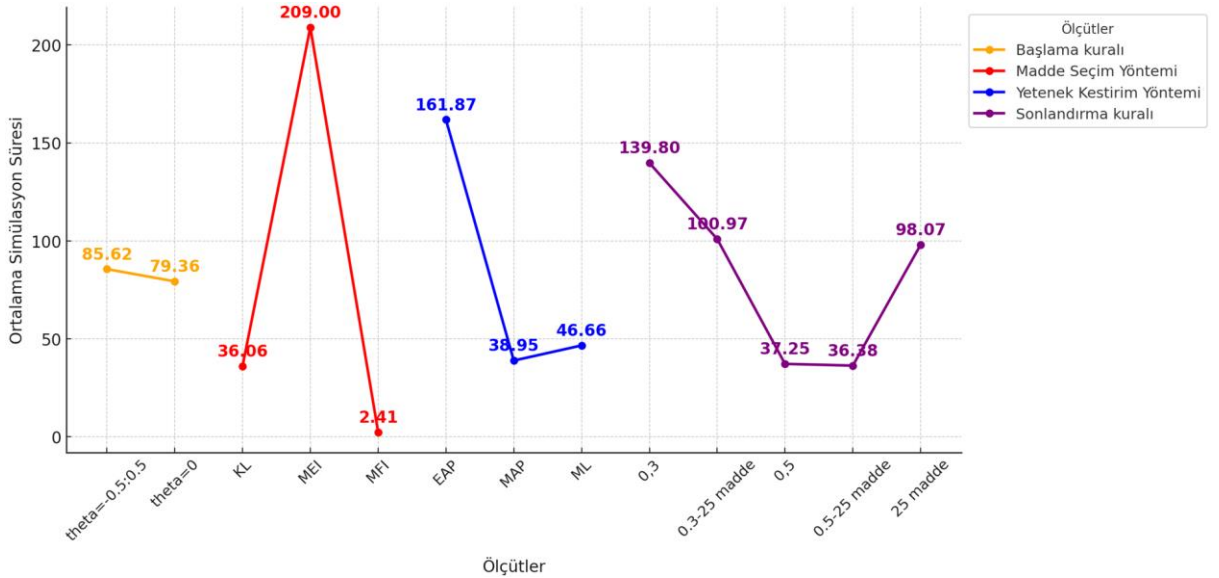
BULGULAR

4.1. Post-hoc Simülasyon Sonuçlarına İlişkin Bulgular

BOBUT uygulamalarında kullanılan farklı başlama kuralları, madde seçim yöntemleri, yetenek kestirim yöntemleri ve sonlandırma kurallarının, test performansı ve sonuçları üzerindeki etkilerini değerlendirmek amacıyla bir dizi analiz gerçekleştirilmiştir. Bu kapsamda, simülasyon süresi, test uzunluğu, korelasyon değerleri, RMSE, bias, madde örtüşme oranı, standart hata ve madde bilgisi gibi çeşitli performans kriterleri incelenmiştir. Tüm post-hoc süreçlerine ilişkin analiz sonuçları EK-11’de gösterilmiştir.

4.1.1. Post-hoc simülasyon genel performans sonuçları

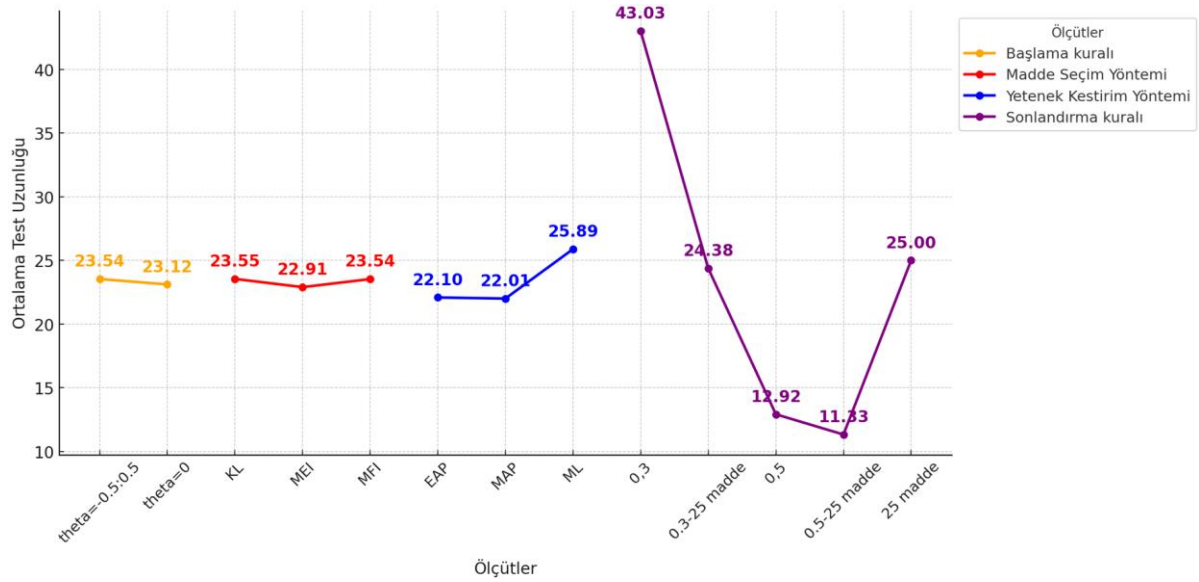
Post-hoc simülasyon kapsamında oluşturulan 90 farklı koşulun simülasyon süresi, test uzunluğu, korelasyon değerleri, RMSE, bias, madde örtüşme oranı, standart hata ve madde bilgisi ölçütlerine göre genel sonuçları Şekil 4.1-Şekil 4.8 arasında verilmiştir.



Şekil 4.1. Post-Hoc Simülasyon Koşullarına göre Ortalama Simülasyon Süreleri

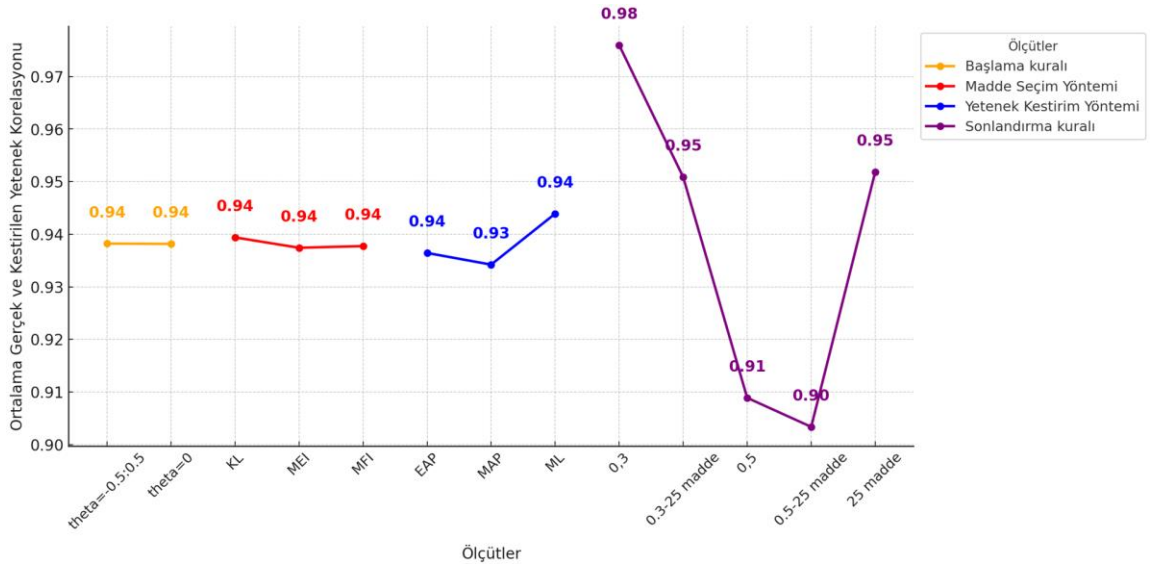
Şekil 4.1’deki simülasyon sonuçlarına göre başlama kuralları arasında $\theta = 0$ seçiminin daha kısa süre gerektirdiği, daha geniş bir başlangıç aralığı seçiminin hesaplama süresini bir

miktar artırdığı görülmüştür. Madde seçim yöntemleri bakımından sonuçlar incelendiğinde, en düşük süreye MFI yöntemiyle ulaşıldığı, MEI yönteminin ise belirgin şekilde en yüksek hesaplama süresini gerektirdiği belirlenmiştir. Hesaplama süresi bakımından yetenek kestirim yöntemleri kıyaslandığında kestirimi en hızlı yapan yöntemin MAP, en uzun sürede yapan yöntemin ise EAP olduğu belirlenmiştir. ML yöntemi ise MAP'ın kestirim süresine göre daha uzun sürede bir sonuç verse de iki yöntemin süre bakımından sonuçları daha benzerdir. Sonlandırma kurallarında ise geniş hata toleransına sahip kuralların hesaplama sürelerini belirgin şekilde kısalttığı, daha hassas kriterlerin ise hesaplama süresini artırdığı tespit edilmiştir.



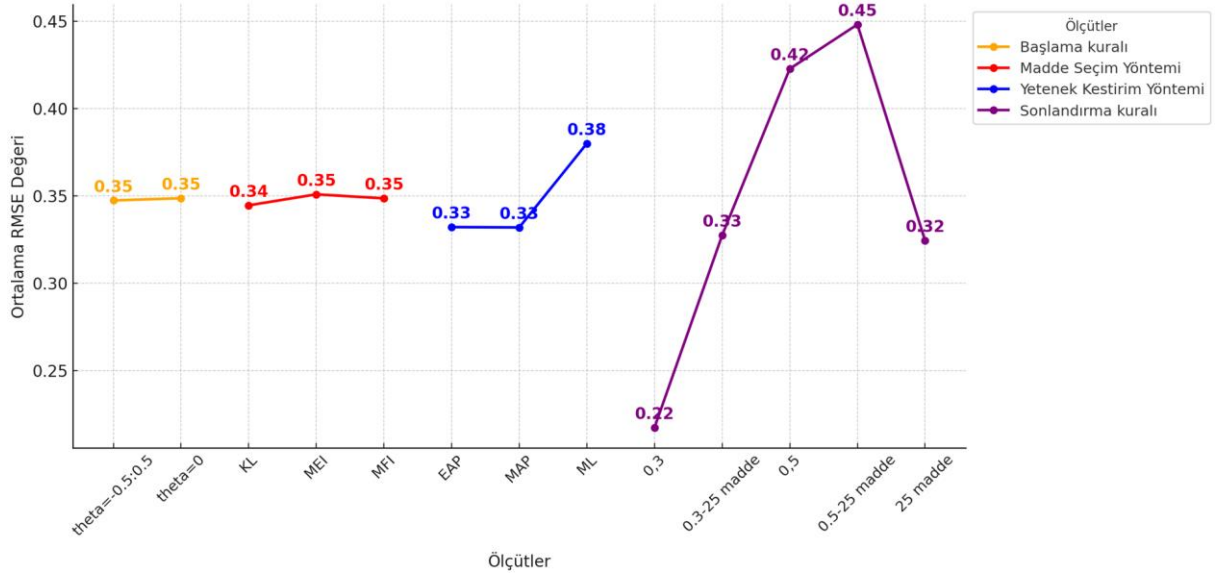
Şekil 4.2. Post-Hoc Simülasyon Koşullarına göre Ortalama Test Uzunlukları

Şekil 4.2'ye göre ortalama test uzunlukları açısından başlama kuralları arasında küçük bir fark bulunsa da bu fark düşük düzeydedir. Madde seçim yöntemleri incelendiğinde, MEI yöntemi diğerlerine göre bir miktar daha kısa test uzunluğu sağlarken MFI ve KL yöntemleri neredeyse eşit test uzunlukları üretmiştir. Yetenek kestirim yöntemlerinde ise MAP ve EAP yöntemleri benzer test uzunluklarına sahipken ML yöntemi diğerlerinden belirgin şekilde daha uzun testler oluşturmuştur. Sonlandırma kurallarına bakıldığında, geniş hata toleransı içeren kurallardan 0,50-25 madde en kısa test uzunluğunu sağlamış, hata toleransı azaldıkça test uzunluğu artmıştır. Özellikle sonlandırma kuralı için standart hatanın 0,30 olması koşulunun test uzunluğunu önemli ölçüde uzattığı belirlenmiştir.



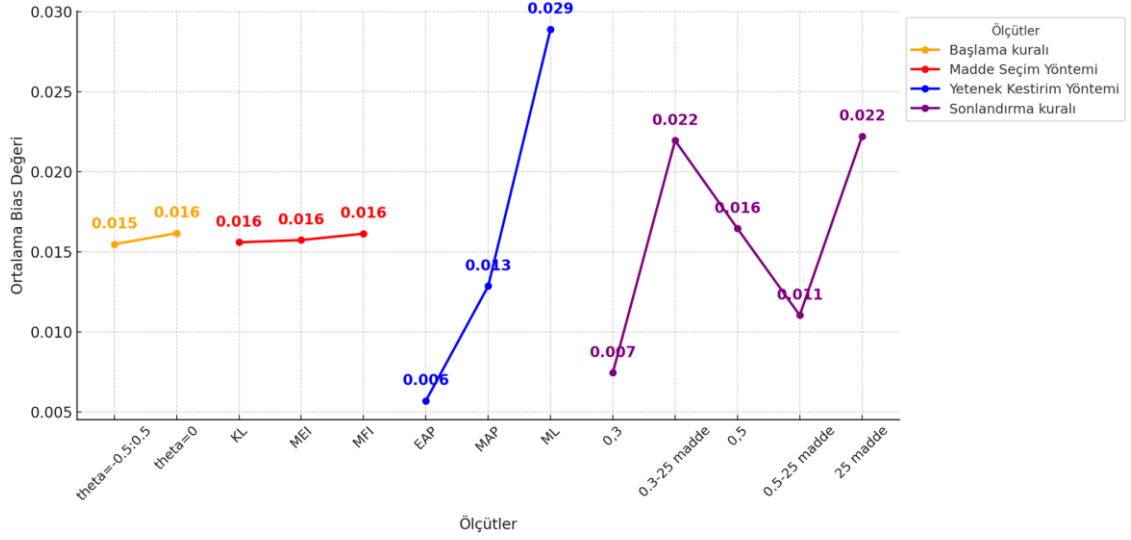
Şekil 4.3. *Post-Hoc Simülasyon Koşullarına göre Gerçek ve Kestirilen Yetenek Arasındaki Ortalama Korelasyon Değerleri*

Şekil 4.3'te başlama kuralının $\theta = 0$ ve $\theta = (-0,5, 0,5)$ aralıkları seçiminden elde edilen yetenek kestirimleri ile gerçek yetenek sonuçlarının her iki durum için de aynı derecede yüksek korelasyon değerleri ürettiği görülmektedir. Bu durum başlangıç noktasının test sonuçları üzerindeki etkisinin minimal düzeyde olduğunu göstermektedir. Madde seçim yöntemleri için de benzer bir durum söz konusu olup her üç yöntemden elde edilen yetenek kestirimleri ile gerçek yetenek sonuçlarına ait korelasyon değerleri yüksek ve birbirine eşit değerdedir. Bu bulgu farklı madde seçim yöntemlerinin yetenek kestirim doğruluğunda aynı derecede etkili olduğu sonucunu ortaya koymaktadır. Yetenek kestirim yöntemlerinin yetenek kestirimleri ile gerçek yetenek puanları arasındaki korelasyonların birbirine eşit ve çok benzer olduğu görülmektedir. Başlama kuralı, madde seçim yöntemi ve sonlandırma kuralları ilgili korelasyon değerlerinde benzer performans gösterirken sonlandırma kurallarında ise daha hassas kuralların korelasyonu artırdığı, özellikle 0,30 tolerans seviyesinin en yüksek korelasyon değerine ulaştığı görülmüştür.



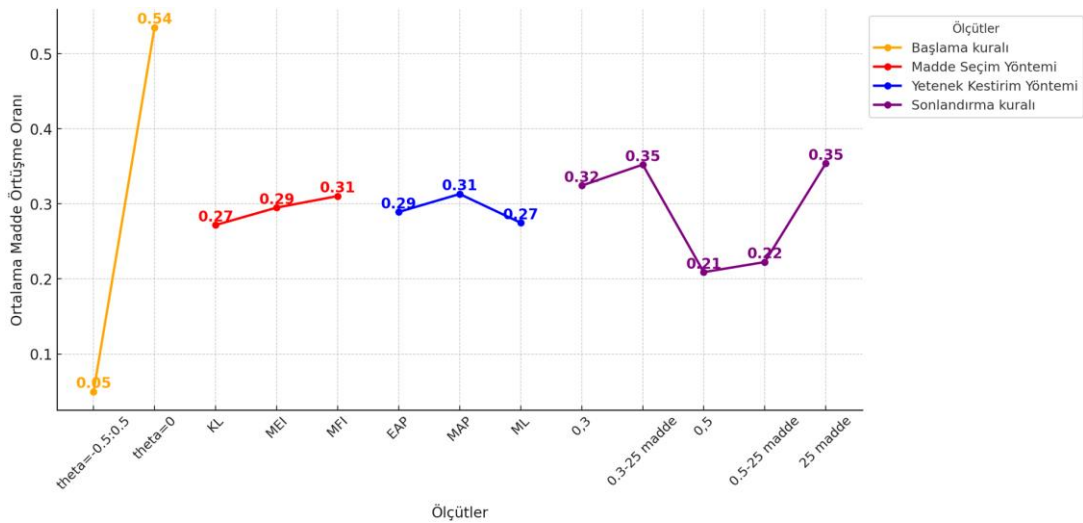
Şekil 4.4. Post-hoc Simülasyon Koşullarına Göre Ortalama RMSE Değerleri

Şekil 4.4'teki bulgular başlama kuralları arasında RMSE değerlerinin eşit olduğu ve doğruluk açısından herhangi bir fark yaratmadığı görülmüştür. Madde seçim yöntemlerinde, KL yöntemi diğerlerinden biraz daha düşük RMSE değeri sağlayarak doğruluk açısından bir avantaj sunmuş, MFI ve MEI yöntemleri ise eşit doğruluk performansı göstermiştir. Yetenek kestirim yöntemlerinde MAP ve EAP yöntemlerinin benzer ve düşük RMSE değerleri sağladığı, ML yönteminin ise daha yüksek bir RMSE değerine sahip olduğu görülmüştür, bu da ML yönteminin tahminlerinde biraz daha az kesinlik sunduğunu göstermektedir. Sonlandırma kurallarına göre, en düşük RMSE değerine 0,30 kriteri ile ulaşılmış, bu da daha hassas bir sonlandırma kuralının doğruluk açısından avantaj sağladığını göstermiştir. Geniş toleranslı sonlandırma kurallarında (örneğin, 0,50-25 madde) RMSE değerinin daha yüksek olduğu ve doğruluğun düştüğü gözlemlenmiştir.



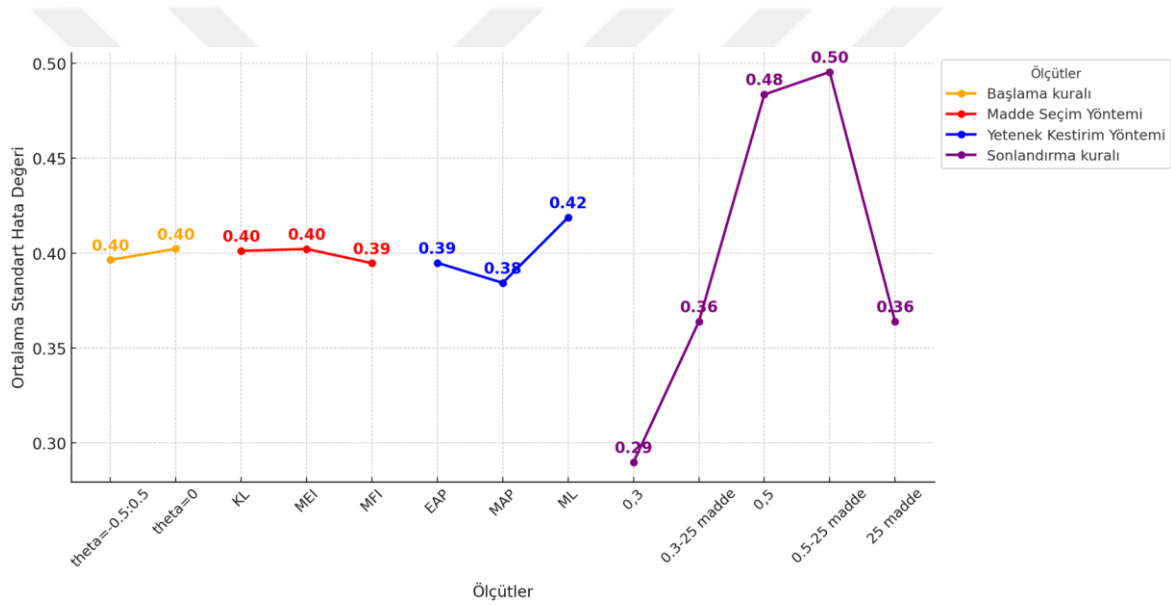
Şekil 4.5. Post-Hoc Simülasyon Koşullarına göre Ortalama Bias Değerleri

Şekil 4.5'te verilen bulgular, başlama kuralları arasında bias değerlerinin çok yakın sonuçlar verdiğini ve pratik açıdan aralarında herhangi bir fark oluşmadığını göstermektedir. Madde seçim yöntemleri arasında KL, MEI ve MFI yöntemleri benzer bias değerleri sağlamış ve bu yöntemlerin doğruluk açısından eşit seviyede performans gösterdiği tespit edilmiştir. Yetenek kestirim yöntemlerinde EAP yöntemi en düşük bias değerine sahipken MAP yöntemi biraz daha yüksek ve ML yöntemi en yüksek bias değeriyle diğer yöntemlerden belirgin şekilde ayrılmıştır. Sonlandırma kurallarında 0,30 kriteri en düşük bias değerini sağlamış ve doğruluk açısından en avantajlı kural olarak öne çıkmıştır. Daha geniş toleranslı kurallar (0,30 ve 25 madde) ise daha yüksek bias değerleri üretmiştir.



Şekil 4.6. Post-hoc Simülasyon Koşullarına göre Ortalama Madde Örtüşme Oranları

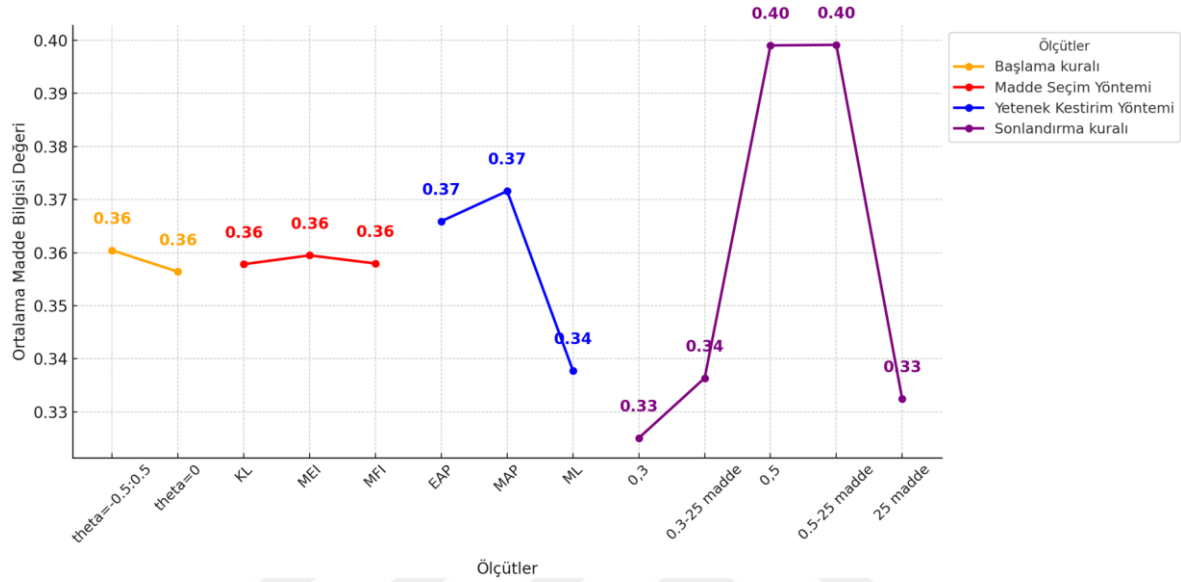
Şekil 4.6'daki başlama kuralı olarak $\theta = 0$ seçiminin, madde örtüşme oranında $\theta = (-0,5, 0,5)$ aralığı seçimine göre belirgin şekilde daha yüksek bir oran sağladığı ve bu başlangıç noktasının madde çeşitliliğini azaltarak daha fazla örtüşmeye neden olduğu görülmektedir. Madde seçim yöntemlerinde MFI yöntemi en yüksek örtüşme oranını üretirken MEI ve KL yöntemleri sırasıyla daha az örtüşme sağlamıştır: Bu durum MFI yönteminin belirli maddeleri daha sık seçtiğine işaret etmektedir. Yetenek kestirim yöntemleri arasında MAP yöntemi en yüksek örtüşme oranını sunmuş, bunu EAP ve ML takip etmiştir; bu durum MAP yönteminin test içeriklerini çeşitlilik açısından sınırlayabileceğini göstermektedir. Sonlandırma kurallarında, daha geniş hata toleransı içeren kuralların daha düşük örtüşme oranları sağladığı, hassas kuralların ise örtüşme oranını artırdığı görülmüştür.



Şekil 4.7. Post-Hoc Simülasyon Koşullarına göre Ortalama Standart Hata Değerleri

Başlama kuralları arasında standart hata değerleri eşit bulunmuş ve başlangıç kuralının hata değerleri üzerinde bir etkisi olmadığı görülmüştür. Madde seçim yöntemleri incelendiğinde, MFI yöntemi diğer yöntemlere göre daha düşük bir standart hata değeri verdiği, KL ve MEI yöntemlerinin ise eşit hata değerleriyle performans açısından aynı sonuçlar ürettiği belirlenmiştir. Yetenek kestirim yöntemlerinden MAP en düşük standart hata değerine sahipken EAP biraz daha yüksek hata değeri vermiştir. ML yöntemi ise en yüksek hata değeriyle diğer yetenek kestirim yöntemlerinden ayrılmıştır. Sonlandırma kurallarında ise 0,30 kriteri en düşük standart hata değerini sağlamış ve en kesin sonuçlar bu kural ile elde edilmiştir. Daha geniş toleranslı sonlandırma kurallarında (örneğin, 0,50 ve 0,50-25 madde) hata değerlerinin belirgin

şekilde arttığı gözlemlenmiştir. Ancak 0,30 ve 0,50 sonlandırma kriterlerindeki hata değerlerinin, araştırma kapsamında sınırlandırılmış hata toleransları olmaları nedeniyle belirgin farklılıklar gösterdiği söylenebilir.

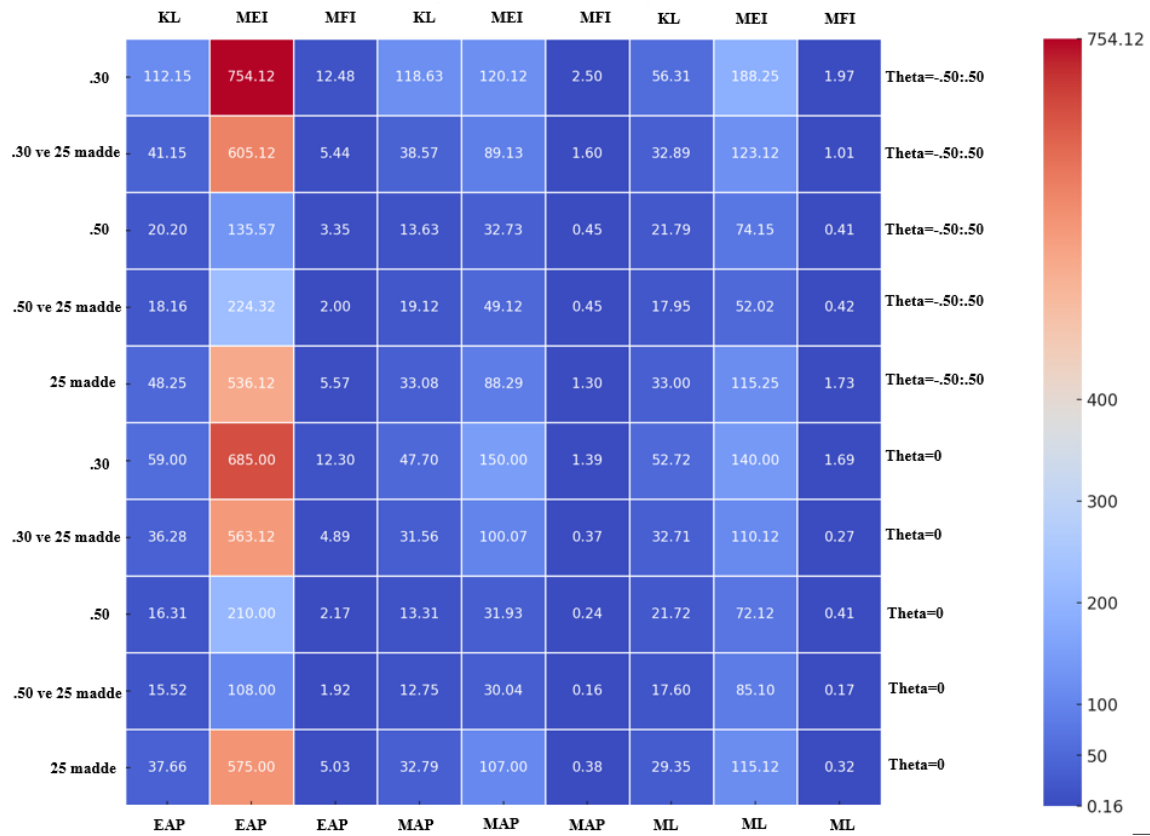


Şekil 4.8. Post-Hoc Simülasyon Koşullarına göre Ortalama Madde Bilgisi Değerleri

Başlama kurallarında arasında ortalama madde bilgisi değerlerinde herhangi bir fark olmadığı görülmüştür. Bu, başlangıç kuralının seçilen maddelerin bilgi değerleri üzerinde bir etkisinin olmadığını göstermektedir. Madde seçim yöntemleri incelendiğinde KL, MFI ve MEI yöntemleri eşit düzeyde madde bilgisi sağlamış, bu da tüm yöntemlerin bilgi açısından dengeli bir performans sunduğunu ortaya koymuştur. Yetenek kestirim yöntemleri arasında MAP ve EAP yöntemleri daha yüksek madde bilgisi değerleri üretirken ML yöntemi diğerlerine göre daha düşük bilgi değeri sağlamıştır. Sonlandırma kurallarına bakıldığında, 0,30 ve 25 madde kriterleri daha düşük madde bilgisi değerlerine sahipken, 0,50 ve 0,50-25 madde kriterleri en yüksek bilgi değerlerini üretmiştir. Bu durum, sonlandırma kurallarının bilgi değerleri üzerindeki etkisinin hassasiyet ve test uzunluğu ile ilişkili olduğunu göstermektedir.

4.1.2. Post-hoc simülasyon tüm koşulların beraber değerlendirilmesine yönelik bulgular

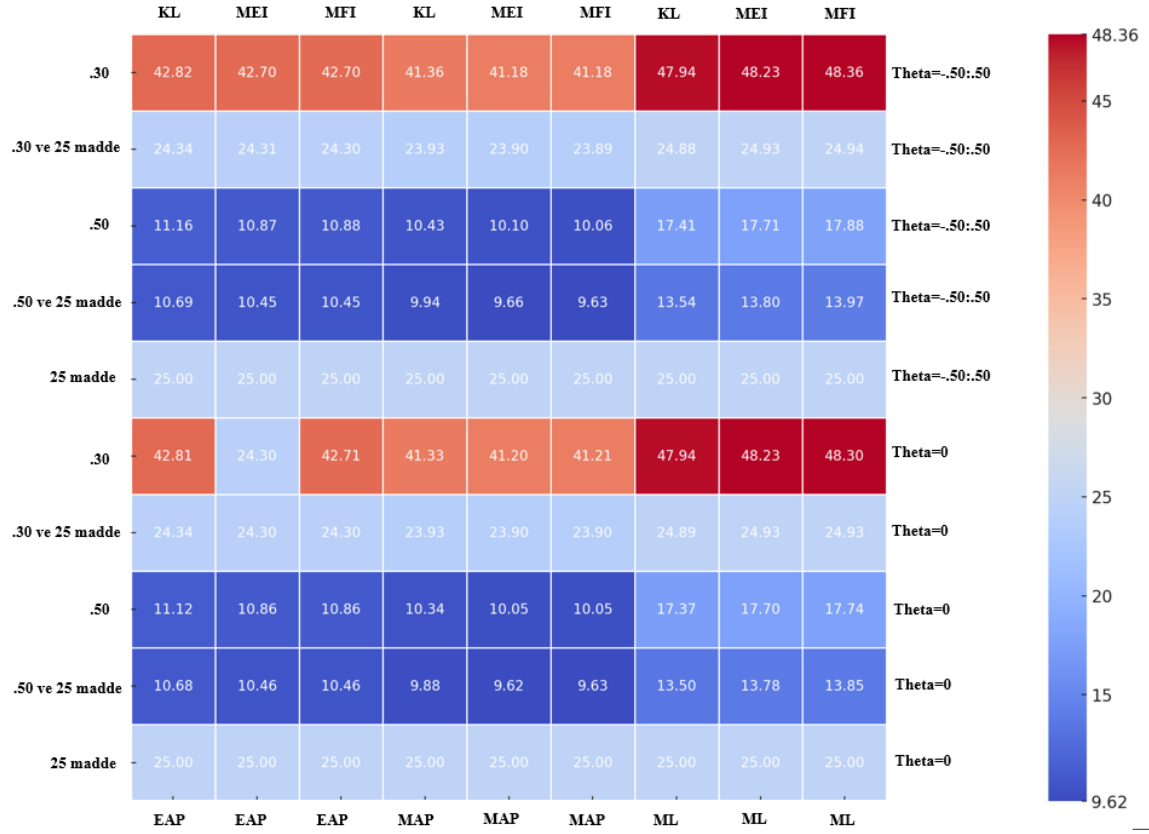
Post-hoc simülasyon kapsamında başlama kuralı, madde seçim yöntemi ve sonlandırma kurallarının farklı yetenek kestirim yöntemleriyle kombinasyonlarında elde edilen sonuçlar Şekil 4.9-Şekil 4.16 arasında verilmiştir.



Şekil 4.9. Yetenek Kestirim Yöntemlerinin Simülasyon Koşullarında Simülasyon Süreleri

Simülasyon süresi kapsamında yapılan analizlerde EAP, MAP ve ML yetenek kestirim yöntemleri arasında belirgin farklılıklar gözlemlenmiştir. MAP yöntemi, simülasyon süresi açısından en hızlı sonuçları sunarak diğer yöntemlere göre avantaj sağlamıştır. Özellikle başlama kuralı $\theta = 0$, madde seçim yöntemi MFI, sonlandırma kuralı 0,50-25 madde kombinasyonunda ($\theta = 0$ / MFI / 0,50-25 madde) MAP yöntemi en kısa sürede tamamlanmıştır. Ayrıca genel olarak MFI madde seçim yönteminin kullanıldığı durumlarda simülasyon sürelerinin ciddi şekilde kısaldığı görülmüştür. Bunun aksine MEI yönteminde sürelerin belirgin bir şekilde arttığı gözlenmiştir, bu da daha MEI'nın karmaşık hesaplama süreçlerinden kaynaklanmaktadır. ML yöntemi ise MAP'a çok yakın bir performans sergileyerek ikinci en hızlı yöntemi temsil etmektedir. EAP yöntemi ise diğer yöntemlere göre

daha uzun simülasyon sürelerine sahip olmasına rağmen $\theta=0$ /MFI/0,50-25 madde kullanıldığında en iyi sonuçlarını vermiştir.

