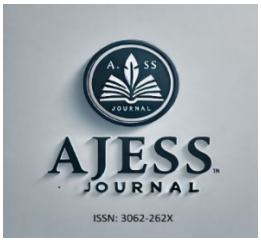
2026, Cilt 3, Sayı 1, 1-12

OPEN ACCESS
ISSN: 3062-262X



Akademik Eğitim ve Sosyal Bilimler Dergisi

(AJESS Journal)

Bireyselleştirilmiş Bilgisayar Testlerinde Madde Seçim Ve Yetenek Kestirim Yöntemlerinin Karşılaştırılması: Simülasyon Tabanlı Bir Çalışma

Comparison Of Item Selection and Ability Estimation Methods In Computerized Adaptive Testing: A Simulation-Based Study

Şeyma ERBAY MERMER , Öğr. Gör. Dr., Bilecik Şeyh Edebali Üniversitesi, sey.erbay@gmail.com

Makale Tarihçesi/Article History

Geliş/Received: 16/02/2025 Düzeltme/Revision: 23/04/2025 Kabul/Accepted: 24/06/2025

Abstract

In this study, the performances of different item selection and ability estimation methods were compared through simulation with Computerized Adaptive Testing. In this context, estimation errors of item selection and ability parameters were calculated using a simulated dataset based on the three-parameter logistic model, and simulations were conducted via SimulCAT software. In the study, Maximum Likelihood Estimation and Bayesian estimation methods were compared for ability estimation; the accuracy of ability estimation and the average standard errors were calculated for each method using Maximum Fisher Information and Maksimum Likelihood Weighted Information as item selection methods. During the testing process, interim theta values of individuals were also examined, and it was recorded which method yielded better results. A fixed-length rule of 20 items was preferred as the test termination criterion, and the test was terminated for each individual after answering 20 questions. The application was carried out with a total of 1,000 individuals, and item pool consisted of 500 items. The average of the results obtained from 25 replications for each method was used in the analyses. According to results, in the Maximum Likelihood Estimation method, the Maksimum Likelihood Weighted Inform function yielded the most accurate ability estimation with the least error; in the Bayesian method, the Maximum Fisher Information criterion provided the most accurate estimation. In interim ability estimation as well, it can be stated that the Maksimum Likelihood Weighted Inform function in the Maximum Likelihood Estimation method and the Maximum Fisher Information criterion in the Bayesian method yielded better results.

Keywords: Computerized adaptive testing, ıtem selection methods, ability estimation

Öz

Mevcut araştırmada, Bireyselleştirilmiş Bilgisayar Test ortamında farklı madde seçme ve yetenek kestirim yöntemlerinin simülasyon verileri aracılığıyla performansları karşılaştırılmıştır. Bu kapsamda, 3 parametreli lojistik model esas alınarak oluşturulan simülatif veri seti ile madde seçim ve yetenek parametrelerinin kestirim hataları hesaplanmış ve simülasyonlar SimulCAT yazılımı aracılığıyla yürütülmüştür. Araştırmada yetenek kestirimi için Maksimum Olabilirlik Kestirim ve Bayesyan yöntemleri karşılaştırılmış; her bir yöntemin soru seçiminde Maksimum Bilgi Ölçütü ve Ağırlıklandırılmış Maksimum Olabilirlik Fonksiyonu yöntemleri ile yetenek kestirimi doğruluğu ve kestirime ilişkin standart hata ortalamaları hesaplanmıştır. Test sürecinde, bireylerin interim theta değerlerine de bakılmış hangi yöntemin daha iyi sonuçlar verdiği kaydedilmiştir. Test sonlandırma kuralı olarak 20 soruluk sabit soru sayısı tercih edilmiş her 20 soruyu cevaplayan birey için test sonlandırılmıştır. Toplam 1000 kişi ile gerçekleştirilen uygulamada soru havuzu toplamda 500 sorudan oluşmuş belirlenen her bir yöntem için 25 replikasyon sonucunda elde edilen değerlerin ortalaması analizlerde kullanılmıştır. Sonuçlara bakıldığında ise yetenek kestirimi için Maksimum Olabilirlik Kestirim Yönteminde Ağırlıklandırılmış Maksimum Olabilirlik Fonksiyonunun, Bayesyan Yönteminde ise Maksimum Bilgi Ölçütü'nün en az hata vererek en iyi yetenek kestirimi yaptığı gözlenmiştir. Ara yetenek kestirim yöntemlerinde de Maksimum Olabilirlik Kestirim Yönteminde Ağırlıklandırılmış Maksimum Olabilirlik Fonksiyonunun, Bayesyan Yönteminde ise Maksimum Bilgi Ölçütü'nün daha iyi sonuçlar verdiği söylenebilir.