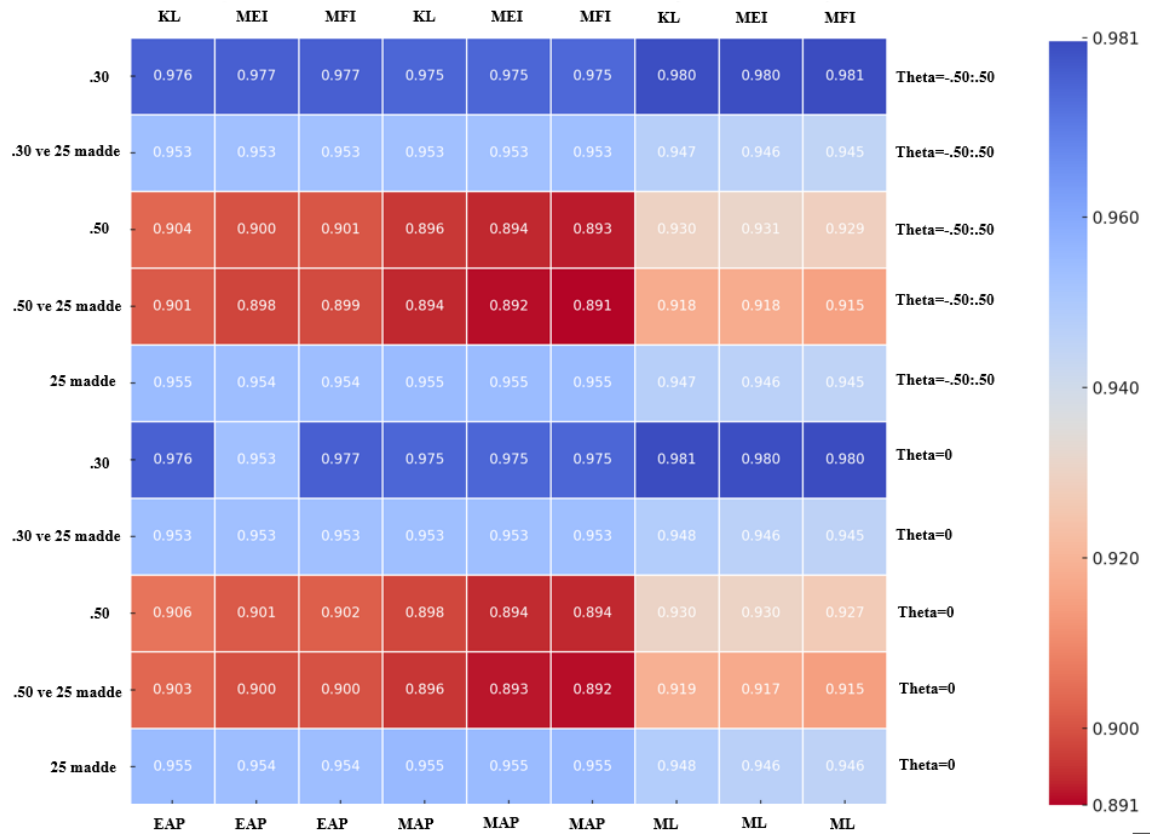


Şekil 4.10. Yetenek Kestirim Yöntemlerinin Simülasyon Koşullarında Ortalama Test Uzunlukları

Ortalama test uzunluğu kapsamında yapılan analizlerde özellikle standart hata 0,30 ölçütünün tüm yetenek kestirim senaryolarında madde sayısının ciddi bir şekilde arttırdığı görülmektedir. Bu durum, standart hata kriterinin daha düşük değerlere ulaşmak için daha fazla madde kullanılmasını gerektirmesi ile açıklanabilir. Bunun aksine 0,50'nin tek başına veya sabit sonlandırma kuralıyla beraber kullanıldığı durumlarda ortalama test uzunluğu tüm yetenek kestirim yöntemlerinde belirgin bir şekilde azalmaktadır. Bu, daha gevşek standart hata kriterlerinin testin adaptif süreçlerinde daha erken sonlanmasını sağladığını göstermektedir.

MAP yöntemi, ortalama test uzunluğu açısından en dengeli sonuçları sunmuştur. $\theta=0$ /MFI/0,50-25 madde sonlandırma kuralı kombinasyonunda MAP yöntemi, test uzunluğunu 9,63 madde ile en kısa sürede tamamlamıştır. ML yöntemi de benzer şekilde kısa test uzunlukları sunarak 9,88 madde ile MAP'a yakın bir performans göstermiştir. ML'nin,

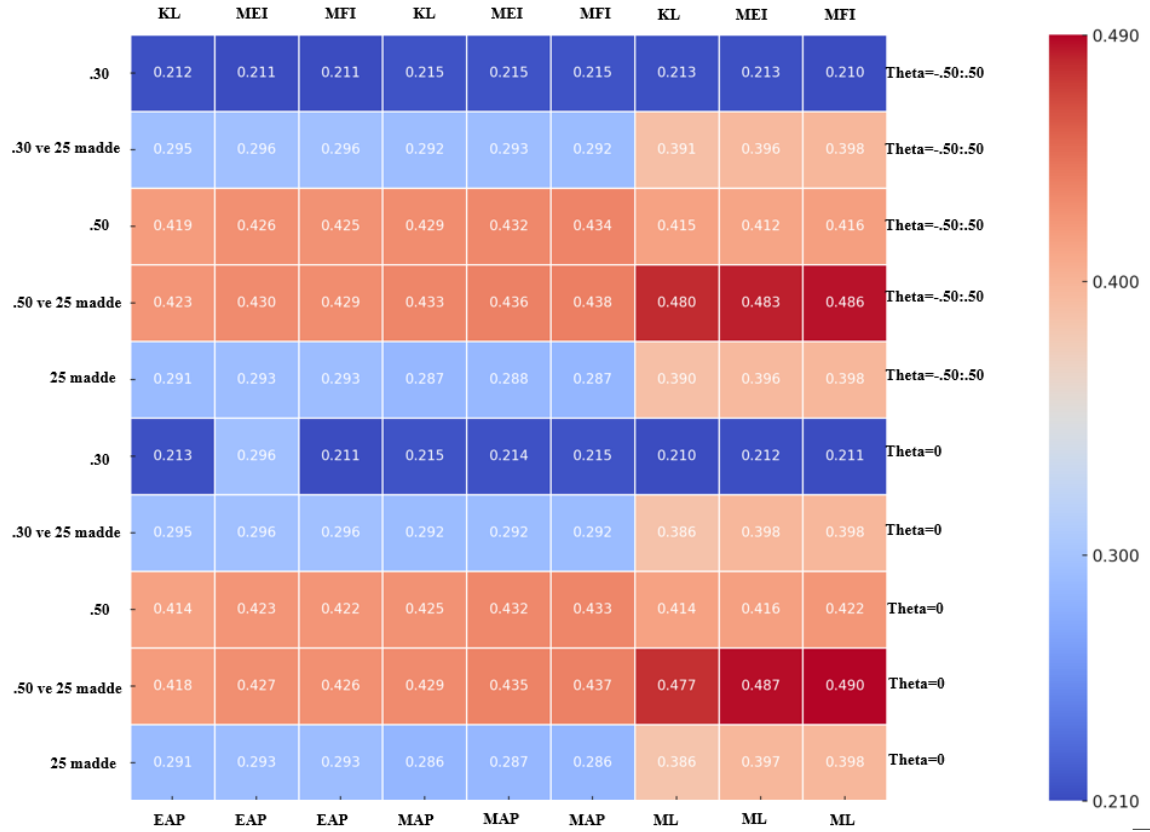
standart hata tabanlı sonlandırma kurallarıyla birlikte kullanıldığında bu avantajı sağladığı görülmektedir. EAP yöntemi ise, ortalama test uzunluğu açısından diğer yöntemlere göre daha yüksek değerler göstermiştir. Ancak $\theta=0$ /MFI/0,50-25 madde sonlandırma kuralı kombinasyonunda EAP yöntemi 10,45 madde ile daha kısa test uzunluğu sunarak kabul edilebilir bir performans sergilemiştir. Bunun dışında MEI yönteminin EAP ile kullanıldığı durumlarda test uzunluğunun belirgin şekilde arttığı gözlenmiştir.



Şekil 4.11. Yetenek Kestirim Yöntemlerinin Simülasyon Koşullarında Gerçek ve Kestirilen Yetenek Arasındaki Korelasyon Değerleri

Gerçek yetenek ile kestirilen yetenek arasındaki korelasyon için yapılan analizlerde MAP ve ML yetenek kestirim yöntemleri, genel olarak daha yüksek korelasyon değerleri sunmuştur. $\theta=0$ /MFI//25 madde kombinasyonunda MAP yöntemi 0,96 gibi yüksek bir korelasyon değeri ile en iyi sonucu vermiştir. ML yöntemi de benzer bir performans sergileyerek 0,95 korelasyon değeri ile MAP'a yakın sonuçlar sunmuştur. EAP yöntemi, genellikle yüksek korelasyon değerleri sunmasına rağmen MEI madde seçim yöntemi ile kullanıldığında korelasyonun düştüğü senaryolar gözlemlenmiştir.

Sonlandırma kuralının etkisi incelendiğinde, standart hata 0,30 gibi daha dar toleranslı kuralların kullanıldığı durumlarda diğer koşullar fark etmeksizin korelasyon değerlerinin 0,98 gibi en yüksek seviyelere ulaştığı görülmektedir. Bunun aksine standart hata 0,50 veya 0,50-25 madde sonlandırma kurallarının devreye girdiği durumlarda korelasyon değerleri düşüş gösterse de yine de bu değerlerin yüksek olduğu söylenebilir.



Şekil 4.12. Yetenek Kestirim Yöntemlerinin Simülasyon Koşullarında RMSE Değerleri

Gerçek yetenek ile kestirilen yetenek arasındaki hata değerleri (RMSE) sonuçlarına göre MAP yöntemi, genel olarak en düşük RMSE değerlerini sunmuş ve en doğru yetenek kestirimini sağlamıştır. theta=0/MFI/0,30 standart hata sonlandırma kuralı kombinasyonunda MAP yöntemi 0,21 RMSE değeri ile en iyi sonucu vermiştir. ML yöntemi de düşük RMSE değerleri ile dikkat çekmektedir. Özellikle sabit 25 madde sonlandırma kuralı ile 0,29 RMSE değeri elde edilmiştir. EAP yöntemi ise diğer yetenek kestirim yöntemlerine göre daha yüksek RMSE değerleri sunmuş ve özellikle madde seçim yöntemi olarak MEI kullanıldığı durumda bu hata oranlarının arttığı gözlenmiştir.

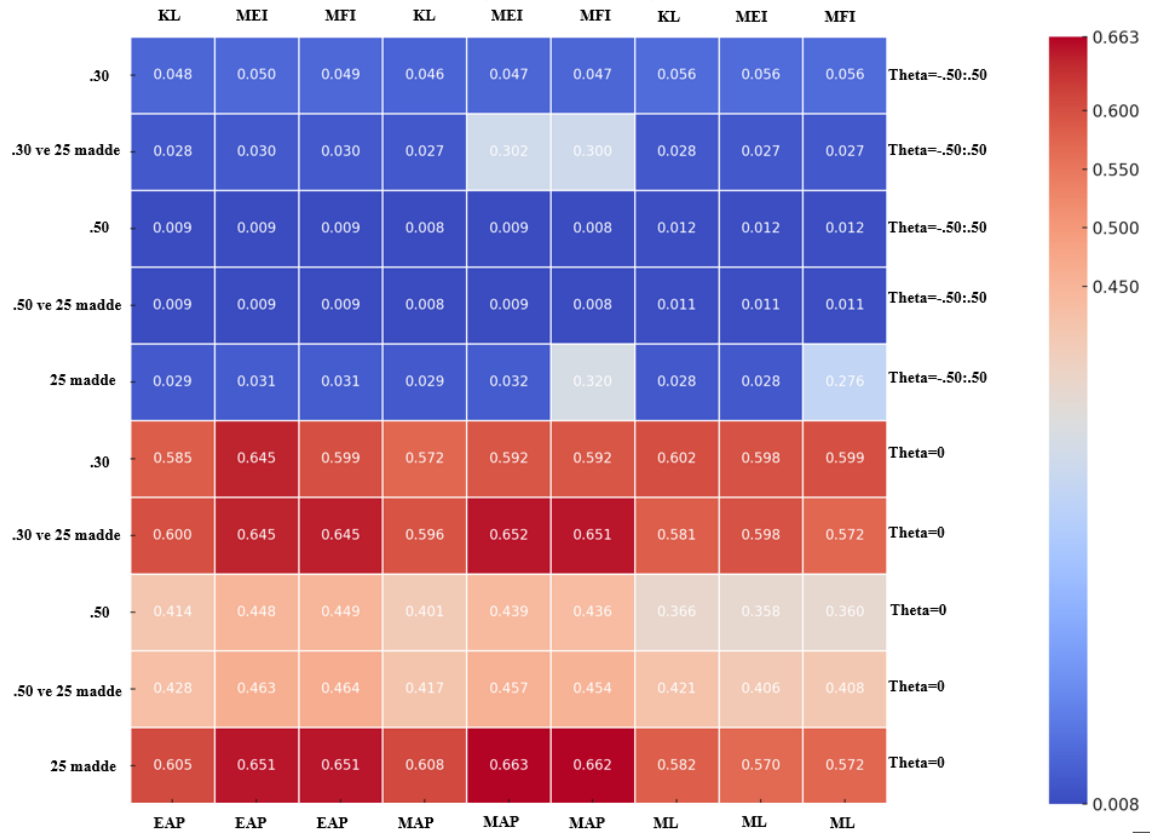
Sonlandırma kurallarının etkisi incelendiğinde, standart hata 0,30 gibi daha dar toleranslı kuralların kullanıldığı durumlarda diğer koşullar fark etmeksizin RMSE değerlerinin en düşük seviyelerde olduğu görülmüştür. Bu, daha sıkı sonlandırma kriterlerinin test doğruluğunu artırdığını ortaya koymaktadır. Ancak standart hata 0,50 veya 0,50-25 madde sonlandırma kurallarında, RMSE değerlerinde artış gözlenmiştir. Bu durum, gevşek kriterlerin doğruluk üzerinde olumsuz bir etkisi olduğunu işaret etmektedir.



Şekil 4.13. Yetenek Kestirim Yöntemlerinin Simülasyon Koşullarında Bias Değerleri

Yetenek kestiriminin sistematik hata içerip içermediğine ve bu hataların hangi koşullarda arttığına (bias) yönelik sonuçlar incelendiğinde MAP ve EAP yetenek kestirim yöntemleri, genel olarak en düşük BIAS değerlerini sunmuş ve en az sistematik hatayı üretmiştir. Özellikle theta=0/MFI/0,30 standart hata sonlandırma kuralı kombinasyonunda MAP yöntemi, bias değerini en düşük seviyeye indirmiştir. EAP yöntemi de benzer şekilde düşük BIAS değerleri sunmuş, aynı koşullarda 0,00-0,01 değer aralığında sonuçlar vermiştir. ML yöntemi ise diğer yöntemlere kıyasla daha yüksek bias değerleri üretmiş ve bazı senaryolarda 0,05 seviyelerine ulaşmıştır. Özellikle madde seçim yöntemi olarak MEI kullanıldığı koşulda, ML'nin BIAS değerlerinin arttığı gözlenmiştir.

Sonlandırma kurallarının etkisi incelendiğinde, standart hata 0,30 gibi daha hassas kriterlerin kullanıldığı durumlarda diğer koşullar fark etmeksizin bias değerlerinin en düşük seviyelere indiği açıkça görülmektedir. Bunun aksine, standart hata 0,50 veya 0,50-25 madde sonlandırma kuralları devreye girdiğinde bias değerlerinde artış gözlenmiştir. Bu durum, daha gevşek kriterlerin sistematik hataları artırabileceğini göstermektedir.

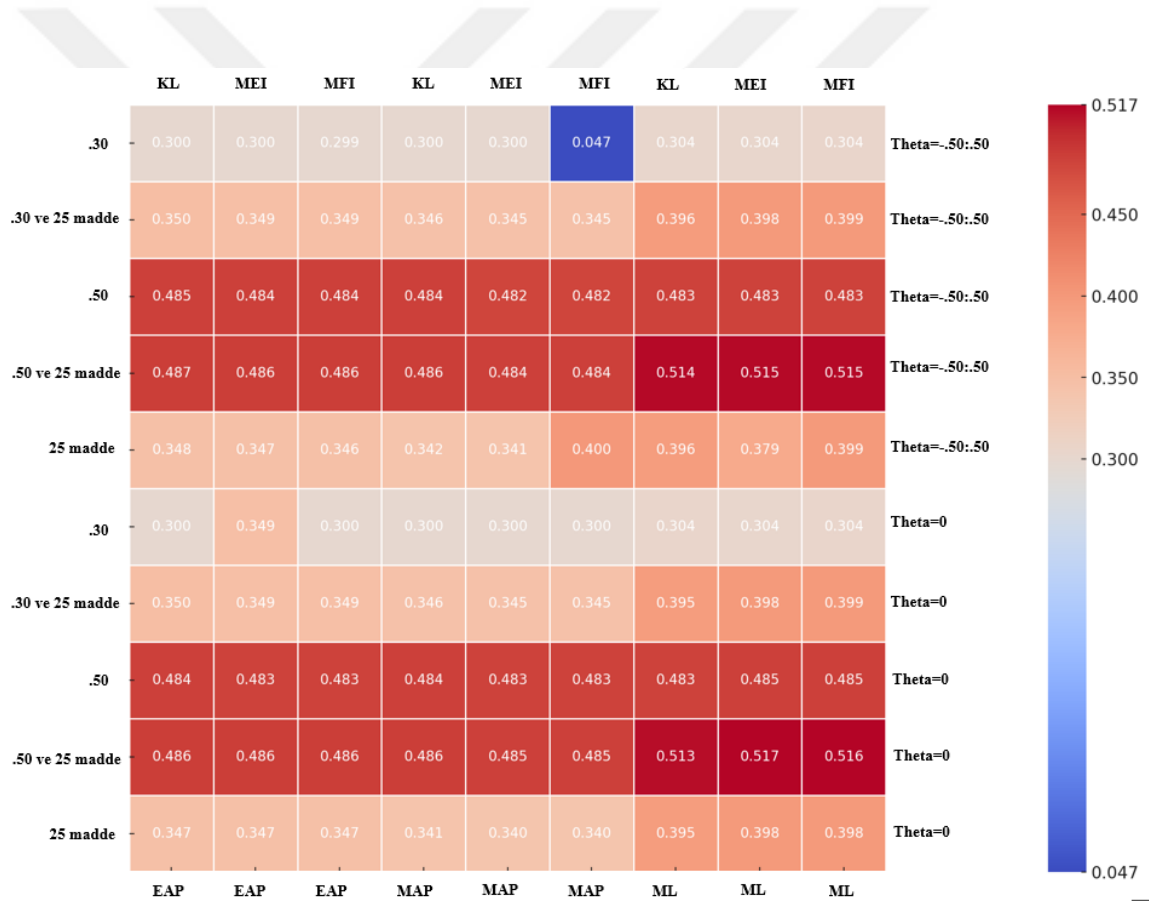


Şekil 4.14. Yetenek Kestirim Yöntemlerinin Simülasyon Koşullarında Madde Örtüşme Oranları

Madde örtüşme oranı sonuçlarına göre yetenek kestirim yöntemleri olarak MAP ve ML tercih edildiğinde genel olarak daha düşük madde örtüşme oranlarına ulaşıldığı görülmektedir. theta=0/MFI/0,30 standart hata sonlandırma kuralı kombinasyonunda MAP yöntemi, madde örtüşme oranını 0,30 ile en düşük seviyeye indirmiştir. Bu durum, MAP'nin adaptif test süreçlerinde daha geniş bir madde havuzu kullandığını ve madde tekrarını azalttığını göstermektedir. EAP yöntemi, genellikle daha yüksek madde örtüşme oranları ile dikkat çekmiş, özellikle MEI madde seçim yöntemi ile kullanıldığında oranların 0,64-0,65 seviyelerine çıktığı gözlenmiştir.

Sonlandırma kurallarının etkisi incelendiğinde, standart hata 0,50 veya 0,50-25 madde gibi daha gevşek kriterlerin kullanıldığı durumlarda diğer koşullar fark etmeksizin madde örtüşme oranlarının daha yüksek seviyelere ulaştığı görülmektedir. Bu durum, daha gevşek kriterlerin testin belirli maddeleri sıkça kullanmasına neden olduğunu göstermektedir. Buna karşın, standart hata 0,30 gibi daha sıkı kriterler kullanıldığında madde örtüşme oranlarının düştüğü gözlemlenmiştir.

Başlama kuralları arasında "theta = 0" başlama kuralı, madde örtüşme oranını genel olarak daha düşük seviyelerde tutmuş, "theta = -0,50:0,50" başlama kuralında ise oranların hafifçe arttığı gözlenmiştir. Bu, geniş bir yetenek aralığından başlama durumunun adaptif süreçte belirli maddelerin daha sık kullanılmasına neden olduğunu göstermektedir.

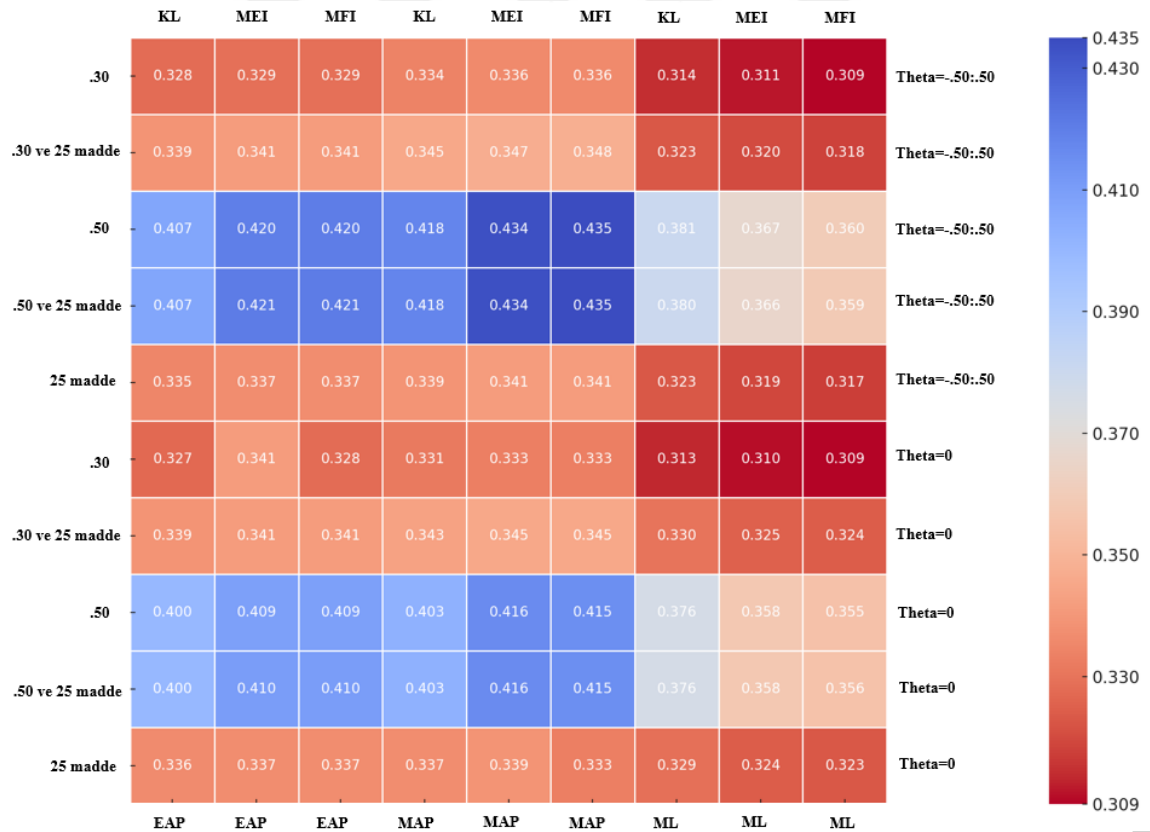


Şekil 4.15. Yetenek Kestirim Yöntemlerinin Simülasyon Koşullarında Ortalama Standart Hata Değerleri

Ortalama standart hata sonuçlarına göre MAP yöntemi, genel olarak en düşük ortalama standart hata değerlerini sunarak en kararlı ve kesin tahminleri sağlamıştır. theta=0/MFI/0,30 standart hata sonlandırma kuralı kombinasyonunda MAP yöntemi, 0,30 ortalama standart hata

değeri ile en iyi sonucu vermiştir. EAP yöntemi, benzer şekilde düşük standart hata değerleri sunmuş ve aynı koşullarda MAP'a yakın sonuçlar sağlamıştır. Ancak MEI madde seçim yöntemi ile kullanıldığında standart hata değerlerinin hafifçe arttığı gözlenmiştir. ML yöntemi ise diğer yöntemlere kıyasla daha yüksek standart hata değerleri sunmuştur. Özellikle MEI madde seçim yöntemi ve gevşek sonlandırma kurallarıyla kullanıldığında standart hata değerleri .52 seviyesine ulaşmıştır. Bu durum, ML'nin belirli senaryolarda daha az kararlı tahminler sağladığını göstermektedir.

Sonlandırma kurallarının etkisi incelendiğinde, standart hata 0,30 gibi daha sıkı toleranslı kuralların kullanıldığı durumlarda diğer koşullar fark etmeksizin ortalama standart hata değerlerinin en düşük seviyelere ulaştığı açıkça görülmektedir. Bunun aksine standart hata 0,50 veya 0,50-25 madde gibi daha gevşek kriterlerin kullanıldığı durumlarda standart hata değerlerinde artış gözlemlenmiştir. Bu, gevşek kriterlerin belirsizlik seviyesini artırarak tahminlerin doğruluğunu düşürebileceğini göstermektedir.



Şekil 4.16. Yetenek Kestirim Yöntemlerinin Simülasyon Koşullarında Ortalama Madde Bilgi Değerleri

Ortalama madde bilgisi sonuçlarına göre MAP yöntemi, genel olarak daha yüksek ortalama madde bilgisi değerleri sunmuştur. $\theta=0/MFI/0,30$ standart hata kombinasyonunda MAP yöntemi, ortalama madde bilgisi değerini 0,43 seviyelerine kadar çıkarmış ve en iyi performansı sağlamıştır. Bu durum, MAP'ın adaptif süreçte en bilgilendirici maddeleri seçmede oldukça etkili olduğunu göstermektedir. EAP yöntemi, MAP'a yakın performans sergileyerek benzer şekilde yüksek madde bilgisi değerleri sağlamıştır. Ancak EAP yöntemi MEI madde seçim yöntemi ile kullanıldığında ortalama madde bilgisi değerleri 0,40 seviyelerinde kalmış, bu da MEI yönteminin bilgi sağlamada daha sınırlı kaldığını göstermiştir. ML yöntemi ise diğer yöntemlere kıyasla daha düşük ortalama madde bilgisi değerleri ile dikkat çekmiştir. Özellikle gevşek sonlandırma kuralları ve MEI madde seçim yöntemi ile kullanıldığında ortalama madde bilgisi değerleri 0,31 seviyesine kadar düşmüştür.

Sonlandırma kurallarının etkisi incelendiğinde, standart hata 0,30 gibi sıkı kriterlerin kullanıldığı durumlarda diğer koşullar fark etmeksizin ortalama madde bilgisinin daha yüksek olduğu gözlemlenmiştir. Bunun aksine, standart hata 0,50 veya 0,50-25 madde gibi gevşek kriterlerin kullanıldığı durumlarda madde bilgisi değerlerinin düştüğü dikkat çekmektedir.

4.1.3. Canlı BOBUT uygulaması koşullarının belirlenmesine yönelik bulgular

Her yetenek kestirim yöntemi için 30 olmak üzere toplam 90 simülasyon senaryosunda ilgili yetenek kestirim yöntemleri için en avantajlı ve en dezavantajlı senaryolar belirlenmiştir. Bu kapsamda her yetenek kestirim yöntemi için ortak olan kabul edilebilir koşullar belirlenmeye çalışılmıştır.

İlk olarak başlama kuralı kapsamında elde edilen sonuçlara görebaşlangıç durumu yetenek kestiriminin doğruluğu veya BOBUT koşullarının avantajı açısından büyük bir farklılaşma gözlenmemiştir. Ancak $\theta = -0,5:0,5$ başlangıç düzeyi tüm yetenek kestirim yöntemlerinin madde örtüşme oranlarını çok ciddi bir şekilde düşürdüğü için canlı BOBUT uygulaması kapsamında tercih edilmiştir.

Madde seçim yöntemlerine ilişkin sonuçlar incelendiğinde ise yine yetenek kestiriminin doğruluğu veya BOBUT koşullarının avantajı açısından madde seçim yöntemleri büyük bir farklılık göstermemekle birlikte MFI yönteminin daha hızlı hesaplama süresi, canlı BOBUT uygulamasında bu yöntemin tercih edilmesine yol açmıştır

Son olarak sonlandırma kuralına ilişkin sonuçlar incelendiğinde temelde 0,30 ve 0,50 hata değerinin baz alındığı durumlarda her ne kadar üç yetenek kestirimi yöntemi yetenek

kestiriminin doğruluğu açısından iyi sonuçlar verse de ortalama test uzunlukları hem öğrenci seviyesi hem de bir BOBUT uygulaması için uygun değerler üretmemiştir. Bu kapsamda hem standart hatanın kullanıldığı hem de sabit sonlandırma kuralının kullanıldığı durumlar her üç yetenek kestirim yöntemi için yetenek kestiriminin doğruluğu veya BOBUT koşulları açısından iyi sonuçlar veren senaryolara odaklanılmıştır. 0,30 ve 25 maddenin birlikte değerlendirildiği durumlar, 0,50 ve 25 maddenin birlikte kullanılma durumlarına göre daha iyi sonuçlar verdiği için canlı BOBUT uygulaması kapsamında tercih edilmiştir. Post-hoc simülasyon kapsamında her üç yetenek kestirimi için belirlenen kurallar ve gösterdikleri performanslar Tablo 4.1’de gösterilmiştir.

Tablo 4.1. *Canlı BOBUT için Seçilen Kriterlerin Post-Hoc Sonuçları*

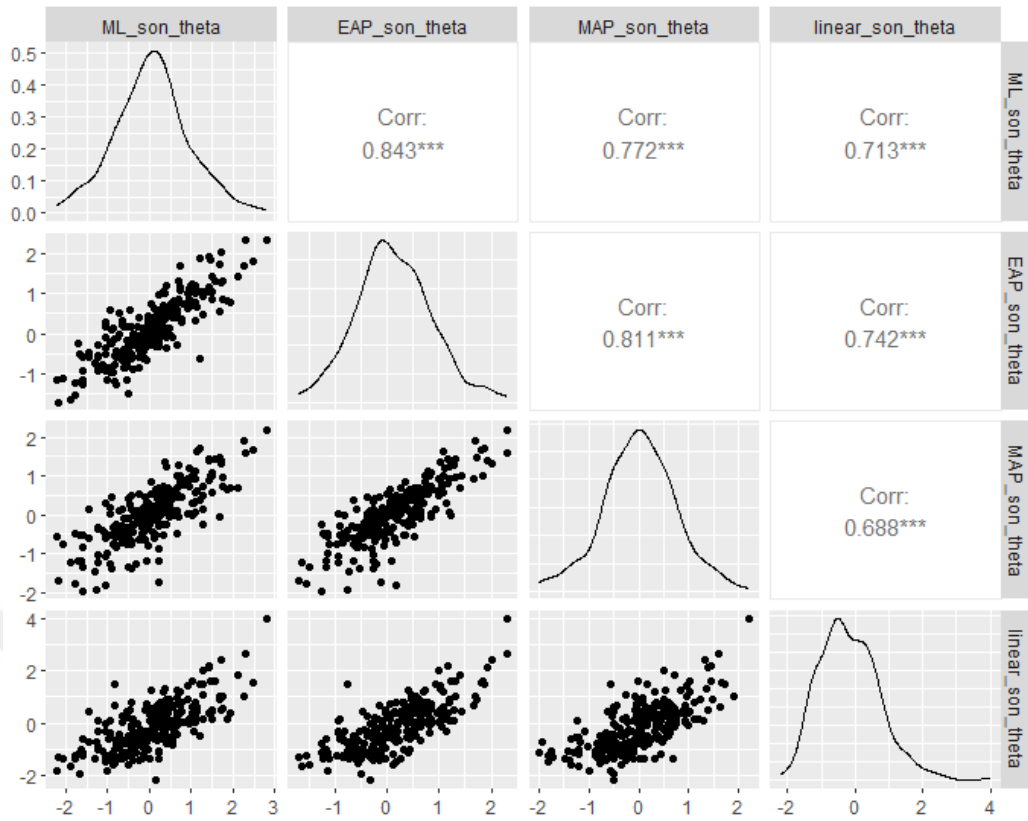
Yetenek Kestirim Yöntemi	ML	MAP	EAP
Simülasyon Süresi (dakika)	1,01	1,60	5,44
Ortalama test uzunluğu	24,94	23,89	24,30
Gerçek ve kestirilen yetenek arası korelasyon	0,944	0,953	0,953
RMSE	0,398	0,292	0,296
Bias	0,043	-0,020	-0,040
Madde Örtüşme Oranı	0,027	0,300	0,030
Ortalama Standart Hata	0,399	0,345	0,349
Ortalama Madde Bilgisi	0,318	0,348	0,341

4.2. Canlı BOBUT ve doğrusal test uygulamalarına yönelik bulgular

İkinci çalışma grubu kapsamında öğrencilere uygulanan üç farklı canlı BOBUT ve doğrusal test uygulamaları sonucu yetenek kestirimleri, güvenilirlik, standart hata, ortalama test uzunluğu, test süresi ve madde kullanım sıklıklarına yönelik bulgulara aşağıdaki başlıklarda yer verilmiştir.

4.2.1. Yetenek kestirimlerine yönelik bulgular

Üç farklı canlı BOBUT ve doğrusal test uygulamaları sonucu öncelikle yetenek düzeylerine yönelik dağılımlar ve korelasyonlar incelenmiş (Şekil 4.17), sonrasında ise yöntemler arasında fark olup olmadığını belirlemek üzere tekrarlı ölçümler varyans analizi gerçekleştirilmiştir (Tablo 4.1).



Şekil 4.17.Yetenek Kestirimleri Arası İlişki Düzeyleri

Şekil 4.17'deki ölçümlere ait dağılımlar incelendiğinde ML yöntemi tahmin değerlerinin genel olarak normal dağılıma yakın olduğu ve ortalama civarında yoğunlaştığı görülürken EAP yönteminin de benzer bir simetrik dağılım sergilediği belirlenmiştir. MAP yöntemine ait tahmin değerlerinin diğer yöntemlere göre daha dar bir aralıkta yoğunlaştığı dikkat çekerken doğrusal test sonuçlarının geniş bir varyansa sahip olduğu gözlemlenmiştir.

Korelasyon analizi bulguları, yöntemler arasında genel olarak pozitif ve anlamlı ilişkiler olduğunu göstermektedir. ML ve EAP yöntemleri arasında yüksek bir korelasyon ($r = 0,843$) bulunurken ML ve MAP yöntemleri arasında bu ilişki biraz daha düşük düzeydedir ($r = 0,772$). ML yöntemi ile doğrusal test arasındaki ilişki ise ($r = 0,713$) adaptif yöntemlere kıyasla daha zayıf bir ilişki olduğunu göstermektedir. EAP ve MAP yöntemleri arasındaki korelasyon ($r = 0,811$) güçlü bir ilişkiyi yansıtırken EAP ile doğrusal test arasındaki ilişki daha düşük düzeyde ($r = 0,742$) olarak ortaya çıkmıştır. MAP ile doğrusal test arasında ise diğer yöntem çiftlerine göre daha zayıf bir ilişki ($r = 0,688$) olduğu gözlenmiştir.

Yöntemler arasındaki çapraz ilişkiler incelendiğinde, ML ile EAP ve ML ile MAP yöntemleri arasındaki ilişkinin oldukça düzenli ve doğrusal bir yapı sergilediği görülmektedir. Bu, tahminlerin tutarlı olduğunu ve yöntemlerin birbirine yakın sonuçlar ürettiğini ifade

etmektedir. Bununla birlikte, doğrusal test ile adaptif yöntemler arasındaki ilişkilerde daha geniş bir saçılma gözlemlenmiştir. Bu durum, doğrusal testin tahmin sonuçlarının adaptif yöntemlere kıyasla daha değişken olduğunu ortaya koymaktadır. Uygulamalar arası yetenek ortalama karşılaştırmasına yönelik tekrarlı ölçümler varyans analizi sonuçları Tablo 4.2’de verilmiştir.