Anahtar Kelimeler: Bireyselleştirilmiş bilgisayar testi, madde seçim yöntemleri, yetenek kestirimi

Atıf için: Erbay Mermer, Ş. (2026) Bireyselleştirilmiş bilgisayar testlerinde madde seçim ve yetenek kestirim yöntemlerinin karşılaştırılması: Simülasyon tabanlı bir çalışma, *Akademik Eğitim ve Sosyal Bilimler Dergisi (AJESS Journal), 3*(1), 1-12. doihttps://doi.org/10.5281/zenodo.15829038

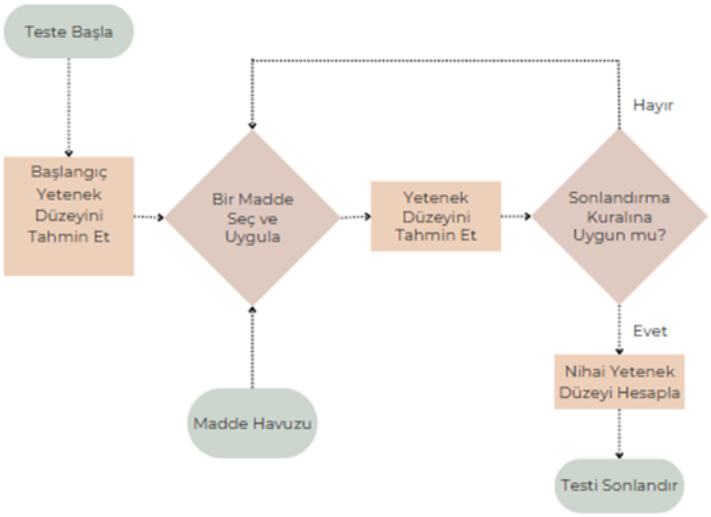
© Erişim ve Kullanım haklarına https://www.ajessjournal.com/index.php/ajess/acikerisim adresinden ulaşabilirsiniz.

Giriş

Teknolojik gelişmelerin etkisiyle eğitimde ölçme ve değerlendirme uygulamaları son yıllarda önemli gelişmeler yaşanmakta ve bu alana yönelik ilgi giderek artmaktadır. Bu gelişmeler ölçme ve değerlendirme süreçlerinde bireysel farklılıkların dikkate alındığı, daha esnek, daha duyarlı ve birey odaklı yenilikçi test uygulamalarına ihtiyacı artmıştır. Bu doğrultuda geliştirilen ve bireylerin yetenek düzeylerindeki farklılıkları esas alarak kişiye özgü test uygulamalarına imkan tanıyan Bireyselleştirilmiş Bilgisayar Testleri (BBT), geleneksel kâğıt-kalem testlerine alternatif bir ölçme aracı olarak öne çıkmakta; test süresini kısaltmakta ve yetenek kestiriminde daha yüksek düzeyde doğruluk sağlamaktadır (Meijer & Nering, 1999; Embretson & Reise, 2000). Bu testler, Madde Tepki Kuramına (MTK) dayalı olarak geliştirilen ve test sürecini bireyin yetenek düzeyine göre dinamik biçimde yönlendiren uygulamalardır. Bu yönüyle BBT, bireylerin gerçek yetenek düzeylerine daha kısa sürede ulaşmayı hedeflemekte, ölçme sürecini daha etkili ve verimli hâle getirmektedir (Kezer & Koç, 2014).