Tablo 4.2. *Uygulamalar Arası Yetenek Düzeyi Ortalama Karşılaştırma Sonuçları*

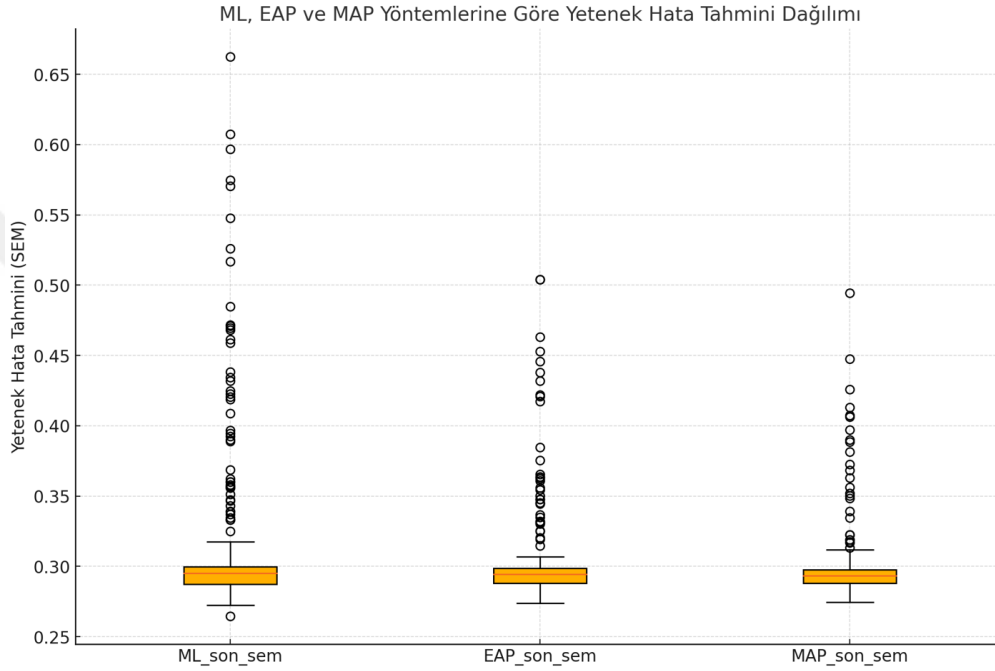
Yöntem	$\bar{\theta}$	S.S	KT	KO	sd	F	p	η^2	Fark
ML	0,052	0,882							
EAP	0,134	0,729			2,29;				EAP > ML,
MAP	0,017	0,724	2,83	1,23	600,16	4,73	0,006	0,018	EAP > MAP,
Doğrusal	0,000	1,000							EAP > Doğrusal

BOBUT yetenek kestirim yöntemlerinin ve doğrusal test uygulaması puanlarının yetenek tahmini üzerindeki etkisini incelemek için tekrarlı ölçümler varyans analizi yapılmıştır. Mauchly'nin Sphericity testi sonuçları, küresellik varsayımının ihlal edildiğini göstermiştir, $W = 0,621$, $\chi^2_{(5)} = 124.005$, $p < 0,001$. Bu nedenle sonuçlar Greenhouse-Geisser düzeltmesi ile raporlanmıştır. Tekrarlı ölçümler ANOVA sonuçlarına göre yöntemler arasında istatistiksel olarak anlamlı bir fark bulunmaktadır [$F(2,29; 600,16) = 4,73$, $p = 0,006$, $\eta^2=0,018$]. Ancak etki büyüklüğünün küçük olduğu ve yöntemler arasındaki farkın toplam varyansın yalnızca %1,8'ini açıkladığı gözlemlenmiştir. Bonferroni düzeltmeli ikili karşılaştırma testleri, hangi yöntemler arasında anlamlı fark olduğunu belirlemek için uygulanmıştır. Buna göre EAP yöntemi ML yöntemine ($\Delta\bar{\theta}=0,82$, $p = 0,033$), MAP yöntemine ($\Delta\bar{\theta} = 0,12$, $p < 0,001$) ve doğrusal test puanlarına ($\Delta\bar{\theta}=0,13$, $p = 0,017$) göre daha yüksek tahminler üretmiştir.

Sonuçlar, EAP yönteminin diğer yöntemlerden daha yüksek yetenek tahminleri ürettiğini göstermektedir. Bu farklılık, EAP'ın tahmin algoritmasının, özellikle de prior bilgi kullanımının, tahmin sonuçlarını etkilediğini göstermektedir. ML ve MAP yöntemleri arasında anlamlı bir fark bulunmaması, bu iki yöntemin yetenek tahminlerinde benzer bir performans sergilediğini ortaya koymaktadır. Doğrusal test puanlarının ML ve MAP ile uyumlu sonuçlar vermesi, doğrusal testin bireye uyarlanmış yöntemlerle karşılaştırılabilir bir yapı sunduğunu göstermektedir. Ancak farkların toplam varyansın yalnızca %1,8'ini açıklaması, bu yöntemler arasındaki pratik farklılıkların minimal ve uygulama bağlamında tüm yöntemlerin kullanılabilir olduğunu düşündürmektedir.

4.2.2. Standart hata değerlerine yönelik bulgular

Üç farklı canlı BOBUT sonucu elde edilen standart hata değerlerinin dağılımları (Şekil 4.18), yetenek düzeyine göre saçılımları (Şekil 4.19) incelenmiştir. Bir sonraki aşamada ise yetenek kestirin yöntemlerinin standart hata değerleri arasında fark olup olmadığını belirlemek üzere tekrarlı ölçümler varyans analizi gerçekleştirilmiştir (Tablo 4.2).

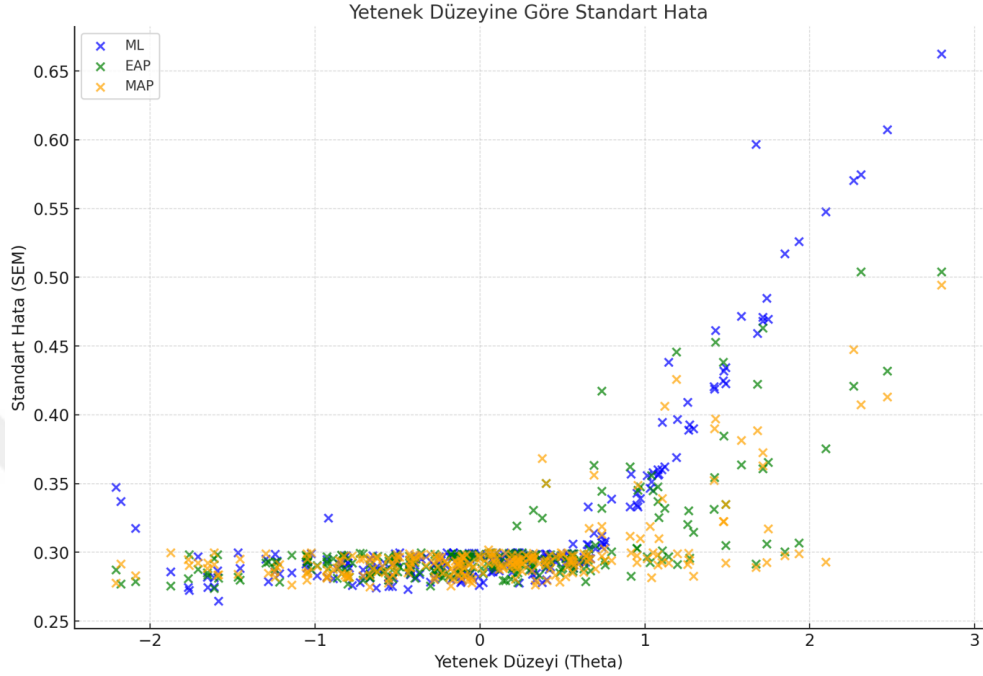


Şekil 4.18. Yetenek Kestirim Yöntemlerinin Standart Hata Dağılımları

Şekil incelendiğinde, EAP ve MAP yöntemlerine ait hata tahminlerinin ortalama olarak en düşük olduğu görülmektedir. Bu durum bu iki yetenek kestirim yönteminin bireylerin yetenek düzeylerini daha kararlı bir şekilde tahmin ettiğini ve ölçüm hatasını minimize edebileceğini göstermektedir. ML yönteminde ise hata tahminlerinin daha yüksek bir seviyede olduğu ve ölçüm hatasında daha fazla değişkenlik bulunduğu söylenebilir.

Grafikte, kutuların genişliği yöntemler arasındaki hata tahminlerinin değişkenliğini yansıtmaktadır. ML yöntemi, diğer yöntemlere kıyasla daha geniş bir dağılıma sahip olup, bu da özellikle düşük yetenek düzeylerinde daha büyük ölçüm hatası ihtimalini işaret etmektedir. Buna karşın diğer yöntemler daha dar bir dağılıma ve düşük aykırı değerlere sahip olmasıyla dikkat çekmektedir. Sonuç olarak daha düşük ölçüm hatası ve tutarlı tahminler ölçütlerinde

EAP ve MAP yöntemi tercih edilebilirken ML yönteminin daha fazla hata ve değişkenlik içerebileceği söylenebilir.



Şekil 4.19. Yetenek Kestirim Yöntemlerinin Yetenek Düzeylerine göre Standart Hata Değerleri

Şekil incelendiğinde genel olarak yetenek düzeyi arttıkça hata miktarının daha çok arttığı ancak orta yetenek düzeyinde ise daha az hataya sahip olduğu görülmektedir. ML yöntemi hem düşük hem de yüksek yetenek düzeylerinde diğer yöntemlere kıyasla daha yüksek hata değerleri sunmaktadır. Bu durum, ML'nin her iki uç noktada da ölçüm doğruluğu açısından daha zayıf bir performans sergilediğini göstermektedir. Özellikle EAP yöntemi, tüm yetenek düzeylerinde daha düşük hata değerleri sunarak en güvenilir yöntemi temsil etmektedir. Düşük ve yüksek yetenek düzeylerinde hata değerlerinin düşük olması, EAP'ın daha tutarlı bir ölçüm sağladığını ve hata tahmininde daha az değişkenlik gösterdiğini ortaya koymaktadır. MAP yöntemi ise ML ve EAP yöntemleri arasında bir performans göstermektedir. Hem düşük hem de yüksek yetenek düzeylerinde hata değerleri ML'ye göre daha düşük olsa da EAP kadar düşük standart hata sunmamaktadır. MAP'ın özellikle orta yetenek düzeylerinde daha dengeli bir performans sergilediği söylenebilir.

Tablo 4.3. Yetenek Kestirim Yöntemlerinin Standart Hata Değerlerinin Karşılaştırma Sonuçları

Yöntem	$\bar{X}$	SS	KT	KO	sd	F	P	η^2	Fark
ML	0,315	0,062							ML>
EAP	0,303	0,035	0,038	0,029	1,29; 338,42	23,86	<0,001	0,08	EAP>
MAP	0,299	0,028							MAP

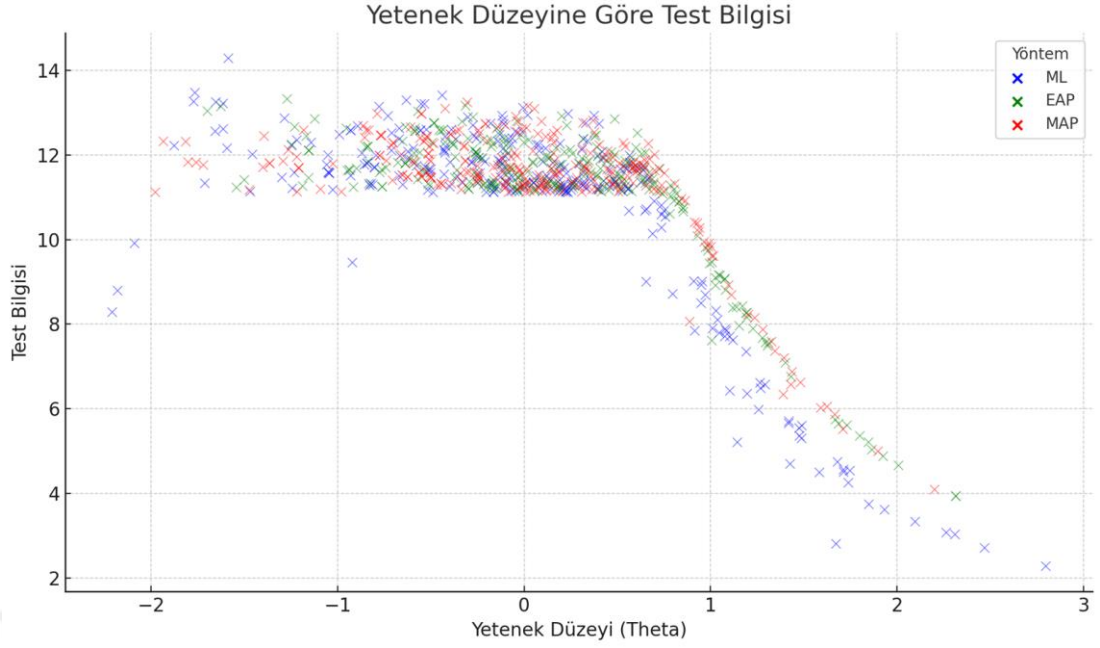
ML, EAP ve MAP kestirim yöntemleri için standart hata miktarlarını karşılaştırmak amacıyla yapılan tekrarlı ölçümler varyans analizi sonuçları küresellik varsayımının ihlal edildiğini göstermiştir, $W = 0,452$, $\chi^2(2) = 207,47$, $p < 0,001$. Bu nedenle analiz sonuçları Greenhouse-Geisser düzeltmesi uygulanarak raporlanmıştır. Tekrarlı ölçümler ANOVA sonuçlarına göre, standart hata değerleri arasında anlamlı farklılık tespit edilmiştir [$F(1,29; 338,42) = 23,86$, $p < 0,001$, $\eta^2 = 0,08$]. Eta kare değeri, kestirim yöntemleri arasındaki standart hata farklarının toplam varyansın %8'ini açıkladığını göstermektedir.

Bonferroni düzeltmeli ikili karşılaştırma testleri sonuçlarına göre ML yöntemi, EAP ($\Delta_{hata} = 0,012$, $p < 0,001$) ve MAP ($\Delta_{hata} = 0,016$, $p < 0,001$) yöntemlerine kıyasla anlamlı derecede daha yüksek standart hata değerine sahiptir. Ayrıca EAP yöntemi, MAP yöntemine kıyasla anlamlı derecede daha yüksek standart hata değerine sahiptir ($\Delta_{hata} = 0,004$, $p = 0,004$).

Sonuç olarak MAP yöntemi, ölçümlerde en düşük standart hata miktarını sağlayarak en kararlı kestirimi sunmaktadır. ML yöntemi ise en yüksek standart hata değerine sahip olup daha fazla değişkenlik göstermiştir. EAP yöntemi, hata miktarı açısından ML ve MAP arasında bir performans sergilemiştir. Bu bulgular, MAP yönteminin ölçümlerde daha hassas sonuçlar sunduğunu, ML yönteminin ise daha geniş bir hata aralığına yol açtığını göstermektedir.

4.2.3. Güvenirlilik değerlerine yönelik bulgular

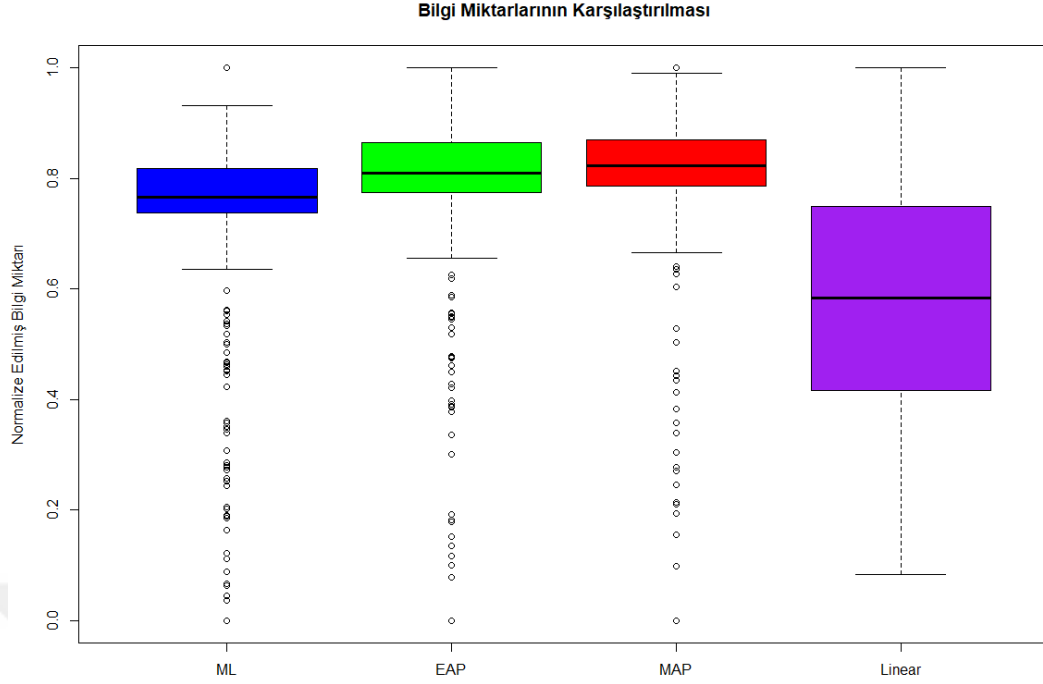
Üç farklı canlı BOBUT ve doğrusal test uygulamaları sonucu elde edilen test bilgi değerlerinin yetenek düzeylerine göre aldıkları değerler (Şekil 4.20) ve dağılımları (Şekil 4.21) incelenmiştir. Bir sonraki aşamada ise elde edilen test bilgi değerleri arasında fark olup olmadığını belirlemek üzere tekrarlı ölçümler varyans analizi gerçekleştirilmiştir (Tablo 4.3).



Şekil 4.20. Yetenek Kestirim Yöntemlerinin Yetenek Düzeylerine göre Test Bilgi Değerleri

Şekil, ML, EAP ve MAP yöntemleri için yetenek düzeyi ile test bilgisi arasındaki ilişkiyi göstermektedir. Tüm yöntemlerde test bilgisinin orta düzey yeteneklerde ($\theta \approx 0$) en yüksek seviyeye ulaştığı, ancak uç yetenek düzeylerinde ($\theta < -2$ veya $\theta > 2$) belirgin bir şekilde azaldığı gözlenmektedir. Bu durum, testlerin en hassas ölçümleri orta düzey yeteneklerde gerçekleştirdiğini ve bu aralıkta güvenilirliğin en yüksek olduğunu göstermektedir. Orta düzey yeteneklerdeki yüksek bilgi miktarı, testlerin genellikle popülasyonun büyük bir kısmını hedeflediği bu bölgede en etkili şekilde çalıştığını ortaya koymaktadır.

Uç yetenek düzeylerinde bilgi miktarının azalması, bu yetenek aralıklarında güvenilirliğin bir miktar düştüğünü göstermektedir. Bu durum, test maddelerinin zorluk düzeylerinin genellikle orta düzey yeteneklere odaklanmasından kaynaklandığı söylenebilir. Özellikle ML yöntemi, uç yetenek düzeylerinde bilgi miktarının en hızlı azaldığı yöntemlerden biri olarak dikkat çekmektedir. EAP yöntemi, daha dengeli bir bilgi dağılımı sunarken ML yöntemi bilgi azalımını en hızlı şekilde göstermiştir. Bu bulgu, ML ve MAP yöntemlerinin uç yetenek düzeylerinde EAP'a kıyasla daha düşük güvenilirlik sağladığını düşündürmektedir.



Şekil 4.21. Uygulamalara Ait Test Bilgi Düzeyleri

Şekil, ML, EAP, MAP ve doğrusal test yöntemleri için normalize edilmiş bilgi miktarlarını göstermektedir. Normalize işlemi, farklı yöntemlerin bilgi miktarlarını aynı ölçekte ifade ederek karşılaştırılabilir hâle getirmiştir. Şekil incelendiğinde ML, EAP ve MAP yöntemleri arasında bilgi miktarlarının oldukça benzer bir dağılım gösterdiği görülmektedir. EAP ve MAP yöntemleri genellikle daha yüksek bilgi miktarı sağlarken ML yöntemi bu iki yönteme göre bilgi miktarının biraz daha düşük yoğunlukta olduğunu göstermektedir. Doğrusal testin normalize edilmiş bilgi miktarı ise diğer yöntemlerden belirgin bir şekilde farklıdır. Doğrusal testte bilgi miktarının geniş bir dağılım sergilediği ve homojen olmadığı dikkat çekmektedir. Bilgi miktarının aralığı geniş olmakla birlikte, ortalama bilgi miktarı diğer yöntemlere kıyasla daha düşük seviyede kalmaktadır. Bu durum, doğrusal testin sabit bir madde yapısına sahip olması nedeniyle farklı yetenek düzeylerindeki bireyler için bilgi sağlama kapasitesinin sınırlı olduğunu göstermektedir.

Sonuç olarak normalize edilmiş bilgi miktarlarına göre bireye uyarlanmış test yöntemlerinin (ML, EAP ve MAP), doğrusal test yönteminden daha dengeli ve yüksek performans sergilediği görülmektedir. Bu yöntemler, bireylerin yetenek düzeylerine uyum sağlayarak ölçüm hassasiyetini artırmaktadır. Öte yandan doğrusal test, sabit madde yapısı nedeniyle bilgi miktarında daha düşük bir performans sergilemektedir.

Tablo 4.4. *Uygulamalar Arası Test Bilgi Düzeylerinin Karşılaştırma Sonuçları*

Yöntem	TI($\bar{\theta}$)	S.S	KT	KO	sd	F	p	η^2	Fark
ML	0,71	0,20							
EAP	0,77	0,18	7,47	4,89	1,53;	76,36	0,006	0,23	MAP>EAP>ML>Doğrusal
MAP	0,80	0,16			400,02				
Doğrusal	0,58	0,20							

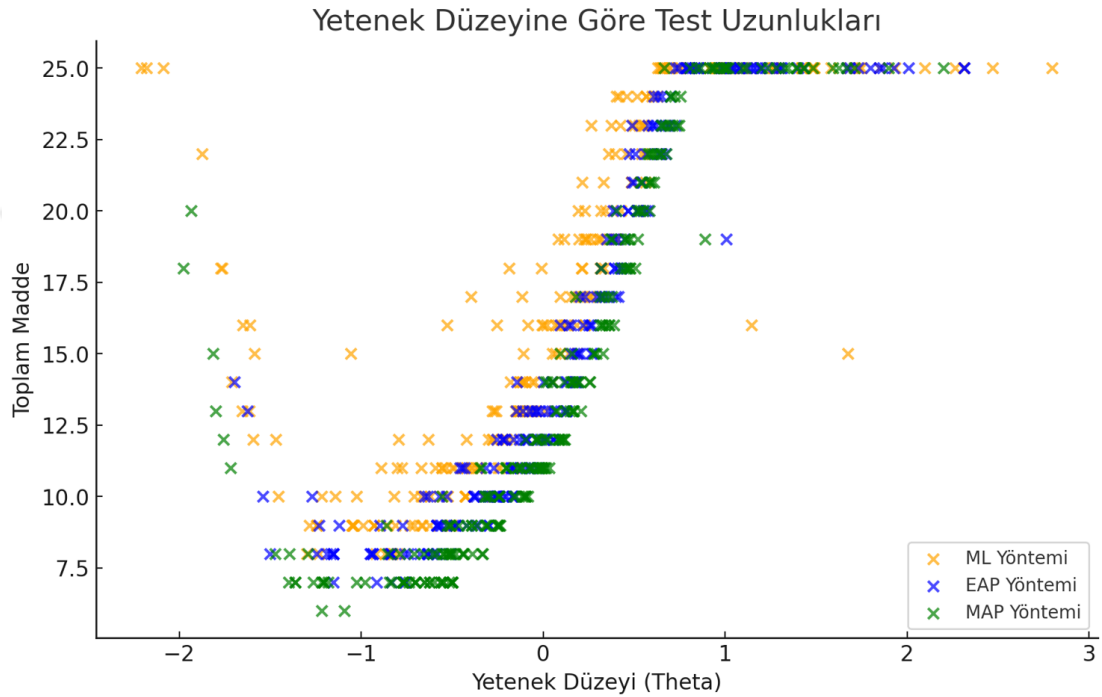
ML, EAP, MAP ve doğrusal test yöntemleri için test bilgi düzeylerini karşılaştırmak amacıyla yapılan tekrarlı ölçümler varyans analizi sonuçları küresellik varsayımının ihlal edildiğini göstermiştir, $W=0,200$, $\chi^2(5) = 419,29$, $p < 0,001$. Bu nedenle analiz sonuçları Greenhouse-Geisser düzeltmesi uygulanarak raporlanmıştır. Tekrarlı ölçümler ANOVA sonuçlarına göre yöntemler arasında test bilgi düzeyleri açısından istatistiksel olarak anlamlı bir fark bulunmuştur, $F(1,53; 400,02) = 76,36$, $p < 0,001$, $\eta^2 = 0,23$). Eta kare değeri, test bilgi düzeylerindeki varyansın %23'ü yöntemler tarafından açıklandığını göstermektedir.

Bonferroni düzeltmeli ikili karşılaştırma testleri sonuçlarına göre, ML yöntemi, EAP ($\Delta_{TI(\bar{\theta})} = -0,064$, $p < 0,001$) ve MAP ($\Delta_{TI(\bar{\theta})} = -0,090$, $p < 0,001$) yöntemlerine göre daha düşük bilgi düzeyine sahiptir. ML yöntemi, doğrusal test yöntemine kıyasla anlamlı derecede daha yüksek bilgi düzeyi sağlamaktadır ($\Delta_{TI(\bar{\theta})} = 0,128$, $p < 0,001$). EAP yöntemi, MAP yöntemine göre anlamlı derecede daha düşük bilgi düzeyi sunmaktadır ($\Delta_{TI(\bar{\theta})} = -0,026$, $p = 0,003$), ancak doğrusal test yöntemine kıyasla daha yüksek bilgi düzeyine sahiptir ($\Delta_{TI(\bar{\theta})} = 0,192$, $p < 0,001$). MAP yöntemi de doğrusal test yöntemine kıyasla anlamlı derecede daha yüksek bilgi düzeyi sağlamaktadır ($\Delta_{TI(\bar{\theta})} = 0,218$, $p < 0,001$).

Sonuçlar, bireye uyarlanmış yöntemlerin (ML, EAP, MAP), doğrusal test yöntemine kıyasla daha yüksek bilgi düzeyi sunduğunu ve yöntemler arasında bilgi düzeyi açısından farklılıklar olduğunu ortaya koymaktadır. Doğrusal test yöntemi, diğer yöntemlere göre daha düşük bilgi düzeyi sağlarken MAP yöntemi en yüksek bilgi düzeyini sunmuştur. EAP yöntemi ise ML ile MAP arasında bir performans sergilemiştir.

4.2.4. Test uzunluklarına yönelik bulgular

Üç farklı canlı BOBUT sonucu elde edilen test uzunluk değerlerinin yetenek düzeyine göre dağılımları (Şekil 4.22) incelenmiştir. Bir sonraki aşamada ise hem canlı BOBUT hem de doğrusal test uygulaması sonucu ortalama test uzunlukları arasında fark olup olmadığını belirlemek üzere tekrarlı ölçümler varyans analizi gerçekleştirilmiştir (Tablo 4.5.).



Şekil 4.22. Yetenek Kestirim Yöntemlerinin Yetenek Düzeylerine göre Test Uzunlukları

Grafikte, farklı yetenek kestirim yöntemlerinin (ML, EAP ve MAP) yetenek düzeyine göre toplam test uzunlukları karşılaştırılmıştır. ML yöntemi, özellikle uç yetenek düzeylerinde (-2'nin altı ve +2'nin üstü) daha fazla maddeye ihtiyaç duymasıyla dikkat çekmektedir. Bu durum, ML yönteminin uç noktalarda yetenek kestirimi için daha fazla bilgiye ihtiyaç duyduğunu göstermektedir. Öte yandan, EAP ve MAP yöntemleri daha dengeli bir test uzunluğu sunmaktadır. Her iki yöntemin merkezî yetenek düzeyleri (-1 ile +1 arası) için oldukça benzer sonuçlar verdiği, test uzunluğu açısından stabil bir seyir izlediği gözlemlenmektedir. Bununla birlikte, MAP yöntemi uç değerlerde EAP'a kıyasla biraz daha az madde kullanmaktadır. Genel olarak, merkezi yetenek düzeylerinde tüm yöntemler benzer bir performans gösterirken, uç noktalarda ML yöntemi daha değişken bir test uzunluğuna sahiptir.

Tablo 4.5. *Uygulamalar Arası Test Uzunluklarının Karşılaştırma Sonuçları*

Yöntem	$\bar{X}$	SS	KT	KO	sd	F	P	η^2	Fark
ML	16,98	5,95							Doğrusal>
EAP	15,60	6,27	18677,52	6225,84	2,53;	424,39	<0,001	0,62	ML>
MAP	14,10	6,21			662,24				EAP>
Doğrusal	25,00	0,00							MAP

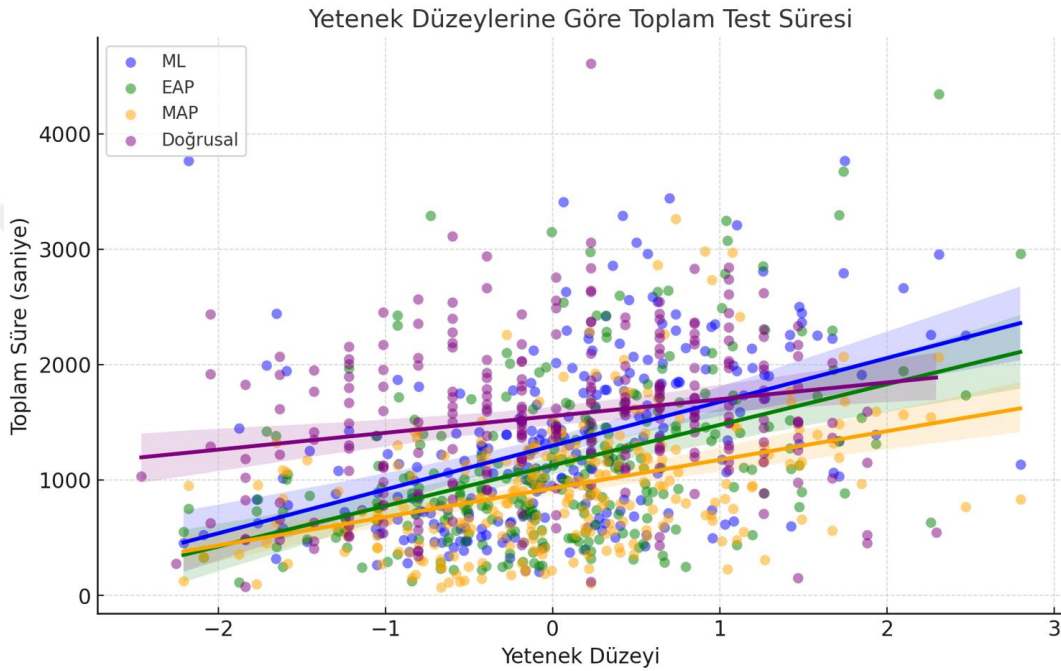
ML, EAP, MAP ve doğrusal test yöntemleri için ortalama madde sayılarını karşılaştırmak amacıyla yapılan tekrarlı ölçümler varyans analizi sonuçları küresellik varsayımının ihlal edildiğini göstermiştir, $W = 0,756$, $\chi^2(5) = 72,90$, $p < 0,001$. Bu nedenle, analiz sonuçları Greenhouse-Geisser düzeltmesi uygulanarak raporlanmıştır. Tekrarlı ölçümler ANOVA sonuçlarına göre yöntemler arasında ortalama madde sayıları açısından istatistiksel olarak anlamlı bir fark bulunmuştur, $F(2,53, 662,24) = 424,39$, $p < 0,001$, $\eta^2 = 0,62$. Eta kare değeri, yöntemlerin ortalama madde sayılarındaki varyansın %62,0'sini açıkladığını göstermekte ve yöntem türünün madde sayısı üzerinde büyük bir etkisi olduğunu ortaya koymaktadır.

Bonferroni düzeltmeli ikili karşılaştırma testleri sonucu doğrusal test yöntemi, diğer tüm yöntemlere kıyasla anlamlı derecede daha yüksek ortalama madde sayısına sahiptir ($p < 0,001$). Bireye uyarlanmış yöntemler arasında ise ML yöntemi, hem EAP yöntemi ($\Delta_{\text{madde sayısı}}=10,38$, $p < 0,001$) hem de MAP yöntemi ($\Delta_{\text{madde sayısı}}=20,89$, $p < 0,001$) ile karşılaştırıldığında anlamlı derecede daha fazla madde ile tamamlanmaktadır. Ayrıca EAP yöntemi, MAP yöntemine kıyasla anlamlı derecede daha yüksek madde sayısına sahiptir ($\Delta_{\text{madde sayısı}}=10,51$, $p < 0,001$).

Sonuçlar, doğrusal test yönteminin sabit bir madde yapısı nedeniyle diğer yöntemlerden belirgin şekilde ayrıldığına, bireye uyarlanmış yöntemlerin (ML, EAP, MAP), bireylerin yetenek düzeyine bağlı olarak değişkenlik gösterdiğine işaret etmektedir. Bu yöntemler arasında ML yöntemi, daha fazla maddeye sahipken MAP yöntemi ise en düşük ortalama madde sayısını sağlamış ve daha az maddeyle yeterli bilgi sağlayarak ölçüm sürecini optimize ettiğini göstermiştir. EAP yöntemi ise bu iki yöntem arasında bir konumda yer alarak madde sayısı açısından dengeli bir yapı sunmuştur.

4.2.5. Test süresine yönelik bulgular

Üç farklı canlı BOBUT ve doğrusal test uygulamaları sonucu toplam test sürelerinin yetenek düzeylerine göre dağılımları (Şekil 4.23) incelenmiştir. Bir sonraki aşamada ise toplam test süreleri arasında fark olup olmadığını belirlemek üzere tekrarlı ölçümler varyans analizi gerçekleştirilmiştir (Tablo 4.6.).



Şekil 4.23. Yetenek Kestirim Yöntemlerinin Yetenek Düzeylerine Göre Toplam Test Süreleri

Şekil 4.23'te yetenek düzeyine göre toplam test süresinin farklı yetenek kestirim yöntemleri (ML, EAP, MAP) açısından karşılaştırması yapılmıştır. Her bir yöntemin ortalama eğilimi ve dağılımı ayrı ayrı incelenmiştir. ML yöntemi, yetenek düzeyleri arasında artan bir toplam süre trendi sergilemektedir. ML yöntemi özellikle yüksek yetenek düzeylerinde (2 ve üzeri) diğer yöntemlere kıyasla daha uzun süreler gerektirebilmektedir. Bu, yöntemin kestirim doğruluğu için daha fazla bilgiye ihtiyaç duyduğunu gösterebilir. EAP yöntemi dengeli ve nispeten daha kısa sürelerde test sonuçları üretmiştir. Özellikle düşük ve orta yetenek düzeylerinde (0 civarı) daha istikrarlı bir performans sergileyerek diğer yöntemlere yakın bir seyir izlemiştir. Yüksek yetenek düzeylerinde sürenin nispeten daha yavaş arttığı görülmektedir. MAP yöntemi toplam test süresi açısından en düşük değerlere sahiptir. Bu durum, MAP yönteminin kestirim için daha az süreye ihtiyaç duyduğunu ve daha hızlı sonuçlar

verdiğini göstermektedir. Yetenek düzeyleri arttıkça süredeki artış diğer yöntemlere göre daha sınırlıdır.

Doğrusal test incelendiğinde ise toplam test süresi ile yetenek düzeyi arasında pozitif bir ilişki olduğunu göstermektedir. Yani yetenek düzeyi arttıkça toplam test süresi de artış göstermektedir. Ancak bu artış, sabit madde sayısının etkisiyle daha yavaş bir eğilim sergilemektedir. Sabit madde sayısı, farklı yetenek düzeylerindeki bireylerin test süresi açısından benzer bir yapı oluşturmalarına neden olmakta ve test süresi üzerindeki değişimin yetenek düzeyine göre daha sınırlı olmasına yol açmaktadır.

Tablo 4.6. *Uygulamalar Arası Test Süreleri Karşılaştırma Sonuçları*

Yöntem	$\bar{X}$	SS	KT	KO	sd	F	p	η^2	Fark
ML	1316,88	737,00							Doğrusal>
EAP	1144,60	717,90	534584104,9	18544263,8	2,88;	54,88	<0,001	0,17	ML>
MAP	940,44	554,69			755,28				EAP>
Doğrusal	1553,88	660,79							MAP

ML, EAP, MAP ve doğrusal test yöntemleri için test sürelerini karşılaştırmak amacıyla yapılan tekrarlı ölçümler varyans analizi sonuçları küresellik varsayımının ihlal edildiğini göstermiştir, $W = 0,935$, $\chi^2(5)=17,580$, $p = 0,004$). Bu nedenle analiz sonuçları, Greenhouse-Geisser düzeltmesi uygulanarak raporlanmıştır. Tekrarlı ölçümler ANOVA sonuçlarına göre yöntemler arasında test süreleri açısından istatistiksel olarak anlamlı bir fark bulunmuştur, $F(2,88, 755,28) = 54,877$, $p < 0,001$, $\eta^2 = 0,173$. Eta kare değeri, test sürelerindeki varyansın %17,3'ünün yöntemler tarafından açıklandığını göstermektedir.