İlk olarak testlerin bilgisayar ortamına aktarılmasıyla başlayan süreç, teknolojik gelişmelere paralel olarak zamanla önemli bir dönüşüm geçirmiştir. Farklı soru formatlarının uygulanabilir hâle gelmesi ve testin tamamlanmasının ardından bireylere anında geri bildirim verilmesi gibi çeşitli avantajlar sayesinde, BBT uygulamaları geleneksel yöntemlere kıyasla daha fazla tercih edilmeye başlanmıştır (Kalender, 2009). Bilgisayar ortamında sunulan testler, esasen kağıt-kalem testlerinin dijital ortama uyarlanmış bir biçimidir ve doğrusal test yapısı kapsamında, tüm katılımcılara aynı sırada ve önceden belirlenmiş sabit bir madde dizisiyle uygulanmaktadır. Bu tür testler, madde puanlama açısından geleneksel testlerle benzerlik gösterse de temel farklılık testin uygulama ortamından kaynaklanmaktadır (Eroğlu & Kelecioğlu, 2015). BBT uygulamalarında ise kağıt-kalem testlerinde karşılaştığımız şekilde standart bir soru formu bulunmamaktadır. BBT, testin uygulanma sürecinin bireylerin yetenek düzeylerine göre biçimlendirilmesine olanak tanır. Bir başka deyişle; bilgisayar, tüm katılımcılara önceden belirlenmiş bir soru setini sıra ile uygulamak yerine, soruları testin gidişatına (testi cevaplayanların verdiği cevaba) göre seçerek sormaktadır (Kalender, 2004).

BBT uygulamalarında testin ölçme gücünü etkileyen en temel değişkenler arasında madde seçim yöntemleri, yetenek kestirim yöntemleri ve test sonlandırılma kuralları yer almaktadır (van der Linden & Glas, 2000; Weiss, 1982). BBT uygulamasında süreç bilgisayar tarafından gerçekleştirilmektedir. Test sürecine ilişkin olarak elde edilen bu üstünlükler BBT uygulamasının birey-madde etkileşimine getirdiği yaklaşım sayesinde elde edilmektedir. BBT uygulamasındaki birey-madde etkileşimini ve BBT uygulamasının işleyişini gösteren akış diyagramı aşağıdaki şekilde verilmiştir.



Şekil 1. BBT akış diyagramı

Şekil 1'de yer alan BBT akış diyagramı, test sürecinin başından sonuna kadar izlenen adımları ve bu adımlar arasındaki işleyişi sistematik biçimde açıklamaktadır. Katılımcıların yetenek tahmini ile başlayan test süreci seçilen maddeye verdiği cevaplara göre güncellenmekte ve bu süreç test sonlandırma kuralı gerçekleşene kadar devam etmektedir. Belirlenen sonlandırma kuralı sağlandığında ise bireyin hangi yetenek düzeyinde olduğu saptanmaktadır (Weiss, 1985; van der Linden & Glas, 2000).

Test başlama kuralı, bireyin test başlangıcında yetenek (theta(θ)) kestiriminin nasıl yapılacağı ve bu kestirime uygun olarak ilk test maddesinin nasıl seçileceğini belirleyen tüm süreçleri kapsar. Başlangıçta θ değeri rastgele seçilebilir, sabit bir değer olarak atanabilir (örneğin 0,00) veya elde buna uygun bir bilgi varsa ek verilere göre belirlenebilir. Bu kural, testin adaptif sürecinin doğru başlatılması açısından kritik bir rol oynar ve testin genel ölçme doğruluğunu doğrudan etkiler (Weiss, 1982; van der Linden & Glas, 2000).

Madde seçme kuralı, BBT uygulamalarında bireyin yeteneğine en uygun maddenin tespit edilmesinde kullanılan bir yöntemdir. Bu süreç, bireyin mevcut yetenek kesimine göre madde havuzundan yetenek kestirimi yüksek ve birey hakkında en fazla bilgiyi verecek maddenin seçilme sürecini kapsar (Weiss, 1982; van der Linden & Glas, 2000; Eggen & Kauchak, 2004). Bu durum mevcut testin ölçme doğruluğunu ve gerçek yetenek düzeyinin ne derece yansıtıldığını doğrudan etkiler.

Test sonlandırma kuralı, bireyin yetenek kestiriminin daha fazla madde sunulmasına gerek kalmadan hangi noktada sonlandırılacağını belirleyen bir ölçüttür. Bu kural, belirlenen sabit madde sayısına ulaşılması, yetenek kestirimindeki hatanın belirlenen bir düzeyin altına inmesi, madde havuzunun tükenmesi gibi farklı ölçütlere dayanabilir bu da testin gereksiz yere uzamasının önüne geçer (van der Linden & Glas, 2000; Wainer, 2000).

Özellikle simülasyon ortamında yürütülen çalışmalarda, bu değişkenlerin ölçme kesinliği üzerindeki etkilerini belirlemek amacıyla farklı stratejilerin karşılaştırılması önem arz etmektedir. Literatürde, sabit test uzunluğu, belirli hata sınırına ulaşma ya da belirli güvenirlik düzeyine erişme gibi farklı sonlandırma kurallarının; en fazla bilgi, rastgele seçim ya da karma stratejilere dayalı madde seçim algoritmalarıyla birlikte değerlendirilmesi, BBT sistemlerinin etkinliğini belirlemeye yönelik olarak sıklıkla kullanılmaktadır (Kezer & Koç, 2014).