Bonferroni düzeltmeli ikili karşılaştırma testleri sonuçlarına göre doğrusal test yöntemi, diğer tüm yöntemlere kıyasla anlamlı derecede daha uzun süre almıştır ($p < 0,001$). ML yöntemi, EAP ($\Delta_{\text{süre}} = 1720,28$, $p = 0,010$) ve MAP ($\Delta_{\text{süre}} = 3760,44$, $p < 0,001$) yöntemlerinden anlamlı bir şekilde daha uzun süre almıştır. EAP yöntemi de MAP yöntemine kıyasla anlamlı derecede daha uzun süre almıştır ($\Delta_{\text{süre}} = 2040,16$, $p < 0,001$).

Sonuçlar, doğrusal test yönteminin sabit madde yapısı nedeniyle diğer yöntemlere kıyasla daha uzun sürede tamamlandığını göstermektedir. Bireye uyarlanmış yöntemler arasında ise testin MAP yöntemiyle en kısa sürede, ML yönteminde ise en uzun sürede tamamlandığı tespit edilmiştir. EAP yöntemi ise süre açısından ML ile MAP arasında bir performans sergilemiştir.

4.2.6. Madde kullanım sıklıklarına yönelik bulgular

3 farklı canlı BOBUT sonucu madde kullanım sıklıkları, toplam kullanılan madde ve madde çift sayısı, madde örtüşme oranları Tablo 4.7’de verilmiştir.

Tablo 4.7. *Kestirim Yöntemlerine Göre Madde Kullanım Sıklığı Değerleri*

Madde No	ML	EAP	MAP
121	215	240	227
79	190	200	205
96	181	9	4
85	158	182	172
112	156	180	155
80	154	169	182
98	154	173	185
99	142	133	107
41	131	103	79
128	123	141	145
86	121	138	102
100	115	130	120
111	113	127	120
60	109	98	125
83	107	125	99
129	106	114	115
73	94	110	119
97	92	99	99
47	90	38	43
55	90	100	78
76	90	104	83
59	89	111	92
42	87	103	98
1	83	1	5
78	79	87	71
117	76	84	85
118	57	71	55
67	55	78	69
Toplam Kullanılan Madde	88	67	70
Madde Çift Sayısı	6,68	7,03	6,19
Madde Örtüşme Oranı	0,00019	0,00020	0,00018

Kestirim yöntemlerine göre madde kullanım sıklıkları ve madde örtüşme oranları değerlendirildiğinde, yöntemler arasında belirgin farklılıklar gözlenmiştir. ML, EAP ve MAP yöntemlerinde madde kullanım sıklığına bakıldığında, her üç yöntem için de en sık kullanılan maddeler arasında 121, 79 ve 85 numaralı maddeler öne çıkmıştır. ML yöntemi, Madde 121’i 215 kez kullanarak en yüksek kullanımı gerçekleştirmiştir. EAP yöntemi ise aynı maddeyi 240 kez kullanarak ML’ye kıyasla daha yoğun bir kullanım göstermiştir. MAP yönteminde ise Madde 121, 227 kez kullanılarak bu yöntem için en sık kullanılan madde olmuştur.

Toplam kullanılan madde sayısı açısından bakıldığında, ML yöntemi 88 madde, EAP yöntemi 67 madde ve MAP yöntemi 70 madde kullanmıştır. Bu durum, ML yönteminin daha fazla maddeyi ölçme sürecine dâhil ettiğini göstermektedir.

Madde çiftleri arasındaki örtüşme oranları da yöntemlerin madde havuzu kullanım stratejilerini değerlendirmek için önemlidir. ML yöntemi için madde çift başına düşen örtüşme oranı 0,00019, EAP yöntemi için 0,00020, ve MAP yöntemi için 0,00018 olarak hesaplanmıştır. Bu oranlar oldukça düşük olup, kullanılan madde havuzunun yöntemler arasında büyük ölçüde çeşitlilik gösterdiğini ortaya koymaktadır. EAP yöntemi, en yüksek madde örtüşme oranına sahip olup bu durum, madde tekrarlarının biraz daha fazla olduğunu işaret etmektedir. MAP yöntemi ise en düşük örtüşme oranına sahip olup daha az tekrar içeren bir madde seçimi gerçekleştirmiştir. Sonuç olarak ML yöntemi daha geniş bir madde kullanımına ve sıklığına sahipken, EAP ve MAP yöntemleri daha az madde kullanarak ölçme sürecini optimize etmiştir.

BÖLÜM V

SONUÇ, TARTIŞMA VE ÖNERİLER

5.1. Sonuç ve Tartışma

5.1.1. Post-hoc simülasyon sonucu başlama kuralı, madde seçim yöntemi ve sonlandırma kurallarına ait sonuç ve tartışma

Başlama kuralı

Başlama kurallarının ($\theta=0$ ve $\theta=-0,5:0,5$) test performansı üzerindeki etkileri post-hoc simülasyon çalışmaları kapsamında detaylı bir şekilde incelenmiştir. Bulgular, başlama kuralları arasında simülasyon süresi ve madde örtüşme oranı dışında anlamlı bir farklılık olmadığını göstermektedir. Özellikle, $\theta=0$ başlama kuralının daha kısa simülasyon süresi gerektirdiği ve madde örtüşme oranını artırdığı belirlenmiş, ancak korelasyon, RMSE, Bias, standart hata ve madde bilgisi değerleri açısından iki başlama kuralı arasında fark bulunmamıştır.

Literatürde, başlama kurallarının testin genel doğruluğu ve güvenilirliği üzerindeki etkisinin minimal olduğu ve diğer faktörlerin (madde seçim yöntemleri ve yetenek kestirim yöntemleri gibi) test performansı üzerinde daha belirleyici olduğu vurgulanmaktadır (Weiss & Kingsbury, 1984; van der Linden & Glas, 2000; Kezer, 2020; Kezer & Koç, 2014). Weiss (2004), MTK çerçevesinde başlama kurallarının testin güvenilirliği ve geçerliliği üzerindeki etkisinin sınırlı olduğunu belirtmiştir. Magis ve Raiche (2012), catR paketi ile yapılan simülasyonlarda başlama kurallarının test performansına etkisinin sınırlı olduğunu ifade ederken, Kimura (2017) bilgisayarlı uyarlamalı testlerde başlama noktasının test doğruluğu üzerindeki etkisinin minimal olduğunu bulmuştur. Ebenbeck ve Gebhardt (2022) tarafından yapılan çalışmada da benzer şekilde başlama kurallarının test sonuçları üzerindeki etkisinin genellikle düşük olduğu ortaya koyulmuştur. Ancak, Thompson ve Weiss (2021), başlama kurallarının özellikle düşük yetenekli katılımcılar üzerindeki etkisinin daha belirgin olabileceğini öne sürmüştür.

Madde Seçim Yöntemi

Madde seçim yöntemlerinin (MFI, KL, MEI) test performansı üzerindeki etkileri incelendiğinde yöntemler arasında bazı farklılıklar olduğunu belirlenmiştir. Simülasyon süresi açısından MFI yöntemi en kullanışlı bulunurken, MEI yöntemi belirgin şekilde daha uzun hesaplama süreleri gerektirmiştir. MFI'nin hızlı hesaplama avantajı, madde bilgi işlevinin doğrudan matematiksel olarak optimize edilmesiyle açıklanabilir (Kezer & Koç, 2014; van der Linden, 2005). Öte yandan, MEI yönteminin daha uzun süre gerektirmesi, bilgi içeriği dağılımını gözetmesinden kaynaklanmakta ve özellikle büyük madde havuzlarında zaman maliyetini artırmaktadır (Özbaşı & Demirtaşlı, 2016; Wainer, 2000).

Test uzunluğu açısından, MEI yöntemi diğer yöntemlere göre biraz daha kısa test uzunluğu sağlamış, MFI ve KL yöntemleri ise benzer sonuçlar üretmiştir. MEI yönteminin daha kısa test uzunluğu sağlaması, bilgi içeriği ve madde dağılımına yönelik optimize edilmiş bir seçim stratejisi izlediğini göstermektedir (Bulut & Kan, 2012). Benzer şekilde, Chen, Ankenmann ve Chang (2000), MEI'nin en bilgilendirici maddeleri seçerek test uzunluğunu kısaltırken ölçme hassasiyetini koruduğunu belirtmiştir.

Korelasyon değerleri tüm yöntemler için eşit bulunmuş, bu da madde seçim stratejilerinin yetenek kestirim doğruluğu üzerinde benzer bir etkiye sahip olduğunu göstermektedir. Chang ve Ying (1996), madde seçim stratejilerinin belirli koşullarda eşdeğer doğruluk sağlayabileceğini ve özellikle yüksek kaliteli madde havuzlarında farkın minimal olduğunu vurgulamıştır. Bununla birlikte, KL yöntemi daha düşük RMSE değeri sağlayarak doğruluk açısından küçük bir avantaj sunmuştur. KL'nin bilgi fonksiyonunu optimize ederken tahmin belirsizliğini azaltması, bu yöntemin yüksek bilgiye sahip maddeleri seçme kapasitesine bağlanabilir (Chang & Ying, 1996).

Madde örtüşme oranı açısından incelendiğinde, MFI yöntemi en yüksek örtüşme oranını sağlamış ve belirli maddeleri daha sık seçme eğiliminde olmuştur. Bu bulgu, Barrada vd. (2008) MFI'nin çeşitliliği sınırlayabileceğine dair bulgularıyla uyumludur. KL ve MEI yöntemleri ise daha düşük örtüşme oranları sağlayarak madde seçiminde daha dengeli bir yaklaşım sergilemiştir (Chang & Ying, 1996). Standart hata açısından, MFI yöntemi daha düşük hata değeri sağlamış, bu da yüksek bilgilendirici maddeleri seçme kapasitesine bağlanmıştır (Kaplan vd., 2015). KL ve MEI yöntemlerinin benzer hata değerleri üretmesi, bu yöntemlerin bilgi fonksiyonunu dengeli bir şekilde optimize ederek doğruluk seviyesini koruyabildiğini göstermektedir (Butterfield, 2018).

Van der Linden ve Pashley (2000), MFI yönteminin hız ve hesaplama verimliliği açısından avantaj sağladığını, ancak madde örtüşme oranının yüksek olabileceğini belirtmiştir. Bu bulgular, mevcut çalışmada MFI'nin en kullanışlı yöntem olarak öne çıkmasıyla tutarlıdır. KL yönteminin daha düşük RMSE değeri sağlaması ise, Chang ve Ying (1996) tarafından yapılan çalışmalarla örtüşmektedir. Hambleton ve Swaminathan (1985), madde seçim yöntemlerinin genel doğruluk performansında eşit düzeyde etkili olabileceğini vurgulamış ve bu görüş, çalışmada bias ve korelasyon değerlerinin tüm yöntemler için benzer bulunmasıyla desteklenmiştir.

Magis ve Raiche (2012), MEI yönteminin daha geniş bir bilgi tabanı sunduğunu, ancak hesaplama süresinin daha uzun olabileceğini ifade etmiştir. Bu durum, mevcut çalışmada MEI yönteminin belirgin şekilde daha uzun simülasyon sürelerine sahip olmasıyla paralellik göstermektedir. Ebenbeck ve Gebhardt (2022), madde seçim yöntemlerinin madde örtüşme oranı üzerindeki etkisini inceleyerek MFI'nin daha sık örtüşme yaratabileceğini belirtmiştir. Bu bulgu, çalışmada MFI'nin en yüksek madde örtüşme oranını sağlamasıyla uyumludur. Huo (2009), madde seçim stratejilerinin test uzunluğu ve bilgi dağılımı üzerindeki etkisinin minimal olduğunu belirtmiş ve bu durum, çalışmada test uzunluğu ve madde bilgisi açısından yöntemler arasında anlamlı fark bulunmamasıyla örtüşmektedir.

Sonlandırma kuralı

Sonlandırma kurallarının test performansı üzerindeki etkileri incelendiğinde, daha geniş hata toleransına sahip kuralların hesaplama sürelerini kısaltırken, daha hassas kriterlerin hesaplama süresini belirgin şekilde artırdığı görülmüştür. Geniş hata toleransının hesaplama süresini kısaltması, daha az hassasiyet gereksiniminden kaynaklanmaktadır (Thissen & Mislevy, 2000). Özellikle, 0,50-25 madde kuralının en kısa test uzunluğunu sağlaması, daha az maddeyle sonuca ulaşılmasını mümkün kılmıştır. Buna karşılık, 0,30 gibi daha hassas hata toleransına sahip kurallar, yüksek doğruluk gereksinimi nedeniyle test uzunluğunu artırmıştır. Bu bulgular, hata toleransı ile test uzunluğu arasındaki ters ilişkiyi doğrulamaktadır (Thissen & Mislevy, 2000).

Sonlandırma kurallarının korelasyon değerleri üzerindeki etkileri incelendiğinde, daha hassas kriterlerin (örneğin, 0,30 toleransı) daha yüksek korelasyon değerleri sağladığı görülmüştür. Weiss ve Kingsbury (1984), hata toleransının düşürülmesinin test doğruluğunu artırdığını, ancak test süresini uzattığını vurgulamıştır. RMSE açısından, 0.3 kriteri en düşük değeri sağlamış ve hassas sonlandırma kurallarının doğruluk açısından avantaj sağladığını

ortaya koymuştur. Geniş hata toleranslı kurallar ise daha yüksek RMSE değerleri üretmiş ve tahmin belirsizliğini artırmıştır (Weiss & Kingsbury, 1984).

Bias açısından, 0,30 kriteri en düşük değeri sağlamış ve yüksek doğruluk hedefli uygulamalar için en avantajlı seçenek olarak öne çıkmıştır. Geniş toleranslı kurallar ise daha yüksek bias değerleri üretmiş ve erken sonlandırmanın tahmin kesinliğini olumsuz etkilediğini göstermiştir. Bu bulgular, hata toleransı ile doğruluk arasındaki ilişkiyi ortaya koyan Weiss ve Kingsbury'nin (1984) çalışmasıyla uyumludur. Örtüşme oranları açısından, geniş hata toleransı içeren kuralların daha düşük, hassas kuralların ise daha yüksek örtüşme oranları ürettiği gözlemlenmiştir. Daha hassas kuralların test uzunluğunu uzatarak belirli maddelerin daha sık seçilmesine neden olduğu bulgusu, Weiss ve Kingsbury'nin (1984) hata toleransı ile test içeriği çeşitliliği arasındaki ilişkiye dair bulgularıyla örtüşmektedir.

Madde bilgisi değerleri açısından, 0,50 ve 0,50-25 madde kriterleri en yüksek madde bilgisi değerlerini üretmiş ve daha bilgilendirici maddelerin seçilmesine olanak tanımıştır. Segall (2005), daha geniş hata toleransına sahip kuralların madde bilgisi üzerinde olumlu etkileri olduğunu belirtmişlerdir. Buna karşılık, 0,30 ve 25 madde kriterlerinin daha düşük bilgi değerleri üretmesi, hassas kriterlerin bilgi optimizasyonunu sınırlayabileceğini göstermektedir.

5.1.2. Post-hoc simülasyon sonucu yetenek kestirim yöntemlerinin genel performansları ve başlama kuralı, madde seçim yöntemi ve sonlandırma kurallarıyla olan kombinasyonlarına ait sonuç ve tartışma

Yetenek kestirim yöntemlerinin (MAP, ML, EAP) performansı incelendiğinde, MAP yöntemi en hızlı sonuç verirken, ML yöntemi daha uzun süre gerektirmiş, EAP yöntemi ise en yüksek hesaplama süresine sahip olmuştur. MAP'in hızlı olması, öncül tanımların etkili kullanımından kaynaklanırken (Segall, 1996), ML yönteminin daha uzun sürede sonuç vermesi, iteratif hesaplama süreçlerine ihtiyaç duymasına bağlanmaktadır (Hambleton & Swaminathan, 1985). EAP'in yüksek hesaplama süresi gerektirmesi ise, olasılık yoğunluk fonksiyonunun integral hesaplamalarından kaynaklanmaktadır (Samejima, 1969).

Test uzunluğu açısından MAP ve EAP yöntemleri benzer sonuçlar üretmiş, ML yöntemi ise belirgin şekilde daha uzun testler oluşturmuştur. Bu durum, ML'nin bilgi elde etme hızını yavaşlatmasıyla açıklanabilir (Segall, 1996). Korelasyon değerleri açısından üç yöntemin de benzer sonuçlar sunduğu, geniş madde havuzlarında eşit doğruluk seviyesine ulaşılabilceği görülmüştür. RMSE değerleri açısından MAP ve EAP daha düşük hata değerleri üretirken, ML

yöntemi daha yüksek RMSE değerine sahip olmuştur (van der Linden & Glas, 2010). MAP ve EAP'ın düşük RMSE değerleri, tahmin hassasiyetini artırma kapasitesini yansıtmaktadır (Eroğlu & Kelecioğlu, 2015). EAP'ın en düşük bias değerine sahip olduğu, MAP'ın biraz daha yüksek, ML'nin ise en yüksek bias değeriyle ayrıştığı görülmüştür. ML'nin en yüksek bias değerine sahip olması, iteratif süreçlerdeki doğruluk kaybına işaret etmektedir.

Madde örtüşme oranı açısından, MAP yöntemi en yüksek örtüşme oranına sahip olmuş, belirli yetenek aralıklarında aynı maddeleri daha sık seçme eğiliminde olduğu gözlemlenmiştir (Glas & van der Linden, 2003). EAP ve ML yöntemleri ise daha düşük örtüşme oranlarıyla test içeriklerini çeşitlendirme açısından avantaj sağlamıştır. Standart hata değerleri açısından, MAP yöntemi en düşük hata değerine ulaşmış ve öncel bilgilerin tahmin belirsizliğini azaltma kabiliyetini göstermiştir (Chen, 1996). ML yöntemi en yüksek hata değerine sahip olurken, EAP daha dengeli bir performans sergilemiştir.

Madde bilgisi açısından MAP ve EAP yöntemleri daha yüksek bilgi değeri üretmiş, ML yöntemi ise diğerlerine göre daha düşük bilgi değerleri sağlamıştır. MAP'ın yüksek bilgi değeri sağlaması, öncel tanımların etkili kullanımıyla açıklanabilir (Chen, 1996). EAP yöntemi, özellikle geniş yetenek aralıklarında bilgi düzeyini artırma kapasitesiyle öne çıkmıştır. ML'nin düşük bilgi değerleri üretmesi ise iteratif süreçlerin düşük bilgilendirici maddeleri seçmesine neden olabilmektedir.

Yetenek kestirim yöntemleri, başlama kuralı, madde seçim yöntemi ve sonlandırma kurallarıyla birlikte değerlendirildiğinde $\theta=0$ başlama kuralı, tüm yöntemler için genellikle daha kısa süreler sağlamıştır. MFI madde seçim yöntemi, simülasyon süresini önemli ölçüde azaltırken, MEI yöntemi hesaplama süresini artırmıştır (Segall, 1996). Sonlandırma kurallarına göre, 0,50 standart hata-25 madde kuralı en hızlı sonuçları sağlarken, 0,30 standart hata kriteri testin sonlanmasını geciktirmiştir (Weiss & Kingsbury, 1984).

Test uzunluğu açısından MFI madde seçim yöntemi, testin daha kısa sürede tamamlanmasını sağlarken, MEI yöntemi test uzunluğunu artırmıştır. $\theta=0$ başlama kuralı, tüm yöntemlerde test uzunluğunu optimize eden en uygun seçenek olarak bulunmuştur. 0,30 standart hata kriteri, tüm yöntemlerde en uzun testleri üretmiş, buna karşın 0,50 standart hata-25 madde kuralı test süresini kısaltmıştır.

Kestirim doğruluğu açısından MAP ve ML yöntemleri en yüksek korelasyon değerlerini sunmuştur. EAP yöntemi, genellikle yüksek doğruluk sağlamış ancak MEI madde seçim yöntemi ile kullanıldığında doğrulukta düşüş gözlenmiştir (Chang & Ying, 1996). $\theta=0$

başlama kuralı, tüm yöntemlerde en yüksek korelasyon değerlerini sunarken, MFI madde seçim yöntemi en yüksek korelasyonu sağlamış, MEI yöntemi ise korelasyonun düşmesine neden olmuştur. Sonlandırma kuralı açısından 0,30 standart hata kriteri en yüksek doğruluğu sağlarken, daha gevşek kriterler kestirim doğruluğunu düşürmüştür.

RMSE açısından, EAP yöntemi genellikle daha yüksek hata değerleri üretmiş ancak MFI yöntemiyle kullanıldığında daha iyi sonuçlar sağlamıştır (Wang vd., 1999). Bias açısından MAP ve EAP yöntemleri en düşük değerleri sunarken, ML yöntemi bazı koşullarda daha yüksek bias seviyeleri göstermiştir. MEI yöntemi, bias değerlerinde artışa neden olmuştur. Başlama kuralı olarak $\theta=0$, en düşük hata ve bias değerlerini sağlamış, 0,30 standart hata kriteri ise en düşük RMSE ve bias değerleriyle kestirim doğruluğunu artırmıştır.

Madde örtüşme oranı açısından MAP yöntemi en düşük örtüşme oranını sağlamıştır. ML yöntemi benzer sonuçlar sunarken, EAP yöntemi özellikle MEI madde seçim yöntemiyle kullanıldığında en yüksek madde örtüşme oranlarını göstermiştir. Bu bulgular, van der Linden ve Glas (2010) tarafından yapılan madde seçim yöntemleri ve madde havuzu ilişkisini ele alan çalışmalarla örtüşmektedir. MFI yöntemi madde havuzunun daha dengeli kullanılmasını sağlarken, MEI yöntemi belirli maddelerin aşırı kullanımına neden olmuştur.

5.1.3. Canlı BOBUT tüm yetenek kestirim yöntemleri için geçerli ortak kuralların belirlenmesine yönelik sonuç ve tartışma

Canlı BOBUT uygulamasına en uygun koşulları belirlemek amacıyla yapılan simülasyon analizleri, başlama kuralı, madde seçim yöntemi ve sonlandırma kuralı bağlamında değerlendirilmiştir. Başlangıç düzeyinin yetenek kestiriminin doğruluğu üzerinde belirgin bir etkisi olmadığı gözlemlenmiştir. Ancak θ başlangıç seviyesinin (-0,5, 0,5) aralığında belirlenmesi, tüm yetenek kestirim yöntemlerinin madde örtüşme oranlarını ciddi şekilde azalttığı için tercih edilmiştir. Literatürde, başlangıç değerinin genellikle test sürecinin başında yeterli bilgi sağlayamayan bir faktör olduğu, ancak belirli başlangıç değerlerinin madde seçim sürecini etkileyebileceği belirtilmiştir (van der Linden & Pashley, 2010). Bu doğrultuda elde edilen bulgu, başlangıç kuralının doğrudan yetenek kestirimi doğruluğuna etkisi olmasa da test verimliliği açısından önem taşıdığını göstermektedir. Madde seçim yöntemleri arasında belirgin bir fark olmadığı gözlemlense de hesaplama süresi açısından MFI yönteminin avantajlı olduğu tespit edilmiştir. Literatürde MFI yönteminin özellikle doğruluk açısından optimal olduğu belirtilmekle birlikte, bazı çalışmalarda KL ve b-optimal yöntemlerinin de belirli koşullar altında daha etkili olduğu ifade edilmektedir (Chang & Ying, 1996). Ancak hesaplama süresi

dikkate alındığında, canlı BOBUT uygulamalarında hızlı yanıt verme gerekliliği göz önünde bulundurulduğunda MFI yönteminin tercih edilmesi rasyonel bir seçim olarak değerlendirilmektedir. Sonlandırma kurallarına ilişkin olarak, 0,30 ve 25 madde ile yapılan kestirimlerin 0,50 ve 25 madde ile yapılan kestirimlere göre daha iyi sonuçlar verdiği belirlenmiştir. Özellikle hem standart hata hem de sabit madde sayısına dayalı sonlandırma kurallarının kullanıldığı senaryoların en uygun sonuçları sağladığı gözlemlenmiştir. Literatürde, test uzunluğunun kontrol edilmesi ve doğruluk arasındaki dengeyi sağlamak adına sabit madde sayısı ve hata bazlı sonlandırma kurallarının birlikte kullanılması önerilmektedir (Weiss & Kingsbury, 1984).

5.1.4. Canlı BOBUT ve doğrusal testlerin yetenek kestirim performanslarına yönelik sonuç ve tartışma

Canlı BOBUT kapsamında yer alan üç yetenek kestirim yöntemiyle doğrusal testlerin yetenek kestirim performanslarına yönelik sonuçlar incelendiğinde ML ve EAP yöntemlerinin tahmin dağılımlarının normal dağılıma oldukça yakın olduğu görülmektedir. Bu durum, literatürde de sıkça vurgulanan bir bulgudur. Özellikle Embretson ve Reise (2000), ML ve EAP yöntemlerinin yeterli gözlem sayısına ulaşıldığında benzer dağılım özellikleri sergilediğini belirtmiştir. MAP yönteminin daha dar bir aralıkta yoğunlaşması ise, bu yöntemin posterior yoğunluk fonksiyonuna dayanarak daha muhafazakâr tahminler ürettiği yönündeki literatürle uyumludur (Baker & Kim, 2004). Doğrusal test sonuçlarının daha geniş bir varyansa sahip olması, KTK ile MTK tabanlı tahmin yöntemleri arasındaki temel farklardan biri olarak değerlendirilebilir. Klasik testlerde her bireyin aynı soruları çözmesi, düşük ve yüksek yetenek seviyelerindeki bireylerde tahmin doğruluğunun azalmasına neden olabilir. Özellikle Lord (1980), klasik testlerin düşük ve yüksek yetenek düzeylerinde hata varyansının arttığını ve güvenilirliğin düştüğünü ifade etmiştir.

Korelasyon analizleri, yöntemler arasında genel olarak pozitif ve anlamlı ilişkiler bulunduğunu göstermektedir. Ancak, adaptif test yöntemleri (ML, MAP ve EAP) kendi aralarında daha yüksek korelasyonlar sergilerken, doğrusal test ile bireye uyarlanmış testler arasındaki korelasyonlar daha düşük seviyelerde kalmıştır. Bu durum, doğrusal testlerin bireyselleştirilmemiş yapısının yetenek tahminlerinde adaptif testlere kıyasla daha fazla hata içerebileceğini göstermektedir. Thissen ve Mislevy (1990), doğrusal testlerin tüm bireyler için aynı maddeleri içermesi nedeniyle, bireysel yetenek seviyelerine uygun olmayan madde dağılımının tahmin doğruluğunu olumsuz etkileyebileceğini belirtmiştir. Özellikle düşük

yetenek seviyesindeki bireylerin testin tamamında zorlandığı, yüksek yetenek seviyesindeki bireylerin ise soruların büyük bir kısmını rahatlıkla çözebildiği göz önüne alındığında, doğrusal testlerin geniş bir varyansa sahip olmasının beklenen bir sonuç olduğu söylenebilir. Bununla birlikte, ML ve EAP yöntemleri arasındaki yüksek korelasyon, bu iki yöntemin benzer tahmin stratejilerine dayandığını ve literatürde belirtilen bulgularla tutarlı olduğunu göstermektedir (Reise & Henson, 2003).

Yetenek kestirim yöntemleri arasındaki farklılıkları test etmek amacıyla gerçekleştirilen tekrarlı ölçümler varyans analizi, yöntemler arasında anlamlı bir fark olduğunu göstermektedir. Ancak, etki büyüklüğünün düşük olması, bu farkın pratik anlamda büyük bir farklılık yaratmadığını düşündürmektedir. Bu bulgu, özellikle adaptif testlerin doğrusal testlerden daha kesin sonuçlar verdiğini, ancak bu farkın uygulama açısından çok büyük olmayabileceğini öne süren çalışmalarla örtüşmektedir (Weiss, 1984). EAP yönteminin diğer yöntemlere kıyasla daha yüksek yetenek tahminleri üretmesi, literatürde de tartışılan bir konudur. Mislevy ve Stocking (1989), EAP yönteminin prior dağılımdan gelen bilgiyi kullanması nedeniyle özellikle düşük yetenek seviyelerinde tahminleri sistematik olarak yukarı kaydırabileceğini belirtmiştir. Aynı şekilde, Samejima (2016) EAP'nin özellikle düşük yetenek seviyelerindeki bireylerde daha az muhafazakâr tahminler sunduğunu ve bu durumun yöntemin matematiksel yapısından kaynaklandığını ifade etmiştir. Bu bulgu, çalışmada elde edilen sonuçlarla doğrudan örtüşmektedir.

ML ve MAP yöntemleri arasında anlamlı bir fark bulunmaması, bu iki yöntemin benzer tahmin yapıları sergilediğini göstermektedir. Baker ve Kim (2004), MAP yönteminin ML yöntemine kıyasla düşük yetenek seviyelerinde biraz daha muhafazakâr tahminler ürettiğini, ancak genel olarak benzer sonuçlar verdiğini belirtmiştir. Bu bağlamda, çalışmanın bulguları, MAP yönteminin tahmin aralığının daha dar olduğunu göstermesi açısından literatürdeki bulgularla örtüşmektedir. Doğrusal test sonuçlarının ML ve MAP ile benzer yetenek tahminleri üretmesi, doğrusal testlerin bireye uyarlanmış testlerle belirli bir seviyede karşılaştırılabilir olduğunu düşündürmektedir. Ancak, bu benzerlik yüksek yetenek seviyelerinde doğrusal testin getirdiği sınırlamalar nedeniyle azalabilir.

5.1.5. Yetenek kestirim yöntemlerinin Canlı BOBUT standart hata performanslarına ait sonuç ve tartışma

Elde edilen standart hata değerleri, yöntemlerin bireylerin yetenek düzeylerini ne kadar hassas tahmin edebildiğine dair önemli bilgiler sunmaktadır. Bulgular, EAP ve MAP yöntemlerinin hata tahminlerinin en düşük olduğu yöntemler olduğunu göstermektedir. Bu durum, literatürde de desteklenen bir bulgudur. Özellikle Bock ve Mislevy (1982), EAP ve MAP yöntemlerinin prior bilgi kullanımına dayandığı için hata varyansını düşürdüğünü ve tahmin edilen yetenek değerlerinin daha az dalgalanma gösterdiğini belirtmiştir. MAP yöntemi, posterior dağılımın en yüksek yoğunluğa sahip olduğu noktada yetenek kestirimi yaparak tahminlerin daha dengeli olmasını sağlarken, EAP yöntemi prior dağılımla birlikte tüm olasılıkları entegre ederek hata oranını düşürmektedir. Bu sonuçlar, Thissen ve Orlando (2001) tarafından yapılan çalışmalarda da desteklenmekte ve özellikle EAP'nin ölçüm hatasını sistematik olarak azalttığı gösterilmektedir.

ML yönteminin daha yüksek hata değerlerine sahip olması ve ölçüm hatasının daha fazla değişkenlik göstermesi, özellikle düşük ve yüksek yetenek seviyelerinde bu yöntemin daha az güvenilir olduğunu ortaya koymaktadır. Bu durum, ML yönteminin prior bilgi kullanmamasından kaynaklanmaktadır. Özellikle Baker ve Kim (2004), ML yönteminin uç değerlerde kararsız sonuçlar üretebildiğini ve hata varyansının arttığını belirtmiştir. Bunun temel nedeni, düşük yetenek seviyelerinde bireylerin düşük madde doğru cevap olasılığına sahip olması ve yüksek yetenek seviyelerinde ise testin yeterince zorluk sunamaması nedeniyle kestirimin daha az hassas hale gelmesidir. Aynı durum, Lord (1980) tarafından da vurgulanmış ve ML'nin uç değerlerde güvenilirliğinin düştüğü belirtilmiştir.

Standart hata değerlerinin yetenek düzeyine göre değişimi incelendiğinde, yetenek düzeyi arttıkça hata değerlerinin azaldığı görülmektedir. Bu bulgu, özellikle MTK tabanlı kestirimlerde sıklıkla vurgulanan bir durumdur. Embretson ve Reise (2000), yetenek düzeyi arttıkça bireylerin doğru cevap verme olasılıklarının arttığını ve bu durumun kestirim güvenilirliğini artırdığını belirtmiştir. Ancak, ML yönteminin her iki uç noktada da diğer yöntemlere kıyasla daha yüksek hata değerleri sunduğu gözlemlenmiştir. Özellikle düşük yetenek seviyelerinde ölçüm hatasının daha büyük olması, yetenek kestiriminin kararsız hale geldiğini ve test sonuçlarının daha az güvenilir olduğunu göstermektedir. MAP yöntemi, özellikle orta yetenek düzeylerinde dengeli bir performans gösterirken, EAP yöntemi tüm yetenek seviyelerinde en düşük hata değerlerini sunarak en güvenilir kestirim yöntemini temsil

etmektedir. Bu durum, MAP yönteminin posterior yoğunluk fonksiyonuna dayalı olarak daha dengeli bir hata dağılımı sunduğunu belirten Samejima (2016) tarafından da desteklenmektedir.

Tekrarlı ölçümler varyans analizi sonuçları, yöntemler arasında standart hata açısından anlamlı farklılıklar olduğunu ortaya koymaktadır. ML yöntemi hem EAP hem de MAP yöntemlerine kıyasla daha yüksek standart hata değerlerine sahiptir. Bu sonuçlar, Weiss (2004) tarafından yapılan çalışmalarda da ortaya konmuş ve ML yönteminin özellikle bireye uyarlanmış testlerde hata varyansının diğer yöntemlere kıyasla daha yüksek olduğu gösterilmiştir. Aynı şekilde, EAP yönteminin hata miktarı açısından ML ve MAP arasında bir konumda bulunması hem teorik hem de deneysel çalışmalarla örtüşmektedir.

5.1.6. Canlı BOBUT ve doğrusal testlerin güvenilirlik performanslarına yönelik sonuç ve tartışma

Bulgular bireye uyarlanmış testlerin, klasik doğrusal testlere kıyasla daha yüksek bilgi düzeyi sunduğunu ve yetenek kestiriminde daha güvenilir ölçümler sağladığını göstermektedir. Literatürde, bireye uyarlanmış testlerin güvenilirliğinin, testin her bireyin yetenek seviyesine göre şekillenmesine bağlı olarak arttığı sıkça vurgulanmaktadır. Özellikle Lord (1980), bireye uyarlanmış testlerin, klasik doğrusal testlere kıyasla daha yüksek bilgi sağladığını ve özellikle orta yetenek seviyelerinde daha hassas ölçüm sunduğunu belirtmiştir. Bulgular, ML, EAP ve MAP yöntemlerinin test bilgisinin en yüksek olduğu bölgenin orta düzey yetenekler olduğunu göstermektedir. Bu durum, MTK tabanlı testlerin genellikle ortalama yetenek seviyesindeki bireyleri en iyi şekilde ölçmek için tasarlandığını ve bu seviyede en yüksek bilgi miktarını sağladığını gösteren Embretson ve Reise (2000) tarafından ortaya konan bulgularla uyumludur.

Uç yetenek seviyelerinde test bilgisinin azalması, bireye uyarlanmış testlerde bile madde dağılımının belirli bir yetenek aralığına odaklanmasının bir sonucu olabilir. Özellikle düşük yetenek seviyelerinde test maddeleri bireylerin büyük bir kısmı için çok zor olabilirken, yüksek yetenek seviyelerinde ise maddeler yeterince zorlayıcı olmayabilir. Thissen ve Mislevy (2000), uç yetenek seviyelerinde madde seçiminin kritik bir öneme sahip olduğunu ve eğer testin madde bankası her yetenek seviyesine uygun maddeler içermiyorsa, bilgi düzeyinin düşebileceğini belirtmiştir. ML ve MAP yöntemleri, uç yetenek seviyelerinde test bilgisinin en hızlı düştüğü yöntemler olarak karşımıza çıkmaktadır. Özellikle ML yöntemi, uç yetenek seviyelerinde ölçüm doğruluğunu kaybetmesiyle bilinen bir yöntemdir (Baker & Kim, 2004). MAP yöntemi ise posterior yoğunluk dağılımını temel aldığı için, yetenek tahminlerinde orta bölgelere daha fazla ağırlık verir ve uç bölgelerde test bilgisi hızla azalabilir. Buna karşın, EAP

yöntemi daha dengeli bir bilgi dağılımı sergilemekte ve uç yetenek seviyelerinde diğer yöntemlere göre daha istikrarlı bilgi sunmaktadır. Bu durum, EAP'nin prior bilgi kullanımına dayalı olması ve tahminlerinde daha dengeli bir yaklaşım sergilemesiyle açıklanabilir (Mislevy & Stocking, 1989).

Doğrusal test yöntemi ile bireye uyarlanmış testler arasındaki farklar incelendiğinde, klasik doğrusal testlerin bilgi miktarı açısından daha geniş bir dağılım sergilediği ve güvenilirliğinin homojen olmadığı görülmektedir. Bu durum, doğrusal testlerin sabit bir madde yapısına sahip olması ve tüm bireyler için aynı maddeleri kullanması nedeniyle ortaya çıkmaktadır. Weiss (1985), doğrusal testlerin özellikle yetenek dağılımının uç noktalarında düşük bilgi sunduğunu ve bireye uyarlanmış testlerin bilgi miktarını daha verimli kullandığını vurgulamaktadır. Bulgular, doğrusal testlerin ortalama bilgi miktarının bireye uyarlanmış yöntemlere kıyasla daha düşük olduğunu ve bireyselleştirilmiş ölçüm yapamaması nedeniyle yetenek kestiriminde daha az hassasiyet sunduğunu göstermektedir.