Mevcut araştırma kapsamında, BBT üzerinde çalışılmış; simülasyon ortamında oluşturulan veri seti aracılığıyla madde seçim yöntemlerinin ve test uzunluklarının yetenek kestirimi ve standart hata ortalamaları karşılaştırmaya çalışılmıştır. Simülatif veriler doğrultusunda, kestirilen yetenek parametreleri arasında farklı madde seçim yöntemlerine göre anlamlı bir farklılaşma olup olmadığı incelenmiştir.

Yöntem

Araştırmanın Modeli

Mevcut çalışma, Madde Tepki Kuramı (MTK) temelli BBT uygulaması simülasyonu şeklinde yapılandırılmıştır. Araştırmanın verileri Monte-Carlo simülasyon yoluyla elde edilmiştir. Analizler için Han (2012) tarafından geliştirilen SimulCAT programından yararlanılmıştır.

Verilerin Analizi

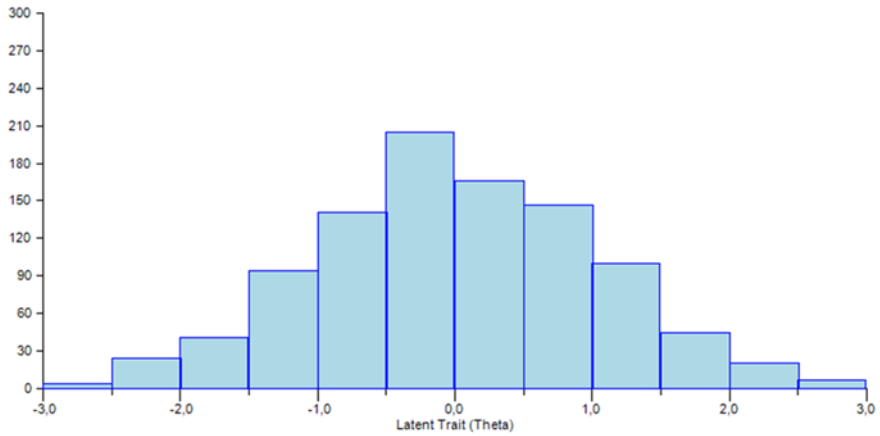
Simülasyon Koşulları

Simülasyon için madde havuzu toplam tek boyutlu 500 maddeden oluşturulmuş ve simülasyon senaryosu için her birey için toplamda 20 madde cevaplanması koşulu belirlenmiştir. Araştırmada ele alınan her koşul için üç parametreli lojistik modellere (3PLM) dayalı veriler elde edilmiştir. 3PLM, bireyin yetenek düzeyi ile maddenin ayırt edicilik parametresi (a_i), güçlük parametresi (b_i) ve şans parametresi (c_i) göz önünde bulundurularak bireyin doğru yanıt verme olasılığını tahmin eder (Baker, 2001; Embretson & Reise, 2000). a parametresi lognormal olarak belirlenmiş ortalaması 0, standart sapması 0,2 olacak şekilde düzenlenmiştir. b parametresinin ortalaması 0, standart sapması 1'dir. c parametresi uniform şekilde en düşük değeri 0 en yüksek değeri 0,2 şeklinde oluşturulmuştur. Her bir oturum tek sefer için 200 kişinin oturuma katılması ve her gün 2 oturum yapılması belirlenen koşullar arasındadır. Ayrıca belirlenen her yöntem için 25 replikasyon sonucunda elde edilen değerlerin ortalaması analizlerde kullanılmıştır. Simülasyon verilerinde 1000 kişi uygulama için belirlenmiş veri setine ait dağılımlar aşağıda verilmiştir.