Test bilgi düzeyleri açısından yöntemler arasında istatistiksel olarak anlamlı farklılıklar olduğunu göstermektedir. Özellikle ML yöntemi, EAP ve MAP yöntemlerine kıyasla anlamlı derecede daha düşük test bilgisi sağlamıştır. Bu bulgu, ML yönteminin özellikle uç yetenek seviyelerinde daha düşük güvenilirlik sunduğunu belirten Baker ve Kim (2004) tarafından desteklenmektedir. EAP yöntemi, MAP yöntemine göre anlamlı olarak daha düşük test bilgisi sağlarken, doğrusal teste kıyasla daha yüksek bilgi sunmuştur. MAP yöntemi ise en yüksek bilgi düzeyine sahip yöntem olarak ortaya çıkmıştır. Bu sonuçlar, özellikle testin güvenilirliğini artırmak isteyen uygulamalarda MAP yönteminin daha avantajlı olabileceğini göstermektedir.

5.1.7. Canlı BOBUT ve doğrusal testlerin test uzunluklarına yönelik sonuç ve tartışma

Bulgular, bireye uyarlanmış testlerin test uzunluğu açısından doğrusal testlere kıyasla nasıl bir avantaj sunduğunu göstermektedir. Literatürde, bireye uyarlanmış testlerin klasik doğrusal testlere kıyasla daha kısa sürede ve daha az maddeyle benzer doğrulukta tahmin yapabildiği vurgulanmaktadır (Weiss, 1985). Bu bağlamda söz konusu bu çalışmada doğrusal testlerin en yüksek ortalama madde sayısına sahip olduğu, bireye uyarlanmış testlerin ise bireylerin yetenek seviyesine göre değişken madde sayılarıyla daha kullanışlı bir ölçüm sağladığı görülmektedir. Özellikle MAP yönteminin en düşük ortalama madde sayısını sunması, bu yöntemin ölçüm sürecini en kullanışlı şekilde optimize ettiğini düşündürmektedir.

Farklı yetenek kestirim yöntemleri karşılaştırıldığında, ML yönteminin uç yetenek düzeylerinde daha fazla maddeye ihtiyaç duyduğu gözlemlenmiştir. Bu durum, ML'nin uç noktalarda yetenek kestirimi yapmak için daha fazla bilgiye ihtiyaç duymasıyla açıklanabilir. Baker ve Kim (2004), ML yönteminin prior bilgi kullanmadığını ve bu nedenle düşük ve yüksek yetenek seviyelerinde kestirimi stabilize etmek için daha fazla maddeye ihtiyaç duyduğunu belirtmiştir. Özellikle -2 ve +2 yetenek seviyelerinin altında ve üstünde ML'nin diğer yöntemlere kıyasla daha fazla maddeyle ölçüm yaptığı görülmektedir. MAP ve EAP yöntemleri ise uç yetenek düzeylerinde daha dengeli bir test uzunluğu sunmuştur. MAP yöntemi, özellikle uç değerlerde EAP'ye kıyasla biraz daha az madde kullanarak ölçüm sürecini optimize etmiş ve daha kullanışlı bir süreç sunduğunu göstermiştir. Bu durum, MAP yönteminin posterior dağılıma dayalı olarak daha kesin kestirimler yapabilmesiyle açıklanabilir.

Test uzunluğu açısından yöntemler arasında anlamlı farklılıklar bulunduğunu göstermektedir. Doğrusal test yönteminin bireye uyarlanmış yöntemlere kıyasla anlamlı derecede daha uzun olduğunu göstermektedir. Bu bulgu, doğrusal testlerin tüm bireylere aynı sayıda madde sunması nedeniyle test süresinin bireyin yetenek seviyesinden bağımsız olarak uzun sürdüğünü belirten Lord (1980) tarafından belirtilmiştir. Bireye uyarlanmış testler arasında karşılaştırma yapıldığında ML yöntemi hem EAP hem de MAP yöntemine kıyasla anlamlı şekilde daha fazla madde kullanmaktadır. Bu sonuç, ML yönteminin düşük ve yüksek yetenek seviyelerinde daha fazla bilgiye ihtiyaç duyması nedeniyle test süresinin uzamasına yol açtığını gösteren çalışmalarla örtüşmektedir (Embretson & Reise, 2000). EAP yöntemi, ML'ye kıyasla daha az maddeyle test sürecini tamamlamaktadır ancak MAP yöntemi en düşük madde sayısı ile ölçümü tamamlayan yöntem olarak öne çıkmaktadır. MAP yönteminin en düşük test uzunluğu ile en yüksek test bilgisini sağlaması, bu yöntemin ölçüm sürecinde en kullanışlı stratejiyi sunduğunu göstermektedir. Samejima (1997), MAP yönteminin özellikle posterior yoğunluk fonksiyonuna dayalı olarak daha kesin kestirimler sunduğunu ve bu nedenle daha az maddeyle yüksek doğruluk sağlayabildiğini belirtmiştir.

5.1.8. Canlı BOBUT ve doğrusal testlerin ölçüm süresine yönelik sonuç ve tartışma

Bulgular bireye uyarlanmış testlerin doğrusal testlere kıyasla ölçüm sürecini nasıl optimize ettiğini ve farklı yetenek kestirim yöntemlerinin test süresi üzerindeki etkisini ortaya koymaktadır. Literatürde, bireye uyarlanmış testlerin test süresini kısaltarak daha kullanışlı bir değerlendirme sunduğu sıkça vurgulanmaktadır. Özellikle Weiss (1982), bireye uyarlanmış testlerin, bireylerin yetenek düzeyine uygun maddeleri seçerek toplam test süresini azalttığını

ve testin daha etkin bir şekilde tamamlanmasını sağladığını belirtmiştir. Söz konusu bu çalışmada da doğrusal test yönteminin en uzun test süresine sahip olduğu, bireye uyarlanmış testlerin ise daha kısa sürede tamamlandığı görülmektedir.

Farklı yetenek kestirim yöntemleri karşılaştırıldığında, ML yönteminin diğer yöntemlere kıyasla daha uzun test süresi gerektirdiği belirlenmiştir. Özellikle yüksek yetenek seviyelerinde ML'nin test süresinde belirgin bir artış gösterdiği görülmektedir. Baker ve Kim (2004), ML yönteminin özellikle uç yetenek seviyelerinde kestirim doğruluğunu artırmak için daha fazla maddeye ihtiyaç duyduğunu ve bu nedenle test süresinin uzadığını belirtmiştir. EAP yöntemi, test süresi açısından ML yöntemine göre daha dengeli ve nispeten daha kısa bir süre sunarken, MAP yöntemi en kısa test süresiyle en kullanışlı kestirimi gerçekleştirmiştir. MAP yönteminin en kısa sürede tamamlanması, bu yöntemin posterior yoğunluk fonksiyonuna dayalı olarak daha kesin kestirimler yapabilmesiyle açıklanabilir. Samejima (1997), MAP yönteminin yetenek tahminlerinde daha kesin sonuçlar sunduğunu ve bu nedenle daha az maddeyle test sürecini tamamlayabildiğini belirtmiştir. MAP yöntemi, en kısa test süresiyle en hızlı ve etkin yöntemi temsil etmektedir. Buna karşılık, EAP yöntemi ML'ye göre daha kısa bir süre sunmasına rağmen, MAP yöntemine kıyasla daha uzun sürede tamamlanmıştır. Bu bulgu, EAP yönteminin tüm yetenek seviyelerinde daha dengeli bir ölçüm sunduğunu gösteren Mislevy ve Stocking (1989) tarafından yapılan çalışmalarla da uyumludur.

Doğrusal test yöntemi incelendiğinde, bireylerin yetenek seviyesine bağlı olarak test süresinin değişmekle birlikte, sabit madde sayısı nedeniyle bu değişimin sınırlı olduğu gözlemlenmiştir. Lord (1980), doğrusal testlerin her bireye aynı madde setini sunduğunu ve bu nedenle test süresinin yetenek seviyelerine tam olarak uyum sağlayamadığını belirtmiştir. Söz konusu bu çalışmada da doğrusal testlerin sabit madde yapısı nedeniyle diğer yöntemlerden belirgin şekilde daha uzun sürede tamamlandığı belirlenmiştir. Sonuçlar doğrusal testin diğer yöntemlerden anlamlı derecede daha uzun süre aldığını göstermektedir.

Test süresi açısından yöntemler arasında istatistiksel olarak anlamlı farklılıklar bulunduğunu göstermektedir. Özellikle ML yöntemi hem EAP hem de MAP yöntemine kıyasla anlamlı derecede daha uzun süre gerektirmiştir. Bu bulgu, ML yönteminin özellikle uç yetenek seviyelerinde daha fazla maddeye ihtiyaç duyması nedeniyle test süresinin arttığını gösteren çalışmalarla örtüşmektedir (Embretson & Reise, 2000). EAP yöntemi, ML'ye kıyasla daha kısa sürede tamamlanırken, MAP yöntemi en düşük test süresini sağlamıştır. Bu sonuçlar, özellikle test süresinin optimize edilmesi gereken uygulamalarda MAP yönteminin en kullanışlı seçenek olabileceğini göstermektedir.

5.1.9. Yetenek kestirim yöntemlerinin Canlı BOBUT madde kullanım durumlarına ait sonuç ve tartışma

Bulgular, bireye uyarlanmış testlerde farklı yetenek kestirim yöntemlerinin madde kullanım stratejilerini nasıl şekillendirdiğini göstermektedir. Literatürde, bireye uyarlanmış testlerin madde seçim sürecinde belirli algoritmalar kullanarak her birey için en uygun maddeleri belirlediği ve bu sürecin kestirim yöntemine bağlı olarak değişebileceği vurgulanmaktadır (van der Linden & Pashley, 2010). Bu çalışmada da ML yönteminin daha geniş bir madde kullanımına sahip olduğu, EAP ve MAP yöntemlerinin ise daha az madde kullanarak ölçme sürecini optimize ettiği görülmektedir.

Madde kullanım sıklıkları incelendiğinde, tüm yöntemler için belirli maddelerin daha sık kullanıldığı görülmektedir. Özellikle Madde 121'in ML, EAP ve MAP yöntemlerinde en sık kullanılan madde olması, madde seçim algoritmalarının belirli maddeleri sıkça tercih ettiğini göstermektedir. Bu durum, testin belirli yetenek düzeylerinde en bilgilendirici maddeleri seçtiği hipoteziyle uyumludur. Thissen ve Mislevy (1990), adaptif testlerde madde seçim sürecinin belirli maddeleri daha sık kullanma eğiliminde olduğunu ve bunun madde havuzunun içeriğine bağlı olarak değişebileceğini belirtmiştir.

Toplam kullanılan madde sayıları incelendiğinde, ML yönteminin diğer yöntemlerden daha geniş bir madde havuzundan yararlandığı görülmektedir. EAP ve MAP yöntemleri ise daha az madde kullanarak ölçme sürecini optimize etmiştir. EAP ve MAP yöntemleri madde ölçüm sürecinde daha az maddeyle yetenek kestirimi yapabildiğini göstermektedir. Bu sonuçlar, Weiss ve Kingsbury (1984) tarafından yapılan çalışmalarda da desteklenmektedir. Araştırmalar, bireye uyarlanmış testlerde kullanılan madde sayısının, kestirim yöntemine bağlı olarak değiştiğini ve bazı yöntemlerin daha az maddeyle benzer doğrulukta sonuçlar üretebildiğini göstermektedir. Çalışmamızda ML yönteminin daha fazla madde kullanması, bu yöntemin kestirim doğruluğunu artırmak için daha fazla bilgiye ihtiyaç duymasından kaynaklanabilir.

Madde örtüşme oranları da yöntemler arasındaki madde seçim stratejilerini değerlendirmek için önemli bir ölçüttür. Söz konusu bu çalışmada tüm yetenek kestirim yöntemleri için örtüşme oranının oldukça düşük olup, yöntemlerin madde havuzundaki çeşitliliği büyük ölçüde koruduğunu göstermektedir. Özellikle EAP yönteminin en yüksek madde örtüşme oranına sahip olması, belirli maddeleri diğer yöntemlere kıyasla daha fazla tekrar ettiğini göstermektedir. MAP yöntemi ise en düşük örtüşme oranına sahip olup, daha

geniş bir madde yelpazesinden seçim yapmıştır. Bu bulgu, MAP yönteminin madde havuzunu daha etkin bir şekilde kullandığını gösteren van der Linden (2005) ait sonuçları desteklemektedir.

5.2. Öneriler

5.2.1. Araştırmacılara Öneriler

Bu çalışmada, başlama kurallarının test performansı üzerindeki etkisinin sınırlı olduğu belirlenmiştir. Ancak, özellikle farklı yetenek seviyelerindeki katılımcılar için özelleştirilmiş başlama kurallarının geliştirilmesi veya önsel bilginin kullanıldığı senaryoların değerlendirilmesi, testin etkinliğini artırabilir. Örneğin, düşük yetenekli bireyler için başlangıç noktasının daha düşük bir theta değeriyle ($\theta = -1$), yüksek yetenekli bireyler için ise daha yüksek bir theta değeriyle ($\theta = 1$) belirlenmesi, test sürecine ve ölçme doğruluğuna nasıl bir katkı sağlayacağı açısından incelenebilir. Ayrıca, adaptif başlama kurallarının kullanılmasıyla, testin bireylerin başlangıç performansına daha iyi uyum sağlayacak şekilde şekillendirilmesi sağlanabilir. Gelecekteki çalışmalarda, farklı başlama stratejilerinin, test süresi, madde sayısı, ölçme duyarlılığı ve bireylerin test deneyimi üzerindeki etkileri detaylı bir şekilde araştırılmalıdır.

Araştırma bulguları, MFI, KL ve MEI yöntemlerinin çeşitli avantaj ve dezavantajlara sahip olduğunu göstermektedir. Her bir yöntemin belirli koşullarda daha iyi performans gösterdiği göz önünde bulundurulduğunda, bu yöntemlerin hibrit bir şekilde kullanılması hem hızlı hesaplama süreci hem de madde çeşitliliğinin sağlanması açısından avantajlı olabilir. Örneğin, testin başlangıcında MFI yöntemi kullanılarak hızlı bir madde seçimi yapılabilir, test ilerledikçe KL veya MEI yöntemleri devreye alınarak daha dengeli ve bilgi açısından zengin bir madde seçimi gerçekleştirilebilir. Bununla birlikte, farklı madde seçim algoritmalarının hibrit modellerle birleştirilmesi ve bu kombinasyonların test performansı üzerindeki etkilerinin incelenmesi, gelecekteki araştırmalar için önemli bir çalışma alanı sunmaktadır.

Sonlandırma kuralları, testin uzunluğu ve ölçüm doğruluğu arasındaki dengeyi sağlamada kritik bir rol oynamaktadır. Araştırma sonuçları, sabit madde sayısına dayalı sonlandırma kuralı ile standart hataya dayalı sonlandırma kuralının birlikte kullanılmasının, test performansını olumlu yönde etkilediğini göstermektedir. Bunun yanı sıra, dinamik sonlandırma kurallarının geliştirilmesiyle test sürecinde daha esnek bir yapı oluşturulabilir. Örneğin, testin başlangıç aşamasında daha geniş bir hata toleransı kullanılırken, test ilerledikçe bu tolerans

kademeli olarak daraltılabilir. Böylece, test süresi optimizasyonu sağlanırken ölçme duyarlılığı da artırılabilir. Gelecekteki çalışmalar, dinamik sonlandırma stratejilerinin adaptif test süreçlerine nasıl entegre edilebileceğini ve bu yaklaşımın bireylerin yetenek kestirimi üzerindeki etkilerini detaylı bir şekilde inceleyebilir.

Bu çalışmada temel olarak test başlama, madde seçme ve test sonlandırma kuralları sabitlenmiş ve farklı yetenek kestirim yöntemleri canlı bir CAT (Bilgisayar Ortamında Bireye Uyarlanmış Test) uygulamasıyla karşılaştırılmıştır. Gelecekteki çalışmalarda, belirli bir yetenek kestirim yönteminin sabitlenerek, diğer değişkenlerin (başlama, madde seçme ve sonlandırma kuralları) farklı kombinasyonlarla karşılaştırılması önerilmektedir. Böylelikle, belirli bir kestirim yönteminin farklı test koşulları altındaki performansı daha iyi analiz edilebilir.

Yapay zeka ve makine öğrenmesi teknikleri, madde seçimi ve yetenek kestirimi süreçlerini daha etkin hale getirmek için önemli fırsatlar sunmaktadır. Örneğin, derin öğrenme modelleri kullanılarak madde seçim süreçleri optimize edilebilir ve bireylerin yetenek seviyeleri daha hassas bir şekilde kestirilebilir. Özellikle doğal dil işleme (NLP) ve takviye öğrenme gibi teknikler kullanılarak, adaptif test sistemlerinin karar mekanizmaları daha dinamik ve özelleştirilmiş hale getirilebilir. Gelecekte, bu tür teknolojilerin adaptif test uygulamalarına entegrasyonu üzerine yapılacak çalışmalar, testlerin ölçme gücünü ve birey uyumunu artırmada önemli katkılar sunabilir.

Madde havuzunun kalitesi ve çeşitliliği, test performansını doğrudan etkileyen en önemli faktörlerden biridir. Özellikle uç yetenek seviyelerindeki bireyler için uygun maddelerin bulunmaması, testin bu bireyler için düşük duyarlılıkla sonuçlanmasına neden olabilir. Mevcut çalışmada kullanılan madde havuzu, uç yetenek seviyelerine hitap eden madde sayısı açısından sınırlı kalmıştır. Gelecek araştırmalarda, uç yetenek seviyelerine özel maddelerin eklenmesi, ölçme doğruluğunu artırmak için değerlendirilebilir. Ayrıca, madde havuzunun sürekli güncellenmesi ve yeni maddelerin eklenmesi, testin dinamik yapısını koruyarak farklı öğrenci grupları için daha dengeli bir ölçme süreci sağlayabilir.

Araştırma bulguları, uç yetenek seviyelerindeki bireyler için test performansının düşük olduğunu göstermektedir. Bu durumu iyileştirmek adına, özellikle uç yetenek seviyelerindeki bireylere uygun madde seçme stratejilerinin geliştirilmesi önem arz etmektedir. Bunun yanı sıra, test sürecinin bu bireyler için daha hassas hale getirilmesi adına, özel sonlandırma kuralları ve daha güçlü madde seçim algoritmaları geliştirilebilir. Örneğin, uç yetenek seviyelerindeki

bireyler için daha karmaşık ve ayırt ediciliği yüksek maddelerin belirlenmesi, test performansını artırabilir.

Madde örtüşme oranının yüksek olması, testin sunduğu madde çeşitliliğini kısıtlayarak ölçme süreçlerini olumsuz etkileyebilir. Bu nedenle, madde seçim algoritmalarının madde örtüşme oranını azaltacak şekilde düzenlenmesi önerilmektedir. Örneğin, belirli maddelerin tekrar kullanımını sınırlayan bir mekanizma eklenebilir veya madde seçiminde çeşitliliği artıracak yeni stratejiler geliştirilebilir. Bu tür modifikasyonların test sürecine etkisini anlamak adına, farklı madde seçim algoritmalarının kıyaslanması yararlı olacaktır.

Son olarak, bu çalışmada elde edilen bulguların farklı popülasyonlarda (örneğin, farklı yaş grupları, eğitim seviyeleri veya kültürel geçmişlere sahip bireyler) geçerliliğinin incelenmesi önerilmektedir. Özellikle, farklı demografik özelliklere sahip bireyler arasında test performansının nasıl değiştiği ve bu popülasyonlara özgü optimizasyon stratejilerinin nasıl geliştirilebileceği önemli bir araştırma konusu olarak öne çıkmaktadır. Farklı örneklem gruplarında yürütülecek çalışmalar, CAT uygulamalarının genel geçerliliğini ve etkinliğini artırmada büyük katkı sağlayacaktır.

5.2.2. Test Uygulayıcılarına Öneriler

BOBUT geleneksel sabit formlu testlere kıyasla daha kısa sürede daha yüksek ölçme doğruluğu sağladığı için test uygulayıcıları ve eğitimciler tarafından tercih edilmelidir. Araştırma, BOBUT'un aynı doğruluk seviyesine ulaşmak için geleneksel testlere kıyasla yaklaşık %50 daha az maddeye ihtiyaç duyduğunu göstermektedir. Bu, hem test süresini kısaltarak aday yorgunluğunu azaltmakta hem de test sürecini daha kullanışlı hale getirmektedir. Özellikle geniş ölçekli sınavlarda ve büyük öğrenci gruplarında, test süresinin kısılması operasyonel maliyetleri önemli ölçüde düşürmekte, değerlendirme süreçlerini hızlandırmakta ve geri bildirim süresini kısaltmaktadır. BOBUT bir diğer önemli avantajı, her bireyin yetenek seviyesine uygun sorularla değerlendirilmesini sağlayarak ölçme duyarlılığını artırmasıdır. Sabit formlu testlerde, düşük yetenek seviyesindeki bireyler için çok zor, yüksek yetenek seviyesindeki bireyler için ise çok kolay maddeler sunulması ölçme hatalarına neden olabilir. BOBUT ise, bireyin yetenek seviyesini belirleyerek ona en uygun soruları seçtiği için ölçme hatasını en aza indirir. Bu durum, özellikle heterojen grupların değerlendirildiği akademik sınavlar ve mesleki yeterlilik testlerinde ölçme sürecinin güvenilirliğini artırmaktadır. Bu nedenle, eğitimcilerin ve sınav uygulayıcılarının BOBUT'u geleneksel test

yöntemlerine kıyasla daha etkili bir ölçme ve değerlendirme aracı olarak değerlendirmeleri önerilmektedir.

BOBUT'un etkin bir şekilde uygulanabilmesi için test geliştirme, uygulama ve değerlendirme süreçlerine yönelik kapsamlı bir eğitim alınması gerekmektedir. BOBUT, geleneksel testlerden farklı olarak madde havuzu tasarımı, madde kalibrasyonu, madde seçim algoritmaları, yetenek kestirim yöntemleri ve sonlandırma kuralları gibi teknik bileşenler içermektedir. Bu nedenle, test geliştiren eğitimcilerin, ölçme ve değerlendirme uzmanlarının ve sınav yöneticilerinin bu alanlarda yetkinlik kazanması büyük önem taşımaktadır. Özellikle, eğitimcilerin MTK ve test optimizasyon süreçleri hakkında bilgi sahibi olmaları, BOBUT'un etkili bir şekilde kullanılmasını sağlayacaktır.

BOBUT'un uzaktan sınav sistemlerinde kullanılması da önemli bir fırsat sunmaktadır. Örneğin, online yapılan akademik sınavlarda ve sertifikasyon programlarında BOBUT kullanılarak adayların yetenek seviyelerine uygun değerlendirmeler yapılabilir. Bu sayede, her adaya aynı soruların sorulması yerine, bireyselleştirilmiş ve güvenilir bir ölçme süreci sağlanabilir.

5.2.3. Kurumlara Öneriler

Nitelikli BOBUT uygulayıcılarının yetişmesi konusunda Millî Eğitim Bakanlığı'na önemli bir sorumluluk düşmektedir. Türkiye'deki eğitim sisteminin ölçme ve değerlendirme süreçlerini daha objektif ve kullanışlı hale getirmek adına, BOBUT'un eğitimde yaygınlaştırılması için hizmet içi eğitimlerin düzenlenmesi gerekmektedir. MEB, öğretmenler, ölçme ve değerlendirme merkezlerindeki uzmanlar ve eğitim yöneticileri için BOBUT testlerinin nasıl geliştirileceği, uygulanacağı ve değerlendirileceği konusunda özel eğitim programları oluşturabilir. Bu eğitimler, öğretmenlerin öğrenci seviyesine uygun bireyselleştirilmiş değerlendirmeler yapmasını sağlayarak eğitimde fırsat eşitliğine katkıda bulunacaktır.

Hizmet içi eğitimlerin içeriği, madde bankası oluşturma, MTK modellerinin kullanımı, adaptif test algoritmalarının işleyişi, BOBUT yazılımlarının kullanımı ve sonuçların analizi gibi temel konuları içerebilir. Ayrıca, öğretmenlerin ve ölçme uzmanlarının geleneksel sınavlardan BOBUT sistemine geçiş sürecini nasıl yönetebileceklerine dair rehberlik sunulmalıdır. Bu eğitimlerin, üniversiteler ve akademik kurumlarla iş birliği içinde yürütülmesi, alanında uzman

akademisyenlerin sürece dahil edilmesi ve uluslararası iyi uygulamaların incelenmesi önerilmektedir.

Ayrıca, MEB tarafından geliştirilecek merkezi sınavlarda BOBUT'un pilot uygulamalarla test edilmesi ve aşamalı olarak adaptif sınav modeline geçiş yapılması mümkündür. Özellikle LGS, bursluluk sınavları ve öğretmen yeterlilik sınavları gibi büyük ölçekli sınavlarda BOBUT'un kullanımı hem test süresini optimize edecek hem de ölçme doğruluğunu artıracaktır. MEB bünyesinde yer alan ölçme ve değerlendirme merkezlerinde BOBUT eğitimi almış uzmanların görev alması ve bu teknolojinin altyapısının oluşturulması, Türkiye'de ölçme ve değerlendirme süreçlerinin çağdaş bir yapıya kavuşmasına katkı sağlayacaktır.

Üniversiteler, özellikle ölçme ve değerlendirme, eğitim teknolojileri ve bilgisayar bilimleri alanlarında çalışan akademisyenler aracılığıyla BOBUT sisteminin Türkiye'de nasıl uygulanabileceği üzerine bilimsel araştırmalar yapmalı ve YÖK'ün politika belirleme süreçlerine katkıda bulunmalıdır. Ayrıca, üniversitelerde ölçme ve değerlendirme yüksek lisans ve doktora programlarında BOBUT üzerine dersler açılarak akademik altyapı güçlendirilmelidir. Bu kapsamda YÖK, BOBUT konusunda uzman yetiştirilmesi için akademik teşvikler sağlamalı ve üniversiteler ile iş birliği yaparak eğitim programları düzenlemelidir.

YÖK ve üniversiteler, pilot uygulamalar başlatarak, geniş ölçekli sınavlarda BOBUT'un nasıl uygulanabileceğini test etmelidir. Örneğin, üniversite içi hazırlık programlarında, yabancı dil yeterlilik sınavlarında ve ders bazlı değerlendirmelerde CABOBUT kullanımı aşamalı olarak yaygınlaştırılabilir. Pilot uygulamalar sonucunda elde edilen veriler analiz edilerek Türkiye'ye özgü bir BOBUT modeli geliştirilmelidir.

BOBUT'un geniş ölçekli sınavlarda uygulanabilmesi için teknik altyapının güçlendirilmesi şarttır. Üniversitelerde ve ÖSYM merkezlerinde güvenli sınav laboratuvarları kurulmalı, sınav salonları bilgisayar tabanlı test uygulamalarına uygun hale getirilmelidir.

KAYNAKÇA

- Akbaş, U., & Koğar, H. (2020). *Nicel araştırmalarda kayıp veriler ve uç değerler*. Pegem Akademi Yayıncılık.
- Algina, J., & Swaminathan, H. (2015). Psychometrics: Classical test theory. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (2nd ed., pp. 423–430). Elsevier Inc. <https://doi.org/10.1016/B978-0-08-097086-8.42070-2>
- Alotaibi, Y. A., & Selouani, S. (2010). Investigating the adaptation of Arabic speech recognition systems to foreign accented speakers. *10th International Conference on Information Science, Signal Processing and Their Applications (ISSPA 2010)*, 646–649. <https://doi.org/10.1109/ISSPA.2010.5605428>
- Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Prentice Hall.
- Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Addison Wesley Longman.
- Angoff, W. H. (1971). *The College Board Admissions Testing Program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Tests*. College Entrance Examination Board.
- Aybek, E. C. (2016). Concerto: Bilgisayar ortamında bireye uyarlanmış test uygulamaları için bir platform [Concerto: A platform for computerized adaptive testing applications]. *Journal of Measurement and Evaluation in Education and Psychology*, 7(2), 251–271.
- Aybek, E. C., & Çıkrıkçı, R. N. (2018). Kendini değerlendirme envanteri'nin bilgisayar ortamında bireye uyarlanmış test olarak uygulanabilirliği [The applicability of the self-assessment inventory as a computerized adaptive test]. *Turkish Psychological Counseling and Guidance Journal*, 8(50), 117–141.
- Aybek, E. C., & Demirtasli, R. N. (2017). Computerized adaptive test (CAT) applications and item response theory models for polytomous items. *International Journal of Research in Education and Science*, 3(2), 475–487.

- Babcock, B., & Weiss, D. J. (2009). Termination criteria in computerized adaptive tests: Variable-length CATs are not biased. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC conference on computerized adaptive testing* (Vol. 14).
- Baker, F. B. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation.
- Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. CRC Press.
- Bandalos, D. L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(1), 102–116. <https://doi.org/10.1080/10705511.2014.859510>
- Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). Wiley.
- Barrada, J. R., Olea, J., & Abad, F. J. (2008). Rotating item banks versus restricted maximum information item selection in computerized adaptive testing. *The Spanish Journal of Psychology*, 11(2), 618–625. <https://doi.org/10.1017/S1138741600004660>
- Belov, D. I., & Armstrong, R. D. (2010). Automatic detection of answer copying via Kullback-Leibler divergence and K-index. *Applied Psychological Measurement*, 34(6), 379–392. <https://doi.org/10.1177/0146621610370453>
- Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88(3), 588–606. <https://doi.org/10.1037/0033-2909.88.3.588>
- Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Addison-Wesley.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1), 29–51. <https://doi.org/10.1007/BF02291411>
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6(4), 431–444. <https://doi.org/10.1177/014662168200600405>

- Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24(1), 1–21. <https://doi.org/10.1080/08957347.2011.532417>
- Bridgeman, B. (2009). Experiences from large-scale computer-based testing in the USA. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment* (pp. 39–44). Office for Official Publications of the European Communities.
- Bringsjord, E. L. (2001). *Computerized-adaptive versus paper-and-pencil testing environments: An experimental analysis of examinee experience* (Doctoral dissertation). State University of New York at Albany.
- Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. ASCD.
- Brookhart, S. M., & Nitko, A. J. (2011). Strategies for constructing assessments of higher-order thinking skills. In G. Schraw & D. R. Robinson (Eds.), *Assessment of higher order thinking skills* (pp. 327–359). Information Age Publishing.
- Brown, G. T., & Hirschfeld, G. H. (2008). Students' conceptions of assessment: Links to outcomes. *Assessment in Education: Principles, Policy & Practice*, 15(1), 3–17. <https://doi.org/10.1080/09695940701876003>
- Brown, J. D. (2013). Classical test theory. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 337–349). Routledge.
- Brummel, B. J., & Drasgow, F. (2010). The effects of estimator choice and weighting strategies on confirmatory factor analysis with stratified samples. *Applied Multivariate Research*, 13(2), 113–132. <https://doi.org/10.22329/amr.v13i2.3019>
- Bugbee, A. C., Jr. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education*, 28(3), 282–299. <https://doi.org/10.1080/08886504.1996.10782176>
- Bulut, O., & Kan, A. (2012). Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Eğitim Araştırmaları-Eurasian Journal of Educational Research*, 49, 61–80. <https://files.eric.ed.gov/fulltext/EJ1059924.pdf>
- Butterfield, M. S. (2016). *Comparing item selection methods in computerized adaptive testing using the rating scale model* (Doctoral dissertation). University of Minnesota.
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage.

- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10
- Chang, H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20(3), 213–229. <https://doi.org/10.1177/014662169602000303>
- Chang, H. H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, 80(1), 1–20. <https://doi.org/10.1007/s11336-014-9401-5>
- Chang, H. H., & Ying, Z. (1999). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 23(3), 211–222. <https://doi.org/10.1177/01466219922031338>
- Chapman, R. (2022). Expected a posteriori scoring in PROMIS®. *Journal of Patient-Reported Outcomes*, 6(1), 59. <https://doi.org/10.1186/s41687-022-00468-5>
- Chen, S. K. (1996). *A comparison of maximum likelihood estimation and expected a posteriori estimation in computerized adaptive testing using the generalized partial credit model* (Doctoral dissertation). The University of Texas at Austin.
- Choi, S. W., & Swartz, R. J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement*, 33(6), 419–440. <https://doi.org/10.1177/0146621608327801>
- Choi, S. W., Grady, M. W., & Dodd, B. G. (2010). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement*, 70(1), 1–24. <https://doi.org/10.1177/0013164409344525>
- Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, 17(1), 31–44. <https://doi.org/10.1111/j.1745-3992.1998.tb00619.x>
- Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 355–386). Praeger.
- Cohen, R. J., Swerdlik, M. E., & Phillips, S. M. (1996). *Psychological testing and assessment: An introduction to tests and measurement*. Mayfield Publishing Co.

- Cömert, M. (2008). *Bireye uyarlanmış bilgisayar destekli ölçme ve değerlendirme yazılımı geliştirilmesi* [Development of a computerized adaptive testing software] (Unpublished master's thesis). Bahçeşehir University, Istanbul.
- Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Sage Publications.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>
- Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları* [Multivariate statistics for social sciences: SPSS and LISREL applications] (Vol. 2). Pegem Akademi.
- Davey, T., & Parshall, C. G. (1995). New algorithms for item selection and exposure control with computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, 20(4), 341–354. <https://doi.org/10.3102/10769986020004341>
- De Ayala, R. J. (1995). An investigation of the standard errors of expected a posteriori ability estimates. *Applied Psychological Measurement*, 19(2), 125–138. <https://doi.org/10.1177/014662169501900202>
- De Ayala, R. J. (2018). Item response theory and Rasch modeling. In B. B. Frey (Ed.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 145–163). Routledge.
- DeMars, C. (2010). *Item response theory*. Oxford University Press.
- Demir, S. (2019). *Çok kategorili bireyselleştirilmiş bilgisayarlı test uygulamalarının farklı madde seçim yöntemlerinde sonlandırma kuralları açısından incelenmesi* [Doktora tezi, Hacettepe Üniversitesi]. YÖK Tez Merkezi.
- DeVellis, R. F. (2006). Classical test theory. *Medical Care*, 44(11), S50–S59. <https://doi.org/10.1097/01.mlr.0000245426.10853.30>
- DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Sage.

- Doğan Şahin, M. (2017). *Gerçek ve üretilmiş veri setlerinde çok boyutlu bireyselleştirilmiş bilgisayarlı test uygulama sonuçlarının incelenmesi* [Doktora tezi, Hacettepe Üniversitesi]. YÖK Tez Merkezi.
- Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37(4), 281–306. <https://doi.org/10.1111/j.1745-3984.2000.tb01088.x>
- Ebenbeck, N., & Gebhardt, M. (2022). Simulating computerized adaptive testing in special education based on inclusive progress monitoring data. *Frontiers in Education*, 7, 945733. <https://doi.org/10.3389/feduc.2022.945733>
- Eggen, T. J. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, 23(3), 249–261. <https://doi.org/10.1177/01466219922031347>
- Eggen, T. J., & Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement*, 30(5), 379–393. <https://doi.org/10.1177/0146621606288026>
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory*. Psychology Press.
- Enders, C. K. (2022). *Applied missing data analysis*. Guilford Publications.
- Eroğlu, S., & Kelecioğlu, H. (2015). Bilgisayar ortamında bireye uyarlanmış testlerde sonlandırma kuralları ve yetenek kestirim yöntemlerinin ölçme kesinliği üzerindeki etkisi [The effect of termination rules and ability estimation methods on measurement precision in computerized adaptive testing]. *Eğitim ve Bilim*, 40(179), 169–182. <https://doi.org/10.15390/EB.2015.1535>
- Ertuna, L. (2020). *The effect of item selection and parameter estimation methods to the accuracy of pretest item parameters on online calibration in CAT* [Doktora tezi, Hacettepe Üniversitesi]. YÖK Tez Merkezi.
- Evans, J. J. (2010). *Comparability of examinee proficiency scores on computer adaptive tests using real and simulated data* (Doctoral dissertation). The State University of New Jersey, USA.

- Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39–50. <https://doi.org/10.1177/002224378101800104>
- Fraenkel, J. R., & Wallen, N. E. (2008). *How to design and evaluate research in education* (7th ed.). McGraw-Hill.
- Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2019). *How to design and evaluate research in education* (10th ed.). McGraw-Hill Education.
- Geerlings, H., Glas, C. A., & van der Linden, W. J. (2011). Modeling rule-based item generation. *Psychometrika*, 76(2), 337–359. <https://doi.org/10.1007/s11336-011-9204-x>
- Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2014). Development of the CAT-ANX: A computerized adaptive test for anxiety. *American Journal of Psychiatry*, 171(2), 187–194. <https://doi.org/10.1176/appi.ajp.2013.13020178>
- Glas, C. A. W., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, 27(4), 247–261. <https://doi.org/10.1177/0146621603027004001>
- Gleser, G. C., Cronbach, L. J., & Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika*, 30(4), 395–418. <https://doi.org/10.1007/BF02289531>
- Gorin, J. S., Dodd, B. G., Fitzpatrick, S. J., & Shieh, Y. Y. (2005). Computerized adaptive testing with the partial credit model: Estimation procedures, population distributions, and item pool characteristics. *Applied Psychological Measurement*, 29(6), 433–456. <https://doi.org/10.1177/0146621605280072>
- Gökcan, M. (2023). *İngilizce kelime bilgisi testinin (VST) bireyselleştirilmiş bilgisayarlı test olarak uygulanabilirliğinin incelenmesi* [Doktora tezi, Hacettepe Üniversitesi]. YÖK Tez Merkezi.
- Gür, R., & Gülleroğlu, H. D. (2020). Bireye uyarlanmış testlerde madde kullanım sıklığı kontrol yöntemlerinin farklı koşullarda ölçme duyarlılığına ve test güvenliğine etkisi [The effect of item exposure control methods on measurement precision and test security in

- computerized adaptive testing under different conditions]. *Eğitim ve Bilim*, 45(202), 1–20. <https://doi.org/10.15390/EB.2020.8454>
- Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Routledge.
- Hambleton, R. K. (1994). Item response theory: A broad psychometric framework for measurement advances. *Psicothema*, 6(3), 535–556.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38–47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer-Nijhoff.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.
- Han, K. C. T. (2018). Conducting simulation studies for computerized adaptive testing using SimulCAT: An instructional piece. *Journal of Educational Evaluation for Health Professions*, 15, 20. <https://doi.org/10.3352/jeehp.2018.15.20>
- Hausler, J., & Sommer, M. (2008). The effect of success probability on test economy and self-confidence in computerized adaptive tests. *Psychology Science*, 50(1), 75–88.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179–185. <https://doi.org/10.1007/BF02289447>
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Huo, Y. (2009). *Variable-length computerized adaptive testing: Adaptation of the a-stratified strategy in item selection with content balancing* (Doctoral dissertation). University of Illinois at Urbana-Champaign.
- Iseri, A. I. (2002). *Assessment of students' mathematics achievement through computer adaptive testing procedures* [Doctoral dissertation, Middle East Technical University]. National Thesis Center.

- Jonassen, D. H., & Grabowski, B. L. (2012). *Handbook of individual differences, learning, and instruction*. Routledge.
- Kalender, İ. (2011). *Effects of different computerized adaptive testing strategies on recovery of ability* [Doctoral dissertation, Middle East Technical University]. National Thesis Center.
- Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement*, 39(3), 167–188. <https://doi.org/10.1177/0146621614559564>
- Kaplan, R. M., & Saccuzzo, D. P. (2017). *Psychological testing: Principles, applications, and issues* (9th ed.). Cengage Learning.
- Kaptan, F. (1993). *Yetenek kestiriminde adaptive (bireyselleştirilmiş) test uygulaması ile geleneksel kâğıt-kalem testi uygulamasının karşılaştırılması* [Doktora tezi, Hacettepe Üniversitesi]. YÖK Tez Merkezi.
- Karasar, N. (2010). *Bilimsel araştırma yöntemi* [Scientific research method] (21st ed.). Nobel Yayın Dağıtım.
- Kezer, F. (2020). *Bilgisayar ortamında bireye uyarlanmış test (BOBUT) uygulamaları ve yetenek kestirim yöntemleri* [Computerized adaptive testing (CAT) applications and ability estimation methods]. Pegem Akademi Yayıncılık.
- Kezer, F., & Koç, E. (2014). Bilgisayar ortamında bireye uyarlanmış testlerde farklı başlatma ve sonlandırma kuralları ile yetenek kestirim yöntemlerinin karşılaştırılması [A comparison of different starting and termination rules and ability estimation methods in computerized adaptive testing]. *Eğitim ve Bilim*, 39(176), 1–14. <https://doi.org/10.15390/EB.2014.2293>
- Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling*, 18(2), 212–228. <https://doi.org/10.1080/10705511.2011.557337>
- Kimura, T. (2017). The impacts of computer adaptive testing from a variety of perspectives. *Journal of Educational Evaluation for Health Professions*, 14, 19. <https://doi.org/10.3352/jeehp.2017.14.19>

- Kingsbury, G. G., & Hauser, C. (2004). Computerized adaptive testing and "No Child Left Behind". *Northwest Evaluation Association*.
- Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, 2(4), 359–375. https://doi.org/10.1207/s15324818ame0204_6
- Kingsbury, G. G., & Houser, R. L. (1988, April). *A comparison of achievement level estimates from computerized adaptive testing and paper-and-pencil testing*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Kocalevent, R. D., Rose, M., Becker, J., Walter, O. B., Fliege, H., Bjorner, J. B., & Klapp, B. F. (2009). An evaluation of patient-reported outcomes found computerized adaptive testing was efficient in assessing stress perception. *Journal of Clinical Epidemiology*, 62(3), 278–287. <https://doi.org/10.1016/j.jclinepi.2008.03.009>
- Koğar, E. Y. (2021). Comparison of testlet effect on parameter estimates using different item response theory models. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 12(3), 254–266. <https://doi.org/10.21031/epod.948227>
- Koğar, H. (2021). *R ile geçerlik ve güvenirlik analizleri*. Pegem Akademi.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer.
- Kosan-Aytuğ, M. A. (2013). *Tıp eğitiminde gelişim sınavı soru bankası oluşturulması ve benzetim verileri ile bilgisayar uyarlamalı test uygulaması* [Doktora tezi, Hacettepe Üniversitesi]. YÖK Tez Merkezi.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2(3), 151–160. <https://doi.org/10.1007/BF02288391>
- Linn, R. L. (1989). *Educational measurement* (3rd ed.). American Council on Education and Macmillan Publishing Company.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

- Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computerized adaptive sequential testing. *Journal of Educational Measurement*, 35(3), 229–249. <https://doi.org/10.1111/j.1745-3984.1998.tb00537.x>
- Lunz, M. E., & Bergstrom, B. A. (1994). An empirical study of computerized adaptive test administration conditions. *Journal of Educational Measurement*, 31(3), 251–263. <https://doi.org/10.1111/j.1745-3984.1994.tb00446.x>
- Lunz, M. E., & Stahl, J. A. (1998). Patterns of item exposure using a randomized CAT algorithm. In *Annual meeting of the National Council on Measurement in Education*, San Diego, CA.
- MacCann, C., & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: Theory and data. *Emotion Review*, 5(4), 382–395. <https://doi.org/10.1177/1754073908095137>
- MacDonald, P. (2003). *Computer-adaptive test for measuring personality factors using item response theory* (Doctoral dissertation). National Library of Canada, Ottawa.
- Magis, D., Beland, S., Raiche, G., & Magis, M. D. (2015, April). Package ‘difR’.
- Magis, D., & Raïche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software*, 48(8), 1–31. <https://doi.org/10.18637/jss.v048.i08>
- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3), 519–530. <https://doi.org/10.1093/biomet/57.3.519>
- McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military design. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 223–236). Academic Press.
- McMillan, J. H., & Hellsten, L. (2010). *Classroom assessment: Principles and practice for effective standards-based instruction*. Pearson Education Canada.
- Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement*, 23(3), 187–194. <https://doi.org/10.1177/01466219922031329>

- Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5–11. <https://doi.org/10.3102/0013189X018002005>
- Miller, D. V. (2003). *Assessment of student achievement: A comparative study of student achievement using paper and pencil assessment and computerized adaptive testing (CAT)* (Doctoral dissertation). Wayne State University, Michigan.
- Milli Eğitim Bakanlığı. (2018). *Ortaöğretim matematik dersi öğretim programı* [Secondary education mathematics curriculum]. Ankara: Talim ve Terbiye Kurulu Başkanlığı.
- Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56(2), 177–196. <https://doi.org/10.1007/BF02294457>
- Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement*, 13(1), 57–75. <https://doi.org/10.1177/014662168901300106>
- Mullis, I. V., Martin, M. O., Kennedy, A., Trong, K., & Sainsbury, M. P. (2015). *Assessment frameworks*. TIMMS and PIRLS International Study Center, Boston College.
- Nitko, A. J. (2013). *Educational assessment of students* (7th ed.). Pearson.
- Nunnally, J. C., & Bernstein, I. H. (1994). The assessment of reliability. In J. C. Nunnally & I. H. Bernstein (Eds.), *Psychometric theory* (3rd ed., pp. 248–292). McGraw-Hill.
- O'Boyle, E. H., Jr., & Aguinis, H. (2012). The best and the rest: Revisiting the norm of normality of individual performance. *Personnel Psychology*, 65(1), 79–119. <https://doi.org/10.1111/j.1744-6570.2011.01239.x>
- OECD. (2019). *PISA 2018 assessment and analytical framework*. OECD Publishing. <https://doi.org/10.1787/b25efab8-en>
- Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). Sage.
- Oswald, F. L., Shaw, A., & Farmer, W. L. (2014). Comparing simple scoring with IRT scoring of personality measures. *Applied Psychological Measurement*, 39(2), 144–154. <https://doi.org/10.1177/0146621614559517>
- Özbaşı, D., & Demirtaşlı, N. (2016). Bilgisayar okuryazarlığı testinin bilgisayar ortamında bireye uyarlanmış test olarak geliştirilmesi [Development of a computerized adaptive

- test for computer literacy]. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 7(1), 220–234. <https://doi.org/10.21031/epod.27718>
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press.
- Penfield, R. D. (2006). Applying Bayesian item selection approaches to adaptive tests using polytomous items. *Applied Measurement in Education*, 19(1), 1–20. https://doi.org/10.1207/s15324818ame1901_1
- Pitkin, A. K., & Vispoel, W. P. (2001). Differences between self-adapted and computerized adaptive tests: A meta-analysis. *Journal of Educational Measurement*, 38(3), 235–247. <https://doi.org/10.1111/j.1745-3984.2001.tb01126.x>
- Railsback, S. F., & Grimm, V. (2019). *Agent-based and individual-based modeling: A practical introduction* (2nd ed.). Princeton University Press.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4(3), 207–230. <https://doi.org/10.2307/1164671>
- Reckase, M. D. (2006). Multidimensional item response theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 607–642). Elsevier.
- Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling*, 52(2), 127–141.
- Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment*, 81(2), 93–103. https://doi.org/10.1207/S15327752JPA8102_01
- Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons.
- Rutten, N., Van Joolingen, W. R., & van der Veen, J. T. (2012). The learning effects of computer simulations in science education. *Computers & Education*, 58(1), 136–153. <https://doi.org/10.1016/j.compedu.2011.08.017>

- Salas, E., Wildman, J. L., & Piccolo, R. F. (2009). Using simulation-based training to enhance management education. *Academy of Management Learning & Education*, 8(4), 559–573. <https://doi.org/10.5465/AMLE.2009.47785471>
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34(4), 1–97.
- Samejima, F. (2016). Graded response models. In W. J. van der Linden (Ed.), *Handbook of item response theory* (pp. 95–107). Chapman and Hall/CRC.
- Sarı, H. İ., & Raborn, A. (2018). What information works best?: A comparison of routing methods. *Applied Psychological Measurement*, 42(6), 499–515. <https://doi.org/10.1177/0146621617752990>
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall/CRC.
- Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, 84(1), 1–66. <https://doi.org/10.1037/0033-295X.84.1.1>
- Scullard, M. G. (2007). *Application of item response theory based computerized adaptive testing to the Strong Interest Inventory* [Doctoral dissertation]. University of Minnesota.
- Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, 61(2), 331–354. <https://doi.org/10.1007/BF02294343>
- Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (Vol. 1, pp. 429–438). Elsevier.
- Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29(7), 4–14. <https://doi.org/10.3102/0013189X029007004>
- Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, 30(4), 298–321. <https://doi.org/10.1177/0146621605285517>
- Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28(3), 237–247. <https://doi.org/10.1111/j.1745-3984.1991.tb00359.x>

- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7(2), 201–210. <https://doi.org/10.1177/014662168300700208>
- Sulak, S. (2013). *Bireyselleştirilmiş bilgisayarlı test uygulamalarında kullanılan madde seçme yöntemlerinin karşılaştırılması* [Doktora tezi, Hacettepe Üniversitesi]. YÖK
- Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2006). Assessing the fit of item response theory models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 683–718). Elsevier.
- Şenel, S. (2017). *Bilgisayar ortamında bireye uyarlanmış testlerin görme engelli öğrencilere uygunluğunun incelenmesi* [Yayımlanmamış doktora tezi, Ankara Üniversitesi].
- Şimşek, A. S. (2017). *Becerilere Güven Mesleki İlgi Envanterinin uyarlanması ve bilgisayarlı bireyselleştirilmiş test uygulamasının geliştirilmesi* [Doktora tezi, Ankara Üniversitesi]. YÖK Tez Merkezi.
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.
- Tat, O. (2019). *Açıklayıcı madde tepki modellerinin bilgisayar ortamında bireye uyarlanmış testlerde kullanımı* [Doktora tezi, Hacettepe Üniversitesi]. YÖK Tez Merkezi.
- Tercan, S. S. (2010). *Çevrimiçi uyarlamalı bir test aracı geliştirilmesi, uygulanması ve öğrenci görüşlerinin belirlenmesi* [Yüksek lisans tezi, Sakarya Üniversitesi]. YÖK Tez Merkezi.
- Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 101–133). Lawrence Erlbaum Associates.
- Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 85–152). Routledge.
- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51(4), 567–577. <https://doi.org/10.1007/BF02295596>
- Thompson, N. A., & Weiss, D. A. (2019). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation*, 16(1), 1–12. <https://doi.org/10.7275/5vvy-8613>

- Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement*, 8(2), 63–70. <https://doi.org/10.1111/j.1745-3984.1971.tb00915.x>
- Tsao, Y., Isotani, R., Kawai, H., & Nakamura, S. (2010). An environment structuring framework to facilitating suitable prior density estimation for MAPLR on robust speech recognition. *2010 7th International Symposium on Chinese Spoken Language Processing* (pp. 29–32). <https://doi.org/10.1109/ISCSLP.2010.5684880>
- Turgut, M. F., & Baykul, Y. (2010). *Eğitimde ölçme ve değerlendirme* (Vol. 2). Pegem Akademi.
- Ueno, M., & Songmuang, P. (2010). Computerized adaptive testing based on decision tree. *2010 10th IEEE International Conference on Advanced Learning Technologies* (pp. 1–3). <https://doi.org/10.1109/ICALT.2010.58>
- University of Cambridge Psychometrics Centre. (2024). *Concerto platform*. Retrieved from <https://www.psychometrics.cam.ac.uk/concerto>
- Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 45, 1-67.
- van der Linden, W. J. (2005). *Linear models for optimal test design*. Springer.
- van der Linden, W. J. (Ed.). (2017). *Handbook of item response theory: Volume 3: Applications*. CRC Press.
- van der Linden, W. J., & Glas, C. A. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Kluwer Academic.
- van der Linden, W. J., & Glas, C. A. (Eds.). (2010). *Elements of adaptive testing*. Springer.
- van der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 1–28). Springer.
- van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. Glas (Eds.), *Elements of adaptive testing* (pp. 3–30). Springer.

- Veerkamp, W. J., & Berger, M. P. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics*, 22(2), 203–226. <https://doi.org/10.3102/10769986022002203>
- Veldkamp, B. P., & Verschoor, A. J. (2019). Robust computerized adaptive testing. In B. P. Veldkamp & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 291–305). Springer. https://doi.org/10.1007/978-3-030-18480-3_15
- Wainer, H. (2000). Introduction and history. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 1–21). Routledge.
- Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge.
- Wang, T., Hanson, B. A., & Lau, C. M. A. (1999). Reducing bias in CAT trait estimation: A comparison of approaches. *Applied Psychological Measurement*, 23(3), 263–278. <https://doi.org/10.1177/01466219922031347>
- Webb, N. (2005). *Depth-of-knowledge levels for four content areas*. Presentation to the Florida Education Research Association, 50th Annual Meeting, Miami, Florida.
- Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, 53(6), 774–789. <https://doi.org/10.1037/0022-006X.53.6.774>
- Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development*, 37(2), 70–84. <https://doi.org/10.1080/07481756.2004.11909753>
- Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21(4), 361–375. <https://doi.org/10.1111/j.1745-3984.1984.tb01040.x>
- Weiss, D. J., & Sahin, A. (2024). *Computerized adaptive testing: From concept to implementation*. Guilford Publications.
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. MESA Press.
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8(2), 125–145. <https://doi.org/10.1177/014662168400800201>

- Zhang, Z. (2016). Multiple imputation with multivariate imputation by chained equation (MICE) package. *Annals of translational medicine*, 4(2), 30. <https://doi.org/10.3978/j.issn.2305-5839.2015.12.63>
- Zhang, J., & Li, J. (2016). Monitoring items in real time to enhance CAT security. *Journal of Educational Measurement*, 53(2), 131–151. <https://doi.org/10.1111/jedm.12104>
- Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. National Defense Headquarters.



EK-1.Kazanımlar

1. Görsellerden ve başlıktan hareketle okuyacağı metnin konusunu tahmin eder.
2. **Okuduklarını özetler.**
3. Metnin konusunu belirler.
4. **Metnin ana fikrini/ana duygusunu belirler.**
5. **Metindeki yardımcı fikirleri belirler.**
6. **Metinle ilgili soruları cevaplar.**
7. Metinle ilgili sorular sorar.
8. Metindeki hikâye unsurlarını belirler.
9. **Metnin içeriğini yorumlar.**
10. **Metnin içeriğine uygun başlık/başlıklar belirler.**
11. **Metinde ele alınan sorunlara farklı çözümler üretir.**
12. **Metinler arasında karşılaştırma yapar.**
13. **Metindeki gerçek ve kurgusal unsurları ayırt eder.**
14. **Metinde önemli noktaların vurgulanış biçimlerini kavrar.**
15. Okudukları ile ilgili çıkarımlarda bulunur
16. **Metin türlerini ayırt eder.**
17. **Görsellerle ilgili soruları cevaplar.**
18. **Grafik, tablo ve çizelgeyle sunulan bilgileri yorumlar.**
19. **Metindeki anlatım biçimlerini belirler.**
20. **Metinde kullanılan düşünceyi geliştirme yollarını belirler.**
21. **Metindeki iş ve işlem basamaklarını kavrar.**
22. **Bağlamdan hareketle bilmediği kelime ve kelime gruplarının anlamını tahmin eder.**
23. **Deyim ve atasözlerinin metne katkısını belirler.**
24. **Metni oluşturan unsurlar arasındaki geçiş ve bağlantı ifadelerinin anlama olan katkısını değerlendirir.**
25. **Metindeki söz sanatlarını tespit eder.**
26. **Anlatım bozukluklarını tespit eder.**
27. **Kelimelerin eş anlamlılarını bulur.**
28. **Kelimelerin zıt anlamlılarını bulur.**
29. **Eş sesli kelimelerin anlamlarını ayırt eder.**
30. **Okuduğu metindeki gerçek, mecaz ve terim anlamlı sözcükleri ayırt eder**

EK-2. Madde, Kazanım ve Bilişsel Alan Örnekleri

Örnek-1.

3. Öznel anlatım, kişisel görüşleri içerir. Nesnel anlatım ise kişisel görüşlerden uzak, kanıtlanabilir yargılara dayalı olur.

Bu açıklamaya göre, aşağıdaki metinlerden hangisinde öznel bir anlatıma başvurulmuştur?

- A. Fransız ve Ruslardan oluşan bir grup araştırmacı, 30.000 yıl öncesinden kalma bir virüs keşfetti. Araştırmacılar bu virüsü Sibirya'nın kuzeyindeki buzullardan, yüzeyin 30 metre altından çıkardılar.
- B. İtalyan besteci Antonio Vivaldi, "Dört Mevsim" adlı eserini 1700'lü yılların başında bir şiirin esinlenerek bestelemiş. Eser, dört keman konçertosundan oluşuyor. Bu konçertoların her birinde de üç bölüm var.
- C. Anadolu sığla ağacı, dünyada yalnızca Türkiye'de yaşayan bir ağaç türü. Muğla'nın içlerinden Fethiye, Köyceğiz, Milas ve Marmaris'e bu ağacın oluşan ormanlar var. Ancak bu ağacın yaşam alanı günümüzde daralmış durumda.
- D. İsrail'de bir şirket çok küçük bir yazıcı üretti. Cepte bile kolaylıkla taşınabilecek kadar küçük olan bu ilginç yazıcı, gücünü şarj edilebilir bir pilinden alıyor. Yazıcı ayrıca akıllı telefon ve bilgisayarlara kolaylıkla bağlanabiliyor.

- **Kazanım:** Metindeki anlatım biçimlerini belirler. (Kazanım 19)
- **Yenilenmiş Bloom Taksonomisi:** *Anlama Düzeyi:* Soru, "öznel" ve "nesnel" anlatımı ayırt etmeyi gerektiriyor, bu da kavramlar arasındaki farkı anlamayı içerir.
- **Webb Taksonomisi:** *Düzy 2-Beceriler ve Kavramlar:* Bu düzey, bilgi sınıflandırma veya açıklama gibi daha derin kavrayışları içerir. Öznel ve nesnel anlatımı karşılaştırmak, kavramsal bir beceridir.
- **PISA Bilişsel Düzeyi:** *Entegrasyon ve Yorumlama:* Bu soru, metinler arasında bir anlam bağlantısı kurmayı ve verilen bilgiyi yorumlamayı gerektirir.
- **PIRLS Bilişsel Düzeyi:** *Fikirleri ve Bilgileri Yorumlama ve Birleştirme:* Metinlerde verilen ifadeleri yorumlama ve öznel-anlatım farkını birleştirerek bir sonuca ulaşma bu düzeye karşılık gelir.

Örnek-2.

8. 1957'den bugüne, iletişim ve uzay keşfi gibi farklı amaçlarla uzaya on binin üzerinde uydu fırlatıldı. Üç binin üzerinde uydu hâlâ aktif durumdayken, diğerleri ömürlerini tamamladıkları için işlevsiz bir şekilde Dünya yörüngesinde dolaşmaktadır. İşlevsiz hale gelen bu uydular uzay çöplüğünü oluşturuyor. Saniyede 8 ila 10 kilometre hızla hareket eden bu uzay çöpleri, görevlerine devam eden uydular için büyük bir tehdit oluşturuyor. Öyle ki 10 cm büyüklüğündeki bir uzay çöpünün, görev halindeki bir uyduya çarpması tehlikeli binlerce yeni uzay çöpünün daha ortaya çıkmasına yol açabilir.

Bu metinde "uzay çöpleri" ile ilgili aşağıdakilerden hangisine değinilmemiştir?

- A. Nasıl temizleneceğine
- B. Etkilerinin ne olduğuna
- C. Hareket hızına
- D. Oluşma nedenlerine

- **Kazanım:** Metinle ilgili soruları cevaplar. (Kazanım 6)
- **Yenilenmiş Bloom Taksonomisi:** *Anlama Düzeyi:* Metni okuyarak "uzay çöpleri" konusundaki bilgileri anlamayı ve metindeki eksik bilginin ne olduğunu tespit etmeyi gerektiriyor.
- **Webb Taksonomisi:** *Düzy 2-Beceriler ve Kavramlar:* Soru, verilen bilgileri analiz etmeyi ve eksik bir noktayı belirlemeyi gerektiriyor. Bu, temel bilgiye ek olarak bir kavramsal anlayış düzeyini içerir.
- **PISA Bilişsel Düzeyi:** *Erişim ve Bilgi Getirme:* Metindeki açıkça belirtilmiş bilgileri seçme ve hatırlama becerisine dayanır.
- **PIRLS Bilişsel Düzeyi:** *Odaklanma ve Açıkça Belirtilen Bilgiyi Getirme:* Metinde açıkça belirtilen bilgilere odaklanmayı ve bunları anlamayı içerir.

Örnek-3.

I. Metin

Natron Gölü, Tanzanya'nın kuzeyinde bulunur. Bazık olan gölün sıcaklığı genellikle kırk derecenin üzerindedir. Bu durum gölde buharlaşmanın çok yoğun olmasına neden olur. Göl suyu buharlaştıkça sodyum karbonatın bir türevi olan natron adı verilen bir karışım açığa çıkar. Göl de adını buradan almıştır. Buharlaşmanın yoğun olduğu bölgelerde fotosentez yapan mikroorganizmalar yaşar. Bu mikroorganizmaların sahip olduğu kırmızı renk pigmentleri gölün kırmızı renkte görünmesine neden olur. Göl çok fazla canlı türüne ev sahipliği yapmamaktadır. Üreme döneminde göle gelen flamingolar, yüksek sıcaklık ve tuz oranına dayanıklı bazı balık türleri gölde yaşamaktadır.

II. Metin

Van Gölü, ülkemizde bulunan en büyük göldür. Bu gölün suyunda çok miktarda sodyum karbonat yani soda adı verilen kimyasal bir madde bulunur. Bu nedenle gölde az sayıda canlı türü yaşayabiliyor. Sazangillerden olan inci kefalı gölde yaşayan canlılardan biridir. Ayrıca Van Gölü'nde dört küçük adacık vardır. Gölün çevresinde ve bazı adalarda bulunan tarihi eserler nedeniyle bölgeye her yıl akın akın turist gelmektedir.

Bu iki metinde bahsedilen göllerin ortak yönü aşağıdakilerden hangisidir?

- A. İçlerinde çok az canlının yaşaması
- B. Bulundukları ülkelerin en büyük gölü olmaları
- C. Sıcaklıklarının çok yüksek olmaları
- D. Her yıl birçok turist tarafından ziyaret edilmeleri

- **Kazanım:** Metinler arasında karşılaştırma yapar. (Kazanım 12)
- **Yenilenmiş Bloom Taksonomisi:** *Analiz Düzeyi:* İki metin arasındaki benzerlikleri belirlemek, metinlerin içeriğini inceleyerek anlamlı ilişkiler kurmayı gerektirir.
- **Webb Taksonomisi:** *Düzye 3-Stratejik Düşünme:* Bu düzey, metinler arasında anlamlı karşılaştırmalar yapmayı ve çıkarımlarda bulunmayı içerir. Bu tür bir karşılaştırma stratejik düşünme gerektirir.
- **PISA Bilişsel Düzeyi:** *Entegrasyon ve Yorumlama:* İki metindeki bilgileri birleştirerek ortak özellikleri tespit etmek, entegre düşünmeyi ve yorumlamayı gerektirir.
- **PIRLS Bilişsel Düzeyi:** *Fikirleri ve Bilgileri Yorumlama ve Birleştirme:* Bu düzey, iki farklı metindeki bilgileri bir araya getirerek karşılaştırma yapmayı ve ortak özellikleri yorumlamayı içerir.

Örnek-4.

45.



Yukarıdaki görselde, Türkiye'nin 2015 yılına ait PISA uygulaması ve sonuçlarına yer verilmiştir. Buna göre, görselden aşağıdakilerden hangisi çıkanamaz?

- A. OECD ülkelerinin de bulunduğu toplam 72 ülkenin katıldığı PISA uygulamasına Türkiye'den 5895 öğrenci katılmıştır.
- B. Ege Bölgesi, PISA sınavlarındaki sıralamada diğer bölgeleri geçerek Türkiye'nin en başarılı bölgesi olmuştur.
- C. Türkiye'nin katıldığı PISA uygulamasında başarı sıralamasının sürekli düştüğü görülmektedir.
- D. Türkiye, 2015 PISA uygulamasında matematik ve okuma alanlarında OECD ülkelerinin ortalamasının altında kalmıştır.

- **Kazanım:** Görsellerle ilgili soruları cevaplar. (Kazanım 17)
- **Yenilenmiş Bloom Taksonomisi:** *Analiz Düzeyi:* Soru, görselde verilen bilgileri analiz etmeyi, farklı unsurları anlamayı ve çıkarım yapmayı gerektiriyor. Özellikle görselden ulaşılamayan bilgiyi tespit etmek, analitik düşünmeyi ifade eder.
- **Webb Taksonomisi:** *Düzye 3-Stratejik Düşünme:* Bu düzey, birden fazla veri kaynağıyla çalışmayı, bunları karşılaştırmayı ve karmaşık bilgi yapılarından anlamlı çıkarımlar yapmayı gerektirir. Soru, görseldeki bilgileri dikkatlice inceleyerek ulaşılamayan seçeneği belirlemeyi gerektirir
- **PISA Bilişsel Düzeyi:** *Entegrasyon ve Yorumlama:* Bu düzey, görselde sunulan bilgilerin birleştirilerek yorumlanmasını ve verilen bilgilere dayalı olarak çıkarımlar yapılmasını içerir.
- **PIRLS Bilişsel Düzeyi:** *Fikirleri ve Bilgileri Yorumlama ve Birleştirme:* Bu düzey, görselde verilen bilgi parçalarını yorumlamayı, anlamlı bir şekilde birleştirmeyi ve soruya uygun bir sonuca ulaşmayı içerir.

Örnek-5.

54. Atasözleri, atalarımızın uzun denemelerine dayanan yargılarını, tecrübelerini kısaca özetleyen, kalıplaşmış ve halka mal olmuş özlü sözlerdir. (I) Bazı atasözleri sosyal ilişkiler üzerine görüş bildirir. (II) Bazı atasözleri ise uzun gözlemlere dayanarak doğa olaylarının günlük yaşam üzerindeki etkilerini anlatır. (III) Denemelere ve mantığa dayanarak doğrudan doğruya öğüt veren atasözleri de vardır. (IV) Bir de birtakım, töre ve gelenekleri bildiren atasözlerimiz de mevcuttur.

Bu metinde numaralanmış cümlelerdeki açıklamalara aşağıdaki atasözlerinden hangisi örnek olarak gösterilemez?

- A. I. cümle: Komşu, komşunun külüne muhtaçtır.
- B. II. cümle: Mart kapıdan baktırır, kazma kürek yaktırır.
- C. III. cümle: Bugünün işini yarına bırakma.
- D. IV. cümle: Sakla samanı gelir zamanı.

- **Kazanım:** Deyim ve atasözlerinin metne katkısını belirler. (Kazanım 23)
- **Yenilenmiş Bloom Taksonomisi:** *Anlama Düzeyi:* Soru, atasözlerinin anlamlarını ve metinle ilişkilerini anlayarak uygun seçeneği belirlemeyi gerektiriyor. Bu da kavramları tanıma ve açıklama becerisini içerir.
- **Webb Taksonomisi:** *Düzy 2-Beceriler ve Kavramlar:* Atasözlerinin açıklamalarla eşleştirilmesi, bilgi sınıflandırma ve kavramsal anlayış gerektirir.
- **PISA Bilişsel Düzeyi:** *Erişim ve Bilgi Getirme:* Metinde verilen atasözleri ile ilgili bilgileri bulmak ve bu bilgileri açıklamalarla eşleştirerek kullanmak bu düzeye karşılık gelir.
- **PIRLS Bilişsel Düzeyi:** *Odaklanma ve Açıkça Belirtilen Bilgiyi Getirme:* Metin içinde açıkça belirtilen açıklamalarla ilgili doğru atasözlerini belirleme ve bu bilgileri eşleştirme becerisi PIRLS'in bu düzeyine karşılık gelir.

Örnek-6.

113. Tarık: Bilim kurgu edebiyatı, geleceğe dair tahminlerde bulunmanın ötesinde, mevcut toplumsal sorunları sorgulama aracı olarak da işlev görüyor. Yazarlar, uzay gemileri ve yapay zekalar aracılığıyla derin felsefi meselelere değiniyorlar.

Leyla: Kesinlikle katılıyorum ve eklemek istiyorum. Bilim kurgu eserleri, teknolojinin insan üzerindeki etkilerini de tartışmaya açıyor. Bu tür eserler, insanlığın geleceği hakkında kritik soruları gündeme getirerek toplumsal bir bilinç yaratıyor.

Bu konuşmaya göre aşağıdakilerden hangisi söylenemez?

- A. Tarık, bilim kurgu edebiyatının geleceğe dair tahminlerde bulunmaması gerektiğini düşünüyor.
- B. Leyla, bilim kurgunun teknolojinin insan hayatındaki rolünü sorguladığını belirtiyor.
- C. Tarık, bilim kurgu eserlerinin farklı alandaki konuları ele aldığını ifade ediyor.
- D. Leyla, bilim kurgu edebiyatının toplumsal farkındalığı artırdığını öne sürüyor.

- **Kazanım:** Metnin içeriğini yorumlar. (Kazanım 9)
- **Yenilenmiş Bloom Taksonomisi:** *Değerlendirme Düzeyi:* Soruda verilen konuşmaları analiz ederek doğru çıkarımları belirleme ve hangi ifadenin söylenemeyeceğini değerlendirme becerisini gerektiriyor.
- **Webb Taksonomisi:** *Düzy 3-Stratejik Düşünme:* Konuşmalardaki ifadeleri analiz ederek farklı seçenekler arasında bir sonuca ulaşmayı içerir. Bu, stratejik düşünmeyi ve eleştirel değerlendirmeyi gerektirir.
- **PISA Bilişsel Düzeyi:** *Yansıtma ve Değerlendirme:* Metindeki ifadelerin doğruluğunu değerlendirmek, farklı fikirleri karşılaştırmak ve çıkarımlarda bulunmak bu düzeye karşılık gelir.
- **PIRLS Bilişsel Düzeyi:** *İçeriği İnceleme ve Değerlendirme:* Metindeki ifadelerin eleştirel bir şekilde değerlendirilmesi, dil ve içerik açısından yorumlanması bu düzeyle uyumludur.

EK-3. Veri1-Veri6 Karşılaştırmasına Ait Mantel-Haenszel DIF Analizi Sonuçları

Madde	χ^2	P	DeltaMH	Etki	Madde	χ^2	P	DeltaMH	Etki	Madde	χ^2	P	DeltaMH	Etki
M1	0,539	0,463	-0,325	A	M53	0,958	0,328	0,278	A	M105	0,556	0,456	-0,042	A
M2	0,056	0,812	0,098	A	M54	1,490	0,222	0,530	A	M106	0,724	0,395	0,026	A
M3	0,002	0,962	0,042	A	M55	0,675	0,411	0,458	A	M107	0,149	0,699	0,065	A
M4	0,290	0,590	-0,295	A	M56	0,177	0,674	0,157	A	M108	1,560	0,212	0,070	A
M5	0,837	0,360	0,281	A	M57	0,376	0,540	0,180	A	M109	1,389	0,239	-0,029	A
M6	0,363	0,547	0,179	A	M58	0,193	0,661	0,133	A	M110	0,003	0,958	0,064	A
M7	0,593	0,441	-0,227	A	M59	0,059	0,809	0,082	A	M111	0,296	0,586	0,025	A
M8	2,272	0,132	-0,827	A	M60	0,739	0,390	0,109	A	M112	0,706	0,401	-0,069	A
M9	0,003	0,960	-0,001	A	M61	1,514	0,219	-0,172	A	M113	0,018	0,895	0,057	A
M10	0,016	0,898	0,057	A	M62	0,023	0,879	-0,040	A	M114	0,222	0,637	0,061	A
M11	0,106	0,745	-0,149	A	M63	0,060	0,807	-0,144	A	M115	0,317	0,574	0,090	A
M12	2,763	0,097	0,530	A	M64	0,110	0,740	0,051	A	M116	0,047	0,829	0,026	A
M13	1,830	0,176	0,486	A	M65	0,055	0,815	0,201	A	M117	3,272	0,071	-0,081	A
M14	0,239	0,625	0,155	A	M66	2,966	0,085	-0,110	A	M118	0,332	0,565	-0,085	A
M15	0,040	0,842	0,081	A	M67	1,256	0,262	-0,101	A	M119	0,586	0,444	-0,030	A
M16	0,474	0,491	0,203	A	M68	0,085	0,770	0,212	A	M120	0,283	0,595	0,109	A
M17	0,005	0,944	0,040	A	M69	0,094	0,759	0,279	A	M121	1,085	0,298	0,072	A
M18	2,509	0,113	0,453	A	M70	0,343	0,558	-0,026	A	M122	0,185	0,667	-0,031	A
M19	0,034	0,855	0,068	A	M71	0,221	0,638	0,202	A	M123	0,089	0,766	-0,004	A
M20	0,514	0,474	-0,215	A	M72	0,000	0,999	0,013	A	M124	0,002	0,964	0,082	A
M21	0,496	0,481	0,218	A	M73	0,060	0,807	0,047	A	M125	0,224	0,636	-0,064	A
M22	0,000	0,993	0,023	A	M74	0,022	0,882	0,067	A	M126	2,736	0,098	-0,036	A
M23	0,322	0,570	-0,172	A	M75	0,001	0,979	0,098	A	M127	0,861	0,354	0,074	A
M24	0,000	0,987	-0,022	A	M76	0,486	0,486	-0,046	A	M128	1,542	0,214	0,001	A
M25	0,343	0,558	0,170	A	M77	0,125	0,723	0,174	A	M129	0,001	0,972	0,067	A
M26	0,112	0,738	-0,196	A	M78	0,659	0,417	-0,029	A	M130	0,436	0,509	-0,081	A
M27	0,856	0,355	0,278	A	M79	0,379	0,538	0,233	A	M131	3,052	0,081	0,071	A
M28	0,127	0,721	0,119	A	M80	2,269	0,132	0,040	A	M132	1,707	0,191	-0,073	A

Madde	X ²	P	DeltaMH	Etki	Madde	X ²	P	DeltaMH	Etki	Madde	X ²	P	DeltaMH	Etki
M29	0,243	0,622	0,239	A	M81	0,392	0,531	0,026	A	M133	0,995	0,319	-0,063	A
M30	0,423	0,515	0,258	A	M82	0,910	0,340	-0,042	A	M134	0,007	0,934	-0,083	A
M31	2,572	0,109	0,494	A	M83	0,710	0,399	0,022	A	M135	3,095	0,079	0,086	A
M32	2,768	0,096	0,488	A	M84	0,022	0,882	-0,032	A	M136	0,772	0,380	0,083	A
M33	0,889	0,346	0,280	A	M85	1,383	0,240	0,076	A	M137	0,027	0,870	0,070	A
M34	0,815	0,367	-0,250	A	M86	0,485	0,486	0,176	A	M138	0,131	0,717	-0,045	A
M35	0,095	0,758	0,097	A	M87	0,036	0,850	-0,071	A	M139	0,040	0,841	-0,071	A
M36	0,268	0,605	0,154	A	M88	1,160	0,282	0,222	A	M140	6,321	0,012	0,002	A
M37	0,029	0,865	0,062	A	M89	0,024	0,877	0,071	A	M141	1,261	0,261	0,020	A
M38	0,028	0,868	-0,066	A	M90	0,007	0,932	-0,061	A	M142	0,427	0,514	0,054	A
M39	0,332	0,565	0,171	A	M91	0,521	0,471	0,073	A	M143	3,414	0,065	0,065	A
M40	1,040	0,308	0,308	A	M92	0,003	0,957	0,008	A	M144	0,802	0,371	0,034	A
M41	0,060	0,807	0,088	A	M93	1,535	0,215	0,086	A	M145	2,972	0,085	0,033	A
M42	1,817	0,178	-0,432	A	M94	0,791	0,374	0,011	A	M146	0,001	0,976	-0,022	A
M43	0,003	0,957	-0,031	A	M95	0,378	0,539	0,091	A	M147	0,081	0,777	-0,073	A
M44	8,406	0,004	0,864	A	M96	0,088	0,767	-0,123	A	M148	5,425	0,020	0,003	A
M45	3,221	0,073	0,526	A	M97	1,252	0,263	0,037	A	M149	0,394	0,530	0,047	A
M46	1,305	0,253	0,358	A	M98	2,842	0,092	-0,049	A	M150	0,039	0,844	0,081	A
M47	1,599	0,206	0,391	A	M99	1,290	0,256	0,157	A	M151	0,008	0,931	-0,087	A
M48	1,924	0,165	0,394	A	M100	0,806	0,369	0,071	A					
M49	0,210	0,647	0,135	A	M101	2,871	0,090	0,078	A					
M50	0,059	0,808	0,090	A	M102	0,200	0,655	-0,021	A					
M51	0,264	0,608	0,162	A	M103	0,715	0,398	-0,015	A					
M52	0,152	0,696	0,370	A	M104	0,003	0,957	-0,020	A					

EK-4. Okuduğunu Anlama Madde Havuzuna Ait Faktör Yükleri

Madde No	Faktör yükü	Madde No	Faktör yükü	Madde No	Faktör yükü	Madde No	Faktör yükü	Madde No	Faktör yükü
M1	0,36	M32	0,12	M63	0,30	M94	0,59	M125	0,31
M2	0,38	M33	0,13	M64	0,46	M95	0,16	M126	0,58
M3	0,50	M34	0,10	M65	0,46	M96	0,17	M127	0,55
M4	0,55	M35	0,29	M66	0,48	M97	0,48	M128	0,36
M5	0,43	M36	0,16	M67	0,65	M98	0,16	M129	0,45
M6	0,29	M37	0,33	M68	0,55	M99	0,42	M130	0,00
M7	0,40	M38	0,48	M69	0,52	M100	-0,27	M131	0,58
M8	0,59	M39	0,35	M70	0,59	M101	0,16	M132	0,59
M9	0,31	M40	0,31	M71	0,33	M102	0,01	M133	-0,15
M10	0,51	M41	0,48	M72	0,52	M103	0,31	M134	0,52
M11	0,63	M42	0,46	M73	0,59	M104	0,49	M135	0,44
M12	0,51	M43	0,38	M74	0,52	M105	0,53	M136	0,31
M13	0,57	M44	0,46	M75	0,30	M106	0,46	M137	0,32
M14	0,45	M45	0,40	M76	0,06	M107	0,39	M138	0,11
M15	0,52	M46	0,42	M77	0,53	M108	0,57	M139	0,31
M16	0,33	M47	0,52	M78	0,56	M109	0,42	M140	0,63
M17	0,49	M48	0,29	M79	0,15	M110	0,52	M141	0,44
M18	0,37	M49	0,30	M80	0,56	M111	-0,04	M142	0,37
M19	0,45	M50	0,52	M81	0,42	M112	0,32	M143	0,52
M20	0,36	M51	0,35	M82	0,61	M113	0,31	M144	-0,06
M21	0,44	M52	0,63	M83	0,50	M114	0,59	M145	0,58
M22	0,05	M53	0,49	M84	0,28	M115	0,55	M146	0,30
M23	0,31	M54	0,33	M85	0,46	M116	0,47	M147	0,32
M24	0,47	M55	0,58	M86	0,30	M117	0,13	M148	0,53
M25	0,30	M56	0,49	M87	0,45	M118	0,52	M149	0,47
M26	0,42	M57	0,14	M88	0,61	M119	0,54	M150	0,18
M27	0,42	M58	0,33	M89	0,62	M120	0,34	M151	0,41
M28	0,37	M59	0,52	M90	0,44	M121	0,43	M152	0,42
M29	0,57	M60	-0,06	M91	0,39	M122	0,40	M153	0,55
M30	0,52	M61	0,52	M92	0,49	M123	-0,13	M154	0,56
M31	0,47	M62	0,45	M93	0,51	M124	0,47	M155	0,49

EK-5. Madde Parametrelerinin Değişmezliği İçin Yapılan DIF Sonuçları

Madde	χ^2	P	DeltaMH	Etki	Madde	χ^2	P	DeltaMH	Etki	Madde	χ^2	P	DeltaMH	Etki
M1	0,000	0,995	-0,027	A	M45	0,060	0,807	0,084	A	M89	0,723	0,395	-0,271	A
M2	3,695	0,055	0,583	A	M46	3,862	0,049	0,550	A	M90	0,160	0,689	-0,123	A
M3	0,971	0,324	0,330	A	M47	0,006	0,939	0,001	A	M91	3,646	0,056	-0,528	A
M4	1,773	0,183	-0,573	A	M48	0,007	0,932	0,049	A	M92	0,196	0,658	-0,148	A
M5	0,878	0,349	0,272	A	M49	0,309	0,579	-0,165	A	M93	0,716	0,397	-0,241	A
M6	0,187	0,665	-0,129	A	M50	0,044	0,834	-0,084	A	M94	2,719	0,099	-0,478	A
M7	0,848	0,357	0,258	A	M51	1,889	0,169	0,414	A	M95	0,147	0,701	0,118	A
M8	2,293	0,130	-0,653	A	M52	0,973	0,324	0,276	A	M96	0,758	0,384	0,260	A
M9	2,371	0,124	-0,408	A	M53	0,096	0,756	-0,108	A	M97	0,000	0,987	0,023	A
M10	3,823	0,051	0,559	A	M54	0,077	0,782	0,093	A	M98	0,201	0,654	0,143	A
M11	0,785	0,376	0,304	A	M55	0,014	0,905	-0,049	A	M99	0,114	0,736	0,107	A
M12	0,264	0,608	0,164	A	M56	0,134	0,715	0,109	A	M100	0,358	0,550	-0,180	A
M13	0,248	0,619	0,176	A	M57	0,334	0,563	0,173	A	M101	0,813	0,367	-0,283	A
M14	0,594	0,441	-0,225	A	M58	0,125	0,723	-0,122	A	M102	0,345	0,557	-0,174	A
M15	0,598	0,439	0,231	A	M59	0,060	0,806	0,083	A	M103	1,405	0,236	-0,328	A
M16	1,224	0,269	0,303	A	M60	0,004	0,950	0,003	A	M104	0,686	0,408	-0,237	A
M17	0,251	0,616	-0,157	A	M61	0,039	0,844	-0,075	A	M105	0,001	0,982	-0,022	A
M18	0,107	0,744	0,102	A	M62	0,091	0,763	0,107	A	M106	0,273	0,602	-0,154	A
M19	0,000	0,995	-0,013	A	M63	1,849	0,174	-0,421	A	M107	0,133	0,715	-0,124	A
M20	0,030	0,862	0,060	A	M64	1,282	0,258	-0,321	A	M108	0,026	0,873	-0,057	A
M21	0,061	0,804	-0,083	A	M65	0,003	0,957	-0,031	A	M109	0,019	0,891	-0,052	A
M22	1,101	0,294	0,300	A	M66	1,725	0,189	-0,420	A	M110	0,001	0,983	-0,025	A
M23	4,269	0,039	0,586	A	M67	0,261	0,610	0,155	A	M111	6,582	0,010	0,764	A
M24	0,268	0,605	-0,146	A	M68	2,607	0,106	0,436	A	M112	2,747	0,097	-0,484	A
M25	2,008	0,157	0,595	A	M69	1,581	0,209	-0,371	A	M113	0,093	0,760	0,097	A
M26	2,300	0,129	0,426	A	M70	0,571	0,450	-0,229	A	M114	0,003	0,957	-0,001	A
M27	0,009	0,924	-0,041	A	M71	2,155	0,142	-0,439	A	M115	0,015	0,901	0,048	A
M28	0,541	0,462	-0,302	A	M72	0,007	0,932	-0,038	A	M116	0,838	0,360	-0,264	A

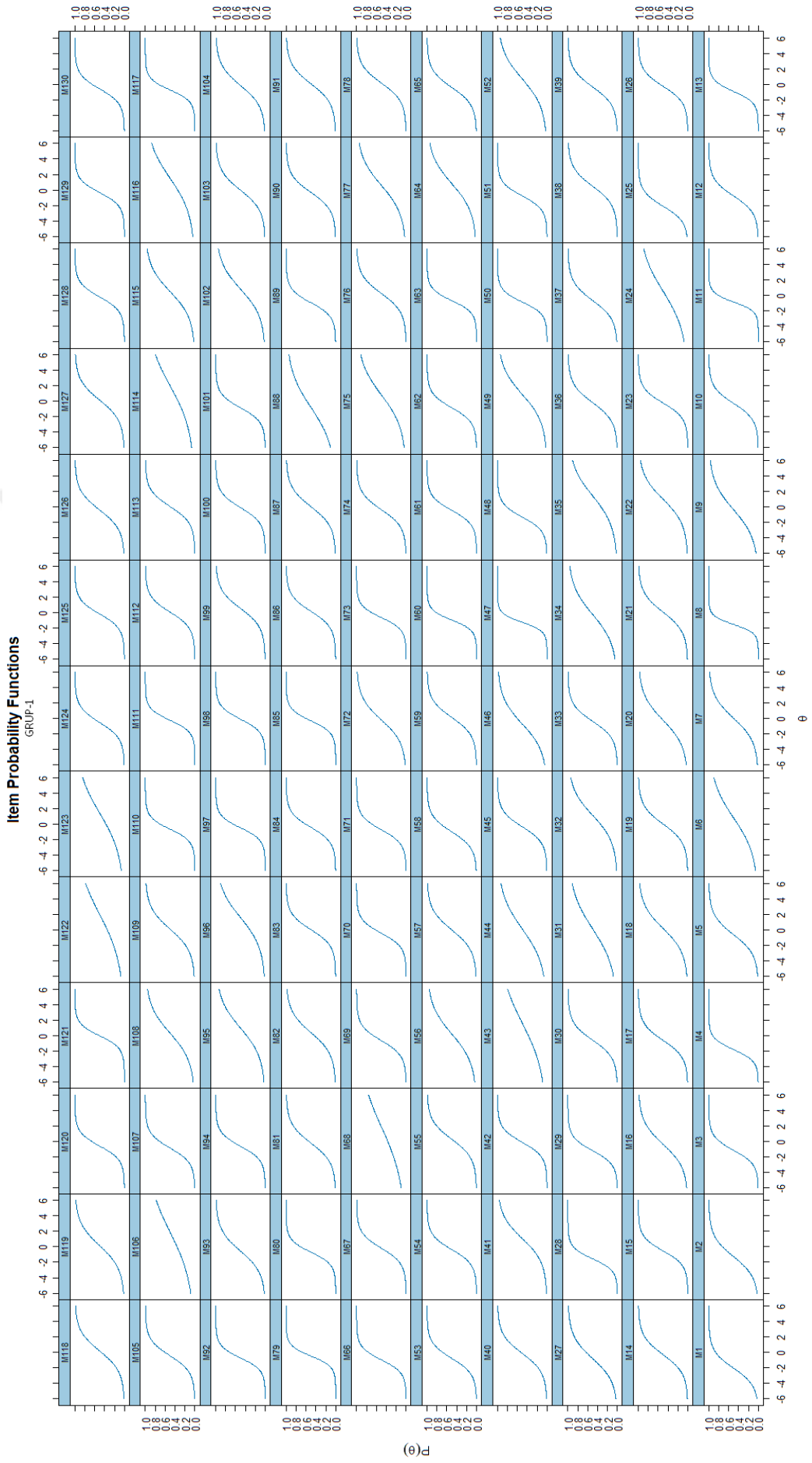
Madde	X ²	P	DeltaMH	Etki	Madde	X ²	P	DeltaMH	Etki	Madde	X ²	P	DeltaMH	Etki
M29	0,012	0,912	-0,056	A	M73	0,076	0,783	-0,102	A	M117	0,091	0,763	-0,113	A
M30	0,128	0,721	0,115	A	M74	0,299	0,585	-0,165	A	M118	0,010	0,920	0,042	A
M31	0,015	0,904	0,045	A	M75	0,377	0,539	0,174	A	M119	0,026	0,872	0,057	A
M32	0,035	0,852	0,066	A	M76	0,880	0,348	0,271	A	M120	0,525	0,469	-0,223	A
M33	2,939	0,087	0,495	A	M77	0,068	0,795	-0,084	A	M121	2,479	0,115	0,470	A
M34	0,053	0,818	-0,075	A	M78	0,739	0,390	-0,247	A	M122	0,110	0,741	-0,106	A
M35	0,522	0,470	0,213	A	M79	0,580	0,446	-0,240	A	M123	0,088	0,767	-0,094	A
M36	1,175	0,279	0,316	A	M80	2,272	0,132	-0,476	A	M124	0,072	0,788	-0,092	A
M37	0,077	0,781	0,092	A	M81	0,017	0,895	-0,052	A	M125	0,767	0,381	0,258	A
M38	1,375	0,241	0,334	A	M82	0,010	0,920	0,043	A	M126	1,203	0,273	-0,310	A
M39	0,012	0,912	0,045	A	M83	1,867	0,172	-0,391	A	M127	1,141	0,285	-0,300	A
M40	0,290	0,591	-0,158	A	M84	1,103	0,294	-0,312	A	M128	0,752	0,386	0,259	A
M41	1,121	0,290	0,320	A	M85	0,030	0,862	-0,067	A	M129	0,188	0,665	-0,139	A
M42	0,001	0,982	-0,010	A	M86	1,724	0,189	-0,378	A	M130	0,015	0,902	-0,050	A
M43	2,051	0,152	0,392	A	M87	0,115	0,735	-0,106	A					
M44	0,670	0,413	0,230	A	M88	0,000	0,998	0,014	A					

EK-6 Madde Parametrelerinin Değişmezliği için Rastgele Bölünen İki Grup için Elde Edilen Madde Parametreleri

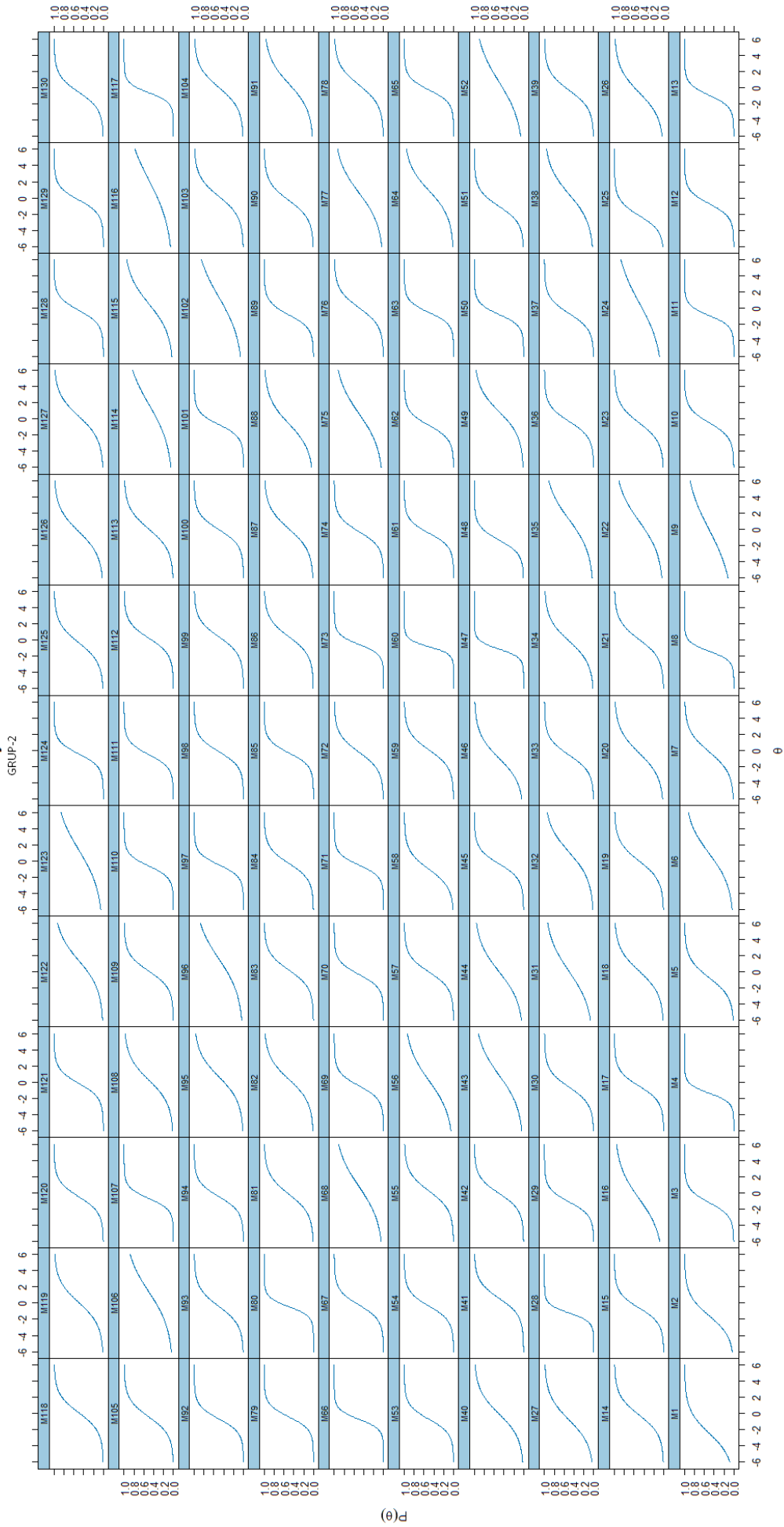
Madde	Grup-1			Grup-2			Madde	Grup-1			Grup-2			Madde	Grup-1			Grup-2		
	a	b		a	b			a	b		a	b			a	b		a	b	
M1	0,885	-2,026	0,677	0,677	-2,427		M45	0,966	-0,347	1,056	1,056	-0,346		M89	1,126	-0,915	1,191	1,191	-0,795	
M2	0,751	-1,268	0,764	0,764	-1,565		M46	0,564	-0,723	0,641	0,641	-0,949		M90	0,875	-0,325	0,860	0,860	-0,313	
M3	1,112	-1,174	1,197	1,197	-1,234		M47	1,738	-1,241	1,947	1,947	-1,120		M91	0,789	0,054	0,588	0,588	0,442	
M4	1,359	-1,818	1,762	1,762	-1,377		M48	1,123	-1,301	1,159	1,159	-1,311		M92	1,277	-0,702	1,191	1,191	-0,643	
M5	0,782	-0,632	0,828	0,828	-0,758		M49	0,536	0,903	0,636	0,636	0,907		M93	0,708	-0,552	0,862	0,862	-0,399	
M6	0,412	0,816	0,483	0,483	0,929		M50	1,286	-0,818	1,403	1,403	-0,742		M94	1,119	-0,653	1,088	1,088	-0,494	
M7	0,658	0,017	0,763	0,763	-0,162		M51	1,141	-0,784	0,973	0,973	-1,052		M95	0,518	0,955	0,634	0,634	0,667	
M8	1,821	-1,518	1,747	1,747	-1,383		M52	0,524	0,685	0,404	0,404	0,675		M96	0,500	1,580	0,434	0,434	1,579	
M9	0,528	-0,348	0,336	0,336	-0,030		M53	0,999	-0,761	1,137	1,137	-0,690		M97	1,270	-0,479	1,363	1,363	-0,457	
M10	0,942	-0,264	1,087	1,087	-0,470		M54	1,037	-0,353	1,099	1,099	-0,364		M98	1,197	-0,106	1,076	1,076	-0,141	
M11	1,670	-0,808	1,532	1,532	-0,874		M55	0,818	0,239	0,860	0,860	0,205		M99	0,811	0,550	0,921	0,921	0,399	
M12	0,923	-0,780	1,217	1,217	-0,719		M56	0,520	0,193	0,463	0,463	0,177		M100	1,011	-0,076	1,025	1,025	-0,010	
M13	1,361	-0,925	1,382	1,382	-0,973		M57	0,747	-0,021	0,983	0,983	-0,052		M101	1,098	-0,842	1,283	1,283	-0,694	
M14	0,822	-0,277	0,819	0,819	-0,172		M58	0,952	-1,012	0,874	0,874	-1,021		M102	0,551	0,859	0,374	0,374	1,399	
M15	1,081	-0,520	1,047	1,047	-0,600		M59	0,875	0,117	0,919	0,919	0,074		M103	0,764	-0,030	0,787	0,787	0,187	
M16	0,657	-0,403	0,461	0,461	-0,749		M60	1,515	-0,952	2,187	2,187	-0,819		M104	0,734	-0,159	0,756	0,756	-0,004	
M17	0,901	-0,639	1,036	1,036	-0,501		M61	1,104	-0,446	1,197	1,197	-0,460		M105	0,947	-0,426	0,896	0,896	-0,397	
M18	0,620	0,170	0,696	0,696	0,060		M62	1,066	-0,837	1,151	1,151	-0,804		M106	0,303	1,904	0,424	0,424	1,411	
M19	0,835	0,005	0,818	0,818	0,006		M63	1,290	-0,646	1,396	1,396	-0,495		M107	1,054	-0,891	1,456	1,456	-0,718	
M20	0,587	-0,242	0,655	0,655	-0,309		M64	0,531	0,810	0,595	0,595	0,900		M108	0,535	0,294	0,656	0,656	0,254	
M21	0,776	0,071	0,851	0,851	0,069		M65	0,882	-0,416	1,290	1,290	-0,309		M109	0,697	0,082	0,958	0,958	0,055	
M22	0,612	1,004	0,463	0,463	1,017		M66	1,214	-0,939	1,665	1,665	-0,684		M110	1,270	-0,657	1,368	1,368	-0,612	
M23	1,017	-0,222	0,829	0,829	-0,484		M67	1,004	-0,178	1,071	1,071	-0,227		M111	1,331	-0,252	1,230	1,230	-0,500	
M24	0,379	0,557	0,354	0,354	0,719		M68	0,278	1,859	0,440	0,440	0,764		M112	0,977	0,069	0,994	0,994	0,233	
M25	1,007	-2,104	1,043	1,043	-2,188		M69	1,028	-0,667	1,182	1,182	-0,457		M113	0,937	0,060	0,757	0,757	0,059	
M26	0,958	-0,268	0,628	0,628	-0,637		M70	1,130	-0,453	1,222	1,222	-0,343		M114	0,336	2,014	0,375	0,375	1,830	
M27	0,665	-0,788	0,638	0,638	-0,780		M71	1,097	-0,731	1,401	1,401	-0,489		M115	0,579	0,524	0,523	0,523	0,583	
M28	1,331	-1,556	1,871	1,871	-1,210		M72	0,671	-0,203	0,885	0,885	-0,196		M116	0,418	1,505	0,333	0,333	2,232	
M29	1,122	-1,241	1,202	1,202	-1,153		M73	1,286	-0,641	1,612	1,612	-0,538		M117	1,360	-0,737	1,664	1,664	-0,649	

Madde	Grup-1			Grup-2			Madde	Grup-1			Grup-2			Madde	Grup-1			Grup-2		
	a	b		a	b			a	b		a	b			a	b		a	b	
M30	0,877	-0,606	0,941	-0,626		M74	0,899	-0,348	1,066	-0,258	M118	0,747	0,196	0,862	0,118					
M31	0,397	0,113	0,442	0,022		M75	0,461	0,834	0,450	0,672	M119	0,654	0,143	0,704	0,084					
M32	0,567	1,129	0,568	1,102		M76	0,833	0,280	0,881	0,093	M120	1,119	-0,622	1,024	-0,504					
M33	0,910	-0,380	0,950	-0,561		M77	0,523	0,587	0,484	0,639	M121	1,261	-0,040	1,126	-0,147					
M34	0,512	-0,235	0,702	-0,126		M78	0,849	0,150	0,793	0,309	M122	0,318	1,988	0,548	1,369					
M35	0,492	1,436	0,478	1,226		M79	1,493	-0,453	1,299	-0,395	M123	0,363	1,424	0,391	1,386					
M36	0,817	-0,173	1,000	-0,303		M80	1,265	-0,688	1,704	-0,439	M124	0,971	-0,534	1,268	-0,415					
M37	0,814	-0,630	1,001	-0,623		M81	0,796	-0,252	0,817	-0,195	M125	1,059	-0,169	0,799	-0,334					
M38	0,816	0,477	0,548	0,462		M82	0,710	0,376	0,694	0,343	M126	0,785	-0,386	0,659	-0,254					
M39	0,847	-0,353	0,874	-0,353		M83	0,971	0,017	0,926	0,182	M127	0,751	0,155	0,693	0,326					
M40	0,808	-0,114	0,654	-0,032		M84	1,040	-0,429	0,993	-0,291	M128	1,086	-0,168	1,168	-0,233					
M41	0,641	1,073	0,936	0,607		M85	1,300	-0,314	1,257	-0,289	M129	1,166	-0,340	1,166	-0,256					
M42	1,085	-0,143	0,976	-0,097		M86	0,949	0,173	0,861	0,403	M130	1,056	-0,604	0,938	-0,593					
M43	0,309	1,606	0,467	0,870		M87	0,782	0,047	0,722	0,073										
M44	0,434	0,116	0,510	-0,018		M88	0,409	-0,789	0,603	-0,599										

EK-7 Madde Parametrelerinin Değişmezliği için Rastgele Bölünen İki Grup için Elde Edilen Madde Karakteristik Eğrileri



Item Probability Functions



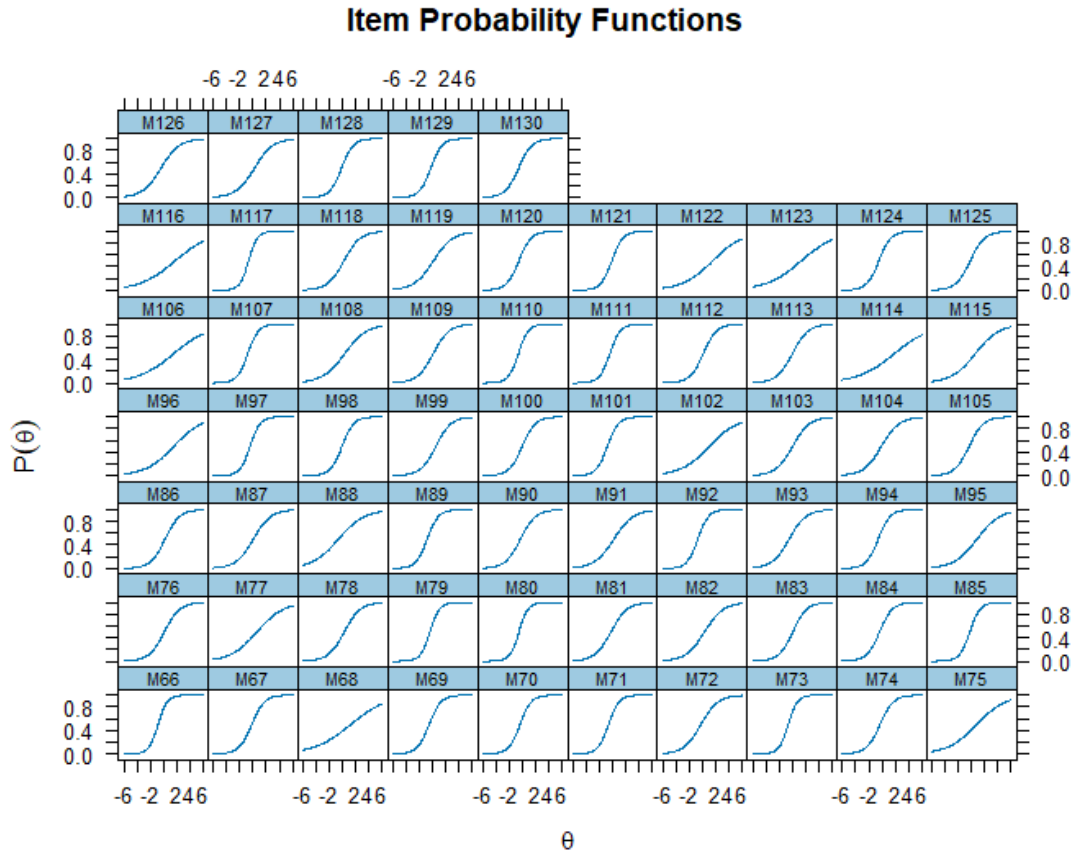
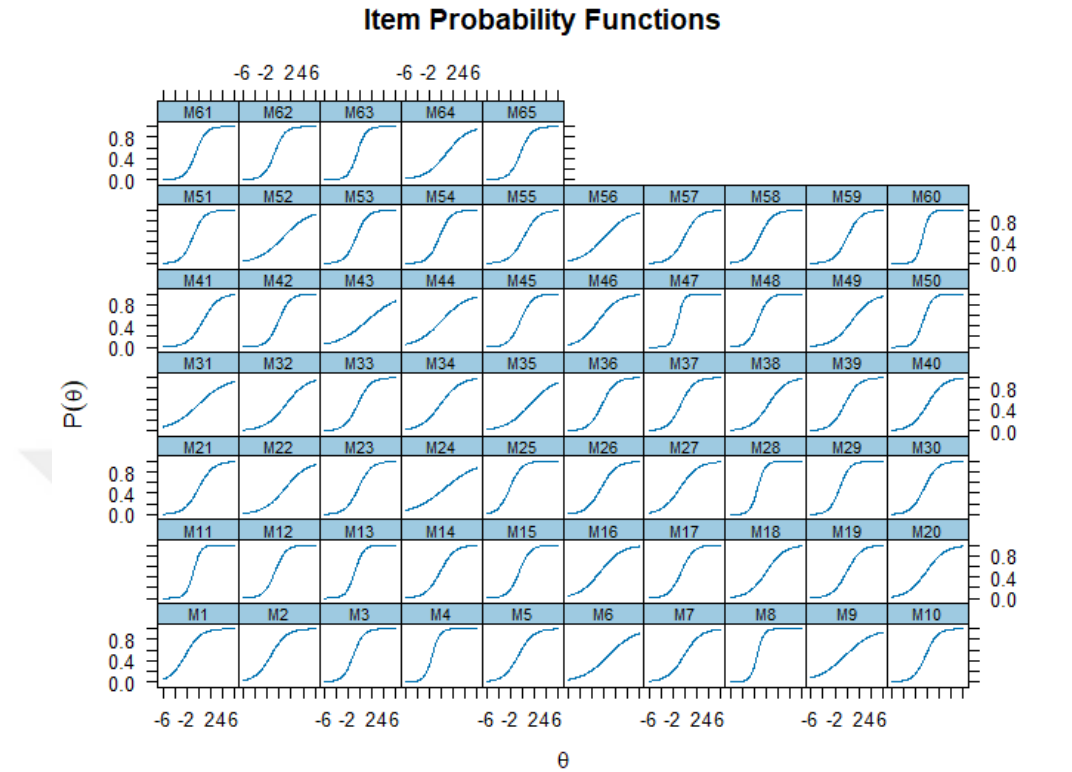
EK-8. Okuduğunu Anlama Testi Madde Havuzuna Ait Madde Parametreleri ve Uyum Değerleri

Madde	α (sh)	b (sh)	Bilgi	RMSEA	Madde	α (sh)	b (sh)	Bilgi	RMSEA
M1	0,775 (0,085)	-2,208 (0,245)	0,494	0,011	M66	1,426 (0,093)	-0,795 (0,061)	1,363	0,006
M2	0,756 (0,071)	-1,417 (0,148)	0,558	0,007	M67	1,034 (0,070)	-0,203 (0,062)	0,946	0,007
M3	1,149 (0,087)	-1,208 (0,093)	1,014	0,000	M68	0,359 (0,053)	1,190 (0,248)	0,172	0,006
M4	1,558 (0,127)	-1,568 (0,102)	1,412	0,013	M69	1,104 (0,075)	-0,555 (0,065)	1,016	0,000
M5	0,802 (0,066)	-0,698 (0,091)	0,656	0,002	M70	1,178 (0,076)	-0,395 (0,059)	1,107	0,005
M6	0,449 (0,054)	0,873 (0,174)	0,259	0,010	M71	1,240 (0,080)	-0,598 (0,061)	1,169	0,000
M7	0,708 (0,060)	-0,077 (0,085)	0,560	0,009	M72	0,773 (0,062)	-0,197 (0,079)	0,636	0,010
M8	1,779 (0,139)	-1,449 (0,087)	1,678	0,016	M73	1,439 (0,088)	-0,585 (0,054)	1,390	0,013
M9	0,427 (0,055)	-0,222 (0,134)	0,243	0,000	M74	0,983 (0,069)	-0,298 (0,066)	0,884	0,000
M10	1,002 (0,071)	-0,373 (0,066)	0,905	0,000	M75	0,455 (0,054)	0,755 (0,157)	0,267	0,000
M11	1,597 (0,103)	-0,841 (0,057)	1,547	0,007	M76	0,855 (0,064)	0,185 (0,073)	0,735	0,016
M12	1,058 (0,077)	-0,747 (0,073)	0,952	0,000	M77	0,502 (0,055)	0,613 (0,131)	0,318	0,012
M13	1,373 (0,092)	-0,949 (0,069)	1,293	0,013	M78	0,820 (0,063)	0,227 (0,077)	0,693	0,006
M14	0,821 (0,064)	-0,225 (0,075)	0,694	0,007	M79	1,393 (0,085)	-0,425 (0,052)	1,346	0,009
M15	1,059 (0,073)	-0,561 (0,068)	0,964	0,000	M80	1,467 (0,090)	-0,549 (0,053)	1,422	0,017
M16	0,553 (0,058)	-0,551 (0,118)	0,374	0,012	M81	0,809 (0,063)	-0,223 (0,076)	0,679	0,008
M17	0,966 (0,071)	-0,566 (0,072)	0,855	0,008	M82	0,701 (0,059)	0,360 (0,090)	0,547	0,017
M18	0,655 (0,059)	0,113 (0,090)	0,496	0,000	M83	0,948 (0,066)	0,098 (0,067)	0,846	0,006
M19	0,826 (0,063)	0,006 (0,074)	0,701	0,000	M84	1,015 (0,070)	-0,361 (0,066)	0,920	0,000
M20	0,620 (0,059)	-0,277 (0,095)	0,454	0,000	M85	1,280 (0,080)	-0,302 (0,054)	1,224	0,000
M21	0,813 (0,062)	0,071 (0,076)	0,685	0,015	M86	0,902 (0,064)	0,284 (0,072)	0,789	0,009
M22	0,534 (0,056)	1,012 (0,154)	0,343	0,000	M87	0,749 (0,061)	0,060 (0,080)	0,608	0,013
M23	0,916 (0,067)	-0,344 (0,071)	0,805	0,008	M88	0,507 (0,057)	-0,672 (0,136)	0,322	0,000

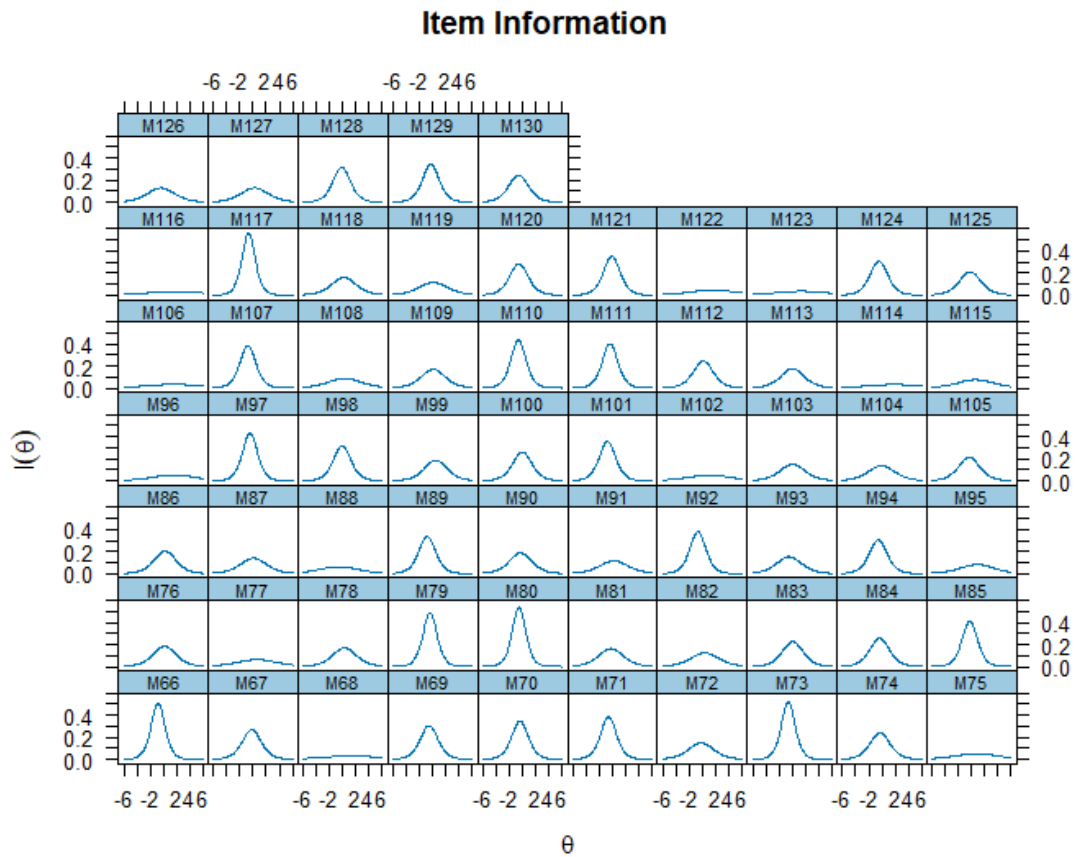
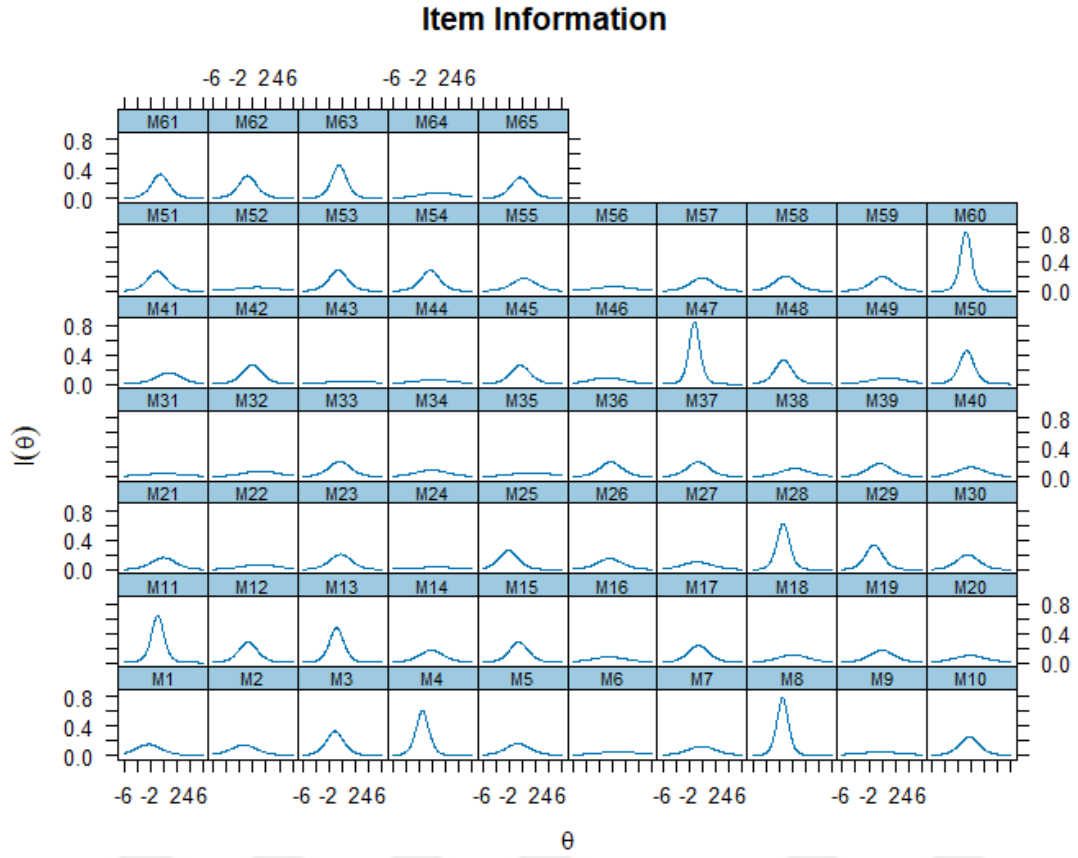
Madde	α (sh)	b (sh)	Bilgi	RMSEA	Madde	α (sh)	b (sh)	Bilgi	RMSEA
M24	0,367 (0,053)	0,635 (0,180)	0,183	0,000	M89	1,159 (0,082)	-0,853 (0,073)	1,060	0,011
M25	1,023 (0,106)	-2,150 (0,202)	0,722	0,017	M90	0,869 (0,065)	-0,319 (0,073)	0,749	0,006
M26	0,780 (0,064)	-0,422 (0,082)	0,641	0,012	M91	0,680 (0,059)	0,226 (0,090)	0,526	0,009
M27	0,650 (0,062)	-0,785 (0,113)	0,478	0,005	M92	1,233 (0,082)	-0,673 (0,062)	1,157	0,000
M28	1,584 (0,119)	-1,360 (0,082)	1,478	0,012	M93	0,784 (0,064)	-0,468 (0,083)	0,644	0,000
M29	1,161 (0,089)	-1,197 (0,090)	1,029	0,012	M94	1,104 (0,076)	-0,574 (0,065)	1,015	0,000
M30	0,907 (0,070)	-0,617 (0,078)	0,784	0,008	M95	0,576 (0,056)	0,797 (0,133)	0,394	0,010
M31	0,419 (0,054)	0,066 (0,133)	0,236	0,000	M96	0,464 (0,056)	1,589 (0,233)	0,258	0,017
M32	0,567 (0,057)	1,116 (0,151)	0,375	0,000	M97	1,311 (0,082)	-0,468 (0,056)	1,254	0,007
M33	0,924 (0,069)	-0,473 (0,072)	0,810	0,000	M98	1,133 (0,073)	-0,124 (0,058)	1,061	0,005
M34	0,605 (0,058)	-0,172 (0,096)	0,438	0,000	M99	0,864 (0,063)	0,472 (0,080)	0,738	0,000
M35	0,484 (0,056)	1,334 (0,194)	0,284	0,000	M100	1,020 (0,069)	-0,042 (0,062)	0,931	0,010
M36	0,903 (0,066)	-0,242 (0,070)	0,791	0,005	M101	1,190 (0,082)	-0,762 (0,068)	1,103	0,000
M37	0,901 (0,069)	-0,627 (0,079)	0,776	0,000	M102	0,457 (0,055)	1,090 (0,185)	0,263	0,002
M38	0,673 (0,059)	0,472 (0,097)	0,512	0,009	M103	0,776 (0,062)	0,080 (0,078)	0,641	0,000
M39	0,859 (0,066)	-0,353 (0,074)	0,736	0,000	M104	0,743 (0,061)	-0,081 (0,080)	0,601	0,008
M40	0,729 (0,061)	-0,077 (0,081)	0,585	0,000	M105	0,922 (0,067)	-0,412 (0,072)	0,809	0,000
M41	0,784 (0,061)	0,805 (0,099)	0,630	0,022	M106	0,363 (0,054)	1,619 (0,307)	0,170	0,012
M42	1,025 (0,069)	-0,122 (0,063)	0,936	0,007	M107	1,238 (0,084)	-0,794 (0,068)	1,155	0,000
M43	0,389 (0,054)	1,162 (0,218)	0,198	0,008	M108	0,594 (0,057)	0,273 (0,102)	0,425	0,003
M44	0,471 (0,055)	0,045 (0,118)	0,289	0,014	M109	0,823 (0,063)	0,069 (0,075)	0,697	0,006
M45	1,011 (0,071)	-0,346 (0,064)	0,915	0,009	M110	1,318 (0,086)	-0,634 (0,059)	1,254	0,004
M46	0,597 (0,060)	-0,846 (0,129)	0,416	0,012	M111	1,266 (0,080)	-0,375 (0,055)	1,206	0,000
M47	1,842 (0,126)	-1,178 (0,067)	1,782	0,007	M112	0,983 (0,068)	0,152 (0,065)	0,887	0,008
M48	1,141 (0,089)	-1,306 (0,101)	0,993	0,010	M113	0,840 (0,063)	0,060 (0,074)	0,718	0,007

Madde	α (sh)	b (sh)	Bilgi	RMSEA	Madde	α (sh)	b (sh)	Bilgi	RMSEA
M49	0,586 (0,057)	0,904 (0,135)	0,402	0,000	M114	0,355 (0,054)	1,916 (0,359)	0,160	0,000
M50	1,346 (0,088)	-0,778 (0,063)	1,276	0,011	M115	0,549 (0,056)	0,554 (0,12)	0,369	0,007
M51	1,052 (0,078)	-0,911 (0,081)	0,934	0,000	M116	0,374 (0,055)	1,836 (0,321)	0,176	0,000
M52	0,461 (0,055)	0,682 (0,150)	0,274	0,000	M117	1,499 (0,094)	-0,691 (0,056)	1,450	0,007
M53	1,066 (0,076)	-0,724 (0,073)	0,963	0,009	M118	0,803 (0,062)	0,156 (0,077)	0,673	0,000
M54	1,066 (0,072)	-0,359 (0,063)	0,980	0,000	M119	0,677 (0,059)	0,113 (0,089)	0,523	0,008
M55	0,838 (0,063)	0,222 (0,075)	0,713	0,014	M120	1,068 (0,075)	-0,565 (0,067)	0,974	0,012
M56	0,489 (0,055)	0,186 (0,120)	0,308	0,010	M121	1,186 (0,075)	-0,092 (0,056)	1,122	0,016
M57	0,859 (0,064)	-0,036 (0,072)	0,741	0,008	M122	0,431 (0,055)	1,597 (0,256)	0,229	0,010
M58	0,913 (0,073)	-1,016 (0,098)	0,766	0,015	M123	0,377 (0,054)	1,406 (0,269)	0,185	0,011
M59	0,894 (0,064)	0,095 (0,070)	0,782	0,000	M124	1,108 (0,075)	-0,468 (0,063)	1,024	0,000
M60	1,803 (0,111)	-0,880 (0,056)	1,765	0,011	M125	0,919 (0,067)	-0,244 (0,069)	0,809	0,000
M61	1,148 (0,075)	-0,453 (0,062)	1,071	0,013	M126	0,720 (0,061)	-0,325 (0,086)	0,571	0,013
M62	1,108 (0,079)	-0,820 (0,075)	1,004	0,005	M127	0,721 (0,060)	0,238 (0,085)	0,574	0,000
M63	1,343 (0,084)	-0,567 (0,058)	1,285	0,000	M128	1,124 (0,072)	-0,201 (0,059)	1,050	0,007
M64	0,563 (0,057)	0,857 (0,133)	0,378	0,000	M129	1,167 (0,076)	-0,298 (0,057)	1,097	0,008
M65	1,071 (0,072)	-0,352 (0,063)	0,985	0,000	M130	0,993 (0,072)	-0,600 (0,072)	0,885	0,003

EK-9. Okuduğunu Anlama Testi Madde Havuzuna Ait Madde Karakteristik Eğrileri



EK-10. Okuduğunu Anlama Testi Madde Havuzuna Ait Maddi Bilgi Fonksiyonları



EK-11. Tüm Simülasyon Koşullarına İlişkin Sonuçlar

Başlama kuralı	Madde seçim yöntemi	Yetenek kestirim yöntemi	Sonlandırma kuralı	Simülasyon süresi	Ortalama test uzunluğu	Gerçek			Bias	Madde örtüşme oranı	Ortalama standart hata	Ortalama madde bilgisi
						yetenek ile kestirilen yetenek korelasyon	RMSE	yetenek				
theta=0	MFI	MAP	0,30	1,39	41,21	0,975	0,215	0,000	0,592	0,300	0,333	
theta=0	MFI	ML	0,30	1,69	48,30	0,980	0,211	-0,017	0,599	0,304	0,309	
theta=0	MFI	EAP	0,30	12,30	42,71	0,977	0,211	0,007	0,599	0,300	0,328	
theta=0	MFI	MAP	0,50	0,24	10,05	0,894	0,433	-0,012	0,437	0,483	0,415	
theta=0	MFI	ML	0,50	0,41	17,74	0,927	0,422	-0,025	0,360	0,485	0,355	
theta=0	MFI	EAP	0,50	2,17	10,86	0,902	0,422	0,012	0,449	0,484	0,409	
theta=0	MFI	ML	25 madde	0,32	25,00	0,946	0,398	0,048	0,572	0,398	0,323	
theta=0	MFI	MAP	25 madde	0,38	25,00	0,955	0,287	-0,018	0,662	0,340	0,333	
theta=0	MFI	EAP	25 madde	5,03	25,00	0,954	0,293	-0,002	0,652	0,347	0,337	
theta=0	MFI	ML	0,3-25 madde	0,27	24,93	0,945	0,398	0,047	0,572	0,399	0,324	
theta=0	MFI	MAP	0,3-25 madde	0,37	23,90	0,953	0,292	-0,017	0,651	0,345	0,345	
theta=0	MFI	EAP	0,3-25 madde	4,89	24,30	0,953	0,296	-0,001	0,645	0,349	0,341	
theta=0	MFI	MAP	0,5-25 madde	0,16	9,63	0,892	0,437	-0,013	0,454	0,485	0,415	
theta=0	MFI	ML	0,5-25 madde	0,17	13,85	0,915	0,490	0,014	0,408	0,517	0,356	
theta=0	MFI	EAP	0,5-25 madde	1,92	10,46	0,900	0,426	0,010	0,464	0,486	0,410	
theta=-0.5:0.5	MFI	ML	0,30	1,97	48,36	0,981	0,210	-0,016	0,056	0,304	0,309	
theta=-0.5:0.5	MFI	MAP	0,30	2,50	41,18	0,975	0,215	-0,001	0,047	0,047	0,336	
theta=-0.5:0.5	MFI	EAP	0,30	12,48	42,70	0,977	0,211	0,006	0,049	0,299	0,329	
theta=-0.5:0.5	MFI	ML	0,50	0,41	17,88	0,929	0,416	-0,026	0,012	0,483	0,360	
theta=-0.5:0.5	MFI	MAP	0,50	0,45	10,06	0,893	0,434	-0,020	0,008	0,482	0,435	
theta=-0.5:0.5	MFI	EAP	0,50	3,35	10,88	0,901	0,425	0,005	0,009	0,484	0,420	
theta=-0.5:0.5	MFI	MAP	25 madde	1,30	25,00	0,955	0,287	-0,020	0,320	0,400	0,341	
theta=-0.5:0.5	MFI	ML	25 madde	1,73	25,00	0,945	0,398	0,043	0,276	0,399	0,317	
theta=-0.5:0.5	MFI	EAP	25 madde	5,57	25,00	0,954	0,293	-0,005	0,031	0,347	0,337	
theta=-0.5:0.5	MFI	ML	0,3-25 madde	1,01	24,94	0,945	0,398	0,043	0,027	0,399	0,318	
theta=-0.5:0.5	MFI	MAP	0,30-25 madde	1,60	23,89	0,953	0,292	-0,020	0,300	0,345	0,348	
theta=-0.5:0.5	MFI	EAP	0,30-25 madde	5,44	24,30	0,953	0,296	-0,004	0,030	0,349	0,341	
theta=-0.5:0.5	MFI	ML	0,50-25 madde	0,42	13,97	0,916	0,486	0,011	0,011	0,515	0,359	
theta=-0.5:0.5	MFI	MAP	0,50-25 madde	0,45	9,63	0,891	0,438	-0,021	0,008	0,484	0,435	

Başlama kuralı	Madde seçim yöntemi	Yetenek kestirim yöntemi	Sonlandırma kuralı	Simülasyon süresi	Ortalama test uzunluğu	Gerçek				Madde örtüşme oranı	Ortalama standart hata	Ortalama madde bilgisi
						yetenek ile kestirilen yetenek korelasyon	RMSE	Bias				
theta=-0.5:0.5	MFI	EAP	0,50-25 madde	2,00	10,45	0,899	0,429	0,004	0,009	0,486	0,421	
theta=0	KL	MAP	0,30	47,70	41,33	0,975	0,215	-0,001	0,572	0,300	0,331	
theta=0	KL	ML	0,30	52,72	47,94	0,981	0,210	-0,018	0,602	0,304	0,313	
theta=0	KL	EAP	0,30	59,00	42,81	0,976	0,213	0,004	0,585	0,300	0,327	
theta=0	KL	MAP	0,50	13,31	10,34	0,898	0,425	-0,013	0,401	0,484	0,403	
theta=0	KL	EAP	0,50	16,31	11,12	0,906	0,414	0,013	0,414	0,484	0,400	
theta=0	KL	ML	0,50	21,72	17,37	0,930	0,414	-0,026	0,366	0,483	0,376	
theta=0	KL	ML	25 madde	29,35	25,00	0,948	0,386	0,047	0,582	0,395	0,329	
theta=0	KL	MAP	25 madde	32,79	25,00	0,955	0,286	-0,017	0,608	0,341	0,337	
theta=0	KL	EAP	25 madde	37,66	25,00	0,955	0,291	-0,004	0,605	0,347	0,336	
theta=0	KL	MAP	0,30-25 madde	31,56	23,93	0,953	0,292	-0,017	0,596	0,346	0,343	
theta=0	KL	ML	0,30-25 madde	32,71	24,89	0,948	0,386	0,047	0,581	0,395	0,330	
theta=0	KL	EAP	0,30-25 madde	36,28	24,34	0,953	0,295	-0,004	0,600	0,350	0,339	
theta=0	KL	MAP	0,50-25 madde	12,75	9,88	0,896	0,430	-0,015	0,417	0,486	0,403	
theta=0	KL	EAP	0,50-25 madde	15,52	10,68	0,904	0,419	0,009	0,428	0,486	0,400	
theta=0	KL	ML	0,50-25 madde	17,60	13,50	0,919	0,478	0,010	0,421	0,513	0,376	
theta=-0.5:0.5	KL	ML	0,30	56,31	47,94	0,980	0,213	-0,017	0,056	0,304	0,314	
theta=-0.5:0.5	KL	EAP	0,30	112,15	42,82	0,976	0,212	0,005	0,048	0,300	0,328	
theta=-0.5:0.5	KL	MAP	0,30	118,63	41,36	0,975	0,215	-0,001	0,046	0,300	0,334	
theta=-0.5:0.5	KL	MAP	0,50	13,63	10,43	0,896	0,429	-0,001	0,008	0,484	0,418	
theta=-0.5:0.5	KL	EAP	0,50	20,20	11,16	0,904	0,419	0,013	0,009	0,485	0,407	
theta=-0.5:0.5	KL	ML	0,50	21,79	17,41	0,930	0,415	-0,027	0,012	0,483	0,381	
theta=-0.5:0.5	KL	ML	25 madde	33,00	25,00	0,947	0,390	0,043	0,028	0,396	0,323	
theta=-0.5:0.5	KL	MAP	25 madde	33,08	25,00	0,955	0,287	-0,017	0,029	0,342	0,339	
theta=-0.5:0.5	KL	EAP	25 madde	48,25	25,00	0,955	0,291	-0,004	0,029	0,348	0,335	
theta=-0.5:0.5	KL	ML	0,30-25 madde	32,89	24,88	0,947	0,391	0,043	0,028	0,396	0,323	
theta=-0.5:0.5	KL	MAP	0,30-25 madde	38,57	23,93	0,953	0,292	-0,017	0,027	0,346	0,345	
theta=-0.5:0.5	KL	EAP	0,30-25 madde	41,15	24,34	0,953	0,295	-0,003	0,028	0,350	0,339	
theta=-0.5:0.5	KL	ML	0,50-25 madde	17,95	13,54	0,918	0,480	0,007	0,011	0,514	0,380	
theta=-0.5:0.5	KL	EAP	0,50-25 madde	18,16	10,69	0,901	0,423	0,010	0,009	0,487	0,407	
theta=-0.5:0.5	KL	MAP	0,50-25 madde	19,12	9,94	0,894	0,433	-0,017	0,008	0,486	0,418	
theta=0	MEI	ML	0,30	140,00	48,23	0,980	0,212	-0,017	0,598	0,304	0,310	
theta=0	MEI	MAP	0,30	150,00	41,20	0,975	0,215	0,000	0,592	0,300	0,333	

Başlama kuralı	Madde seçim yöntemi	Yetenek kestirim yöntemi	Sonlandırma kuralı	Simülasyon süresi	Ortalama test uzunluğu	Gerçek				Ortalama standart hata	Ortalama madde bilgisi
						yetenek ile kestirilen yetenek korelasyon	RMSE	Bias	Madde örtüşme oranı		
theta=0	MEI	EAP	0,30	685,00	24,30	0,953	0,296	-0,001	0,645	0,349	0,341
theta=0	MEI	MAP	0,50	31,93	10,05	0,894	0,432	-0,013	0,439	0,483	0,416
theta=0	MEI	ML	0,50	72,12	17,70	0,930	0,416	-0,028	0,358	0,485	0,358
theta=0	MEI	EAP	0,50	210,00	10,86	0,902	0,423	0,012	0,448	0,483	0,409
theta=0	MEI	MAP	25 madde	107,00	25,00	0,955	0,287	-0,018	0,663	0,340	0,339
theta=0	MEI	ML	25 madde	115,12	25,00	0,946	0,397	0,048	0,570	0,398	0,324
theta=0	MEI	EAP	25 madde	575,00	25,00	0,954	0,293	-0,002	0,651	0,347	0,337
theta=0	MEI	MAP	0,30-25 madde	100,07	23,90	0,953	0,292	-0,018	0,653	0,345	0,345
theta=0	MEI	ML	0,30-25 madde	110,12	24,93	0,946	0,398	0,048	0,598	0,398	0,325
theta=0	MEI	EAP	0,30-25 madde	563,12	24,30	0,953	0,296	-0,001	0,645	0,349	0,341
theta=0	MEI	MAP	0,50-25 madde	30,04	9,62	0,893	0,435	-0,014	0,457	0,485	0,416
theta=0	MEI	ML	0,50-25 madde	85,10	13,78	0,917	0,487	0,011	0,406	0,517	0,358
theta=0	MEI	EAP	0,50-25 madde	108,00	10,46	0,900	0,427	0,010	0,463	0,486	0,410
theta=-0.5:0.5	MEI	MAP	0,30	120,12	41,18	0,975	0,215	0,000	0,047	0,300	0,336
theta=-0.5:0.5	MEI	ML	0,30	188,25	48,23	0,980	0,213	-0,017	0,056	0,304	0,311
theta=-0.5:0.5	MEI	EAP	0,30	754,12	42,70	0,977	0,211	0,006	0,050	0,300	0,329
theta=-0.5:0.5	MEI	MAP	0,50	32,73	10,10	0,894	0,432	-0,014	0,009	0,482	0,434
theta=-0.5:0.5	MEI	ML	0,50	74,15	17,71	0,931	0,412	-0,034	0,012	0,483	0,367
theta=-0.5:0.5	MEI	EAP	0,50	135,57	10,87	0,900	0,426	0,005	0,009	0,484	0,420
theta=-0.5:0.5	MEI	MAP	25 madde	88,29	25,00	0,955	0,288	0,019	0,032	0,341	0,341
theta=-0.5:0.5	MEI	ML	25 madde	115,25	25,00	0,946	0,396	0,043	0,028	0,379	0,319
theta=-0.5:0.5	MEI	EAP	25 madde	536,12	25,00	0,954	0,293	-0,004	0,031	0,347	0,337
theta=-0.5:0.5	MEI	MAP	0,30-25 madde	89,13	23,90	0,953	0,293	-0,018	0,302	0,345	0,347
theta=-0.5:0.5	MEI	ML	0,30-25 madde	123,12	24,93	0,946	0,396	0,043	0,027	0,398	0,320
theta=-0.5:0.5	MEI	EAP	0,30-25 madde	605,12	24,31	0,953	0,296	-0,004	0,030	0,349	0,341
theta=-0.5:0.5	MEI	MAP	0,50-25 madde	49,12	9,66	0,892	0,436	-0,015	0,009	0,484	0,434
theta=-0.5:0.5	MEI	ML	0,50-25 madde	52,02	13,80	0,918	0,483	0,004	0,011	0,515	0,366
theta=-0.5:0.5	MEI	EAP	0,50-25 madde	224,32	10,45	0,898	0,430	0,004	0,009	0,486	0,421

EK-12. ETİK KURUL VE MEB ARAŞTIRMA İZİN BELGELERİ

Evrak Tarih ve Sayısı: 29.11.2023-791769



T.C
AKDENİZ ÜNİVERSİTESİ REKTÖRLÜĞÜ
Sosyal ve Beşeri Bilimler Bilimsel Araştırma ve Yayın Etiği Kurulu
KURUL KARARI



TOPLANTI TARİHİ : 28.11.2023

TOPLANTI SAYISI : 21

KARAR SAYISI : 519

Üniversitemiz Eğitim Fakültesi Eğitim Bilimleri Bölümü öğretim elemanı Arş. Gör. İbrahim Hakkı TEZCİ'nin araştırmacılığını, Doç. Dr. Bayram BIÇAK'ın danışmanlığını üstlendiği, *"Farklı Yetenek Kestirim Yöntemleriyle Bilgisayar Ortamında Bireye Uyarlanmış Testlerin (BOBUT) Çeşitli Koşullara Göre İncelenmesi"* konulu çalışmanın, fikri hukuki ve telif hakları bakımından metot ve ölçeğine ilişkin sorumluluğun başvurucaya ait olmak üzere, proje süresince uygulanmasının etik olarak uygun olduğuna oy birliği ile karar verilmiştir.

Prof. Dr. Hilmi DEMİRKAYA
Kurul Başkanı

Başkan
Prof. Dr.
Hilmi DEMİRKAYA

Başkan Yrd.
Prof. Dr.
Sibel PAŞAOĞLU YÖNDEM

Üye
Prof. Dr.
Ebru İÇİGEN

Üye
Prof. Dr.
Nurşen ADAK

Üye
Prof. Dr.
Taner KORKUT

Üye
Prof. Dr.
Gökhan AKYÜZ

Üye
Prof.
Ceren HEPYÜCEL



T.C.
AKDENİZ ÜNİVERSİTESİ REKTÖRLÜĞÜ
Eğitim Bilimleri Enstitüsü Müdürlüğü



Sayı : E-36380087-605.01-813518
Konu : İbrahim Hakkı TEZCİ Araştırma İzni

27.12.2023

DAĞITIM YERLERİNE

Eğitim Bilimleri Ana Bilim Dalı Eğitimde Ölçme ve Değerlendirme Doktora Programı Prof.Dr. Bayram BIÇAK danışmanlığındaki 202054027001 numaralı öğrencisi İbrahim Hakkı TEZCİ'nin "Farklı Yetenek Kestirim Yöntemleriyle Bilgisayar Ortamında Bireye Uyarlanmış Testlerin (BOBUT) Çeşitli Koşullara Göre İncelenmesi" isimli araştırmasını İlimiz merkez ilçelerinde bulunan resmi-özel ortaokul 7. ve 8.sınıf öğrencilerine uygulama isteğinin uygun görüldüğüne ilişkin Antalya İl Milli Eğitim Müdürlüğü'nün 18.12.2023 tarih ve 92346325 sayılı yazısı Ek'te gönderilmiştir.

Ayrıca araştırmanın bitiminde, sonuç raporunun bir örneğinin CD ortamında başvuru sahibinin İl Milli Eğitim Müdürlüğü'nün ekte örneği bulunan dilekçe ile Ar-Ge bürosuna gönderilmesi hususunda;

Bilgilerini ve gereğini rica ederim.

Doç. Dr. Güçlü ŞEKERCİOĞLU
Müdür

Ek: Antalya İl Milli Eğitim Müdürlüğü'nün Yazısı

Dağıtım:
Eğitim Bilimleri Anabilim Dalı Başkanlığına
Sayın Prof. Dr. Bayram BIÇAK

ÖZ GEÇMİŞ

Kişisel Bilgiler

Adı Soyadı: İbrahim Hakkı TEZCİ

Eğitim Durumu

Lisans: Dokuz Eylül Üniversitesi, Buca Eğitim Fakültesi, Sınıf Eğitimi (2008-2012)

Yüksek Lisans: Akdeniz Üniversitesi, Eğitim Bilimleri Enstitüsü, Eğitimde Ölçme ve Değerlendirme Ana Bilim Dalı, Eğitimde Ölçme ve Değerlendirme Yüksek Lisans Programı (2017-2020)

Doktora: Akdeniz Üniversitesi, Eğitim Bilimleri Enstitüsü, Eğitimde Ölçme ve Değerlendirme Ana Bilim Dalı, Eğitimde Ölçme ve Değerlendirme Doktora Programı (2020-)

Bilimsel Faaliyetler

Makaleler

Asma, B., & Tezci, İ. H. (2023). Academician perspective on the implementation of distance education in universities during the COVID-19 pandemic. *Journal of Qualitative Research in Education*, (34). 86-113. <https://doi.org/10.14689/enad.34.912>

Canbulat, T., Kırıktaş, H., Tezci, İ. H., & İlhan, E. (2020). İlkokula hazırbulunuşluk ölçeği: yapısal güçlendirme çalışması. *International Journal of Current Approaches in Language, Education and Social Sciences*, 2(2), 675-687.

Canbulat, M., Direkci, B., Çorapçıgil, A., Şimşek, E. E., Asma, B., Tezci, İ. H., & Şimşek, B. (2020). The psychometric properties of school belonging scale for primary school students: A validity and reliability study. *Elementary Education Online*, 19(3), 1422-1438.

Canbulat, M., Direkçi, B., Akbulut, S., Çorapçıgil, A., Tezci, İ. H., Şimşek, E. E., & Asma, B. (2020). Okul müdürlerinin gözünden “Okul 2023” projesinin incelenmesi. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, 20(1), 570-584.

Direkci, B., Canbulat, M., Tezci, I., & Akbulut, S. (2020). The psychometric properties of school belonging scale for middle school students. *International Journal of Assessment Tools in Education*, 7(2), 159-176.

Bildiriler

Tezci, İ. H., & Bıçak, B. (2024). Yapay zekâ ve nitel veri: Geleceğin analiz yöntemleri. *Uluslararası Ölçme, Seçme ve Yerleştirme Sempozyumu*, 04-06 Ekim 2024, Ankara, Türkiye.

Tezci, İ. H. (2024). Yapay zekâ destekli değerlendirme: Açık uçlu sınav sorularının puanlanmasında bir karşılaştırma. 9. *Uluslararası Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi*, EPOD – 2024, Eskişehir, Türkiye.

Tezci, İ. H. (2024). AI CAT-GPT: Yapay zeka destekli bilgisayar ortamında bireye uyarlanmış test uygulamaları. 9. *Uluslararası Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi*, EPOD – 2024, Eskişehir, Türkiye.

Tezci, İ. H., & Bıçak, B. (2023). Kağıt kalem ve bilgisayar ortamında bireye uyarlanmış testlerinin karşılaştırılması. 1. *Bireye Uyarlanmış Test Araştırmaları Ulusal Sempozyumu*, Boğaziçi Üniversitesi, İstanbul,

Tezci, İ. H., Bıçak, B., & Çobanoğlu Aktan, D., (2020). Öğretmenlerin ölçme davranışlarının geçerliğinin incelenmesi . 7.*Uluslararası Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi*, Ankara, Türkiye.

Tezci, İ. H., (2019). Öğretmen Adaylarına Yönelik Tutum Ölçeği Geliştirme Çalışması . I. *Uluslararası Bilim, Eğitim, Sanat ve Teknoloji Sempozyumu* (pp.40-43). İzmir, Türkiye

Tezci, İ. H., (2019). Kümeleme Analizi, Açıklayıcı ve Doğrulamalı Faktör Analizi ile Elde Edilen Yapı Geçerliliği Sonuçlarının Karşılaştırılması . *INES* (pp.30-32). Antalya, Türkiye.

Projeler

Eğitim Bilimleri Alanında Lisansüstü Eğitimine Devam Eden Öğrencilerin Çevrimiçi Ölçme ve Değerlendirme Becerilerinin Geliştirilmesi (Eğitmen, TÜBİTAK 2237 Bilimsel Eğitim Etkinliklerini Destekleme Programı)

Kitap Bölümleri

Tezci, İ. H. (2024). Hayat bilgisi dersinde ölçme ve değerlendirme. S. Z. Genç, T. Canbulat, & F. M. Özbilen (Eds.), *Hayat bilgisi dersinde güncel konu ve araştırmalar*. Vizetek Yayıncılık.

Tezci, İ. H. (2022). Webb'in bilgi derinliği taksonomisi. E. Yeşilyurt (Ed.), *Eğitim taksonomileri* (ss. 183-202). Vizetek.

Tezci, İ. H. (2022). Geniş ölçekli sınavların bilişsel yapı taksonomisi. E. Yeşilyurt (Ed.), *Eğitim taksonomileri* (ss. 203-224). Vizetek.

BENZERLİK RAPORU

FARKLI YETENEK KESTİRİM YÖNTEMLERİYLE BİLGİSAYAR ORTAMINDA BİREYE UYARLANMIŞ TESTLERİN ÇEŞİTLİ KOŞULLARA GÖRE İNCELENMESİ

ORJİNALLİK RAPORU

% 5	% 4	% 3	% 0
BENZERLİK ENDEKSİ	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCİ ÖDEVLERİ

BİRİNCİL KAYNAKLAR

1	acikbilim.yok.gov.tr İnternet Kaynağı	% 1
2	docplayer.biz.tr İnternet Kaynağı	% 1
3	kitaplar.ankara.edu.tr İnternet Kaynağı	% 1
4	openaccess.hacettepe.edu.tr İnternet Kaynağı	<% 1
5	"Keyword Index", Neuropsychopharmacology, 2012. Yayın	<% 1
6	dspace.ankara.edu.tr İnternet Kaynağı	<% 1
7	www.openaccess.hacettepe.edu.tr:8080 İnternet Kaynağı	<% 1
8	Özbaşı, Durmuş. "Bilgisayar Okuryazarlığı Testinin Bilgisayar Ortamında Bireye	<% 1

BİLDİRİM

Hazırladığım tezin tamamen kendi çalışmam olduğunu ve her alıntıya kaynak gösterdiğimi taahhüt eder, tezimin kâğıt ve elektronik kopyalarının Akdeniz Üniversitesi Eğitim Bilimleri Enstitüsü arşivlerinde saklanmasına izin verdiğimi onaylarım.

- ☐ Tezimin/Raporumun tamamı her yerden erişime açılabilir.
- ☐ Tezim/Raporum sadece Akdeniz Üniversitesi yerleşkelerinden erişime açılabilir.
- ☐ Tezimin/Raporumun 6 ay süreyle erişime açılmasını istemiyorum. Bu sürenin sonunda uzatma için başvuruda bulunmadığım takdirde, tezimin/raporumun tamamı her yerden erişime açılabilir.

Şubat 2025

İbrahim Hakkı TEZCİ