Tablo 1. *Veri setine ait dağılımlar*

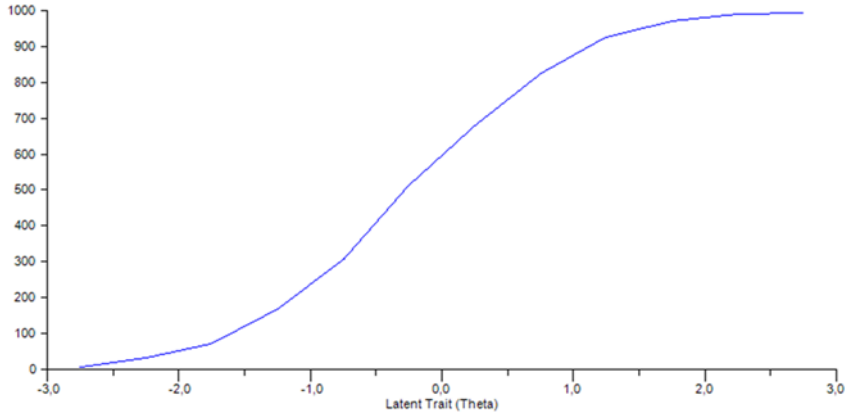
N	Ortalama	ss	Skewness	Kurtosis(ex)
1000	0,009	1,052	0,01	0,02

Yukarıdaki tablo incelendiğinde 1000 kişiye ait ortalamanın 0,009 olduğu standart sapmanın 1,052 olduğu gözlenmiştir. Ayrıca çarpıklık ve basıklık katsayılarının da -1,5 ve +1,5 arasında olduğu görülerek veri setinin normal dağılımda olduğu söylenebilir (Tabachnick & Fidell, 2013). Veri setinin yetenek düzeyine ait dağılım grafiği aşağıdaki gibidir.



Şekil 2. Yetenek düzeyine ait dağılım grafiği

Latent trait (theta(θ)), gizil yetenek düzeyi olarak geçen ve doğrudan gözlenemeyen ancak testler aracılığıyla dolaylı olarak ölçülen bireysel özelliklerdir (Carter & Wilkinson, 1984). Şekil 2'de SimulCAT ile üretilmiş 1000 kişiye ait veri setinin θ değerleri yer almaktadır. Histogram grafiği incelendiğinde dağılımın çan eğrisi formunda olduğu ve normal dağılım özellikleri gösterdiği söylenebilir. Kümülatif olarak bireylerin yetenek kestirimlerinin grafiği aşağıdaki gibidir.



Şekil 3. Yetenek düzeyine ait kümülatif dağılım grafiği

Sigmoid eğriler (S şeklinde), normal dağılım gibi birçok sürekli dağılımın kümülatif dağılım fonksiyonlarında görülür (Hunter, 2020). Şekil 3'teki dağılım grafiği normal dağılımın kümülatif dağılım fonksiyonunun karakteristik yapısıyla tutarlıdır.

Madde Seçim Yöntemleri

Tek boyutlu CAT uygulamalarında madde seçim yöntemleri birbirinden farklılık göstermektedir. Bu yöntemler Geleneksel Yöntemler ve Bayesyan yöntemler olmak üzere iki farklı yaklaşıma göre gruplanmaktadır (Ho, 2010; Thissen & Mislevy, 2000). Fisher'in Maksimum Bilgi Ölçütü (MFI) ve Ağırlıklandırılmış Maksimum Olabilirlik Fonksiyonu Geleneksel Yöntemlerde yer almakta, Sonsal Dağılımla Ağırlıklandırılmış Beklenen Maksimum Bilgi ve Beklenen Maksimum Bilgi, Beklenen Minimum Sonsal varyans ve Sonsal Dağılımla Ağırlıklandırılmış Beklenen Maksimum Bilgi de Bayesyan Yöntemler arasındadır (Fisher, 1925; Gelman et al., 2013; Overstall, 2020).

En sık kullanılan madde seçme yöntemlerinden ilki MFI yöntemidir. Bu yöntemde belirli bir yetenek düzeyi için en yüksek bilgiyi (yetenek düzeyi kestirimi standart hatasında en büyük düşüşü) sağlayacak olan madde seçilir (Kalender, 2009). Ağırlıklandırılmış Maksimum Olabilirlik Fonksiyonu (MLWI) maddenin seçiminde olabilirlik fonksiyonu ile ağırlıklandırılmış maksimum bilgiyi kullanan yöntemdir. Bu yöntem maksimum bilgi veren maddenin belirlenmesinde, Fisher'ın bilgi fonksiyonu ve olabilirlik fonksiyonundan elde edilecek bilgiyi esas almaktadır (Veerkamp & Berger, 1997).

Bu çalışma kapsamında Veerkamp ve Berger (1997) tarafından geliştirilen Fisher'in Maksimum Bilgi Ölçütü (The Maksimum Fisher Information- MFI) ve Eggen (1999) tarafından geliştirilen Ağırlıklandırılmış Maksimum Olabilirlik Fonksiyonu (The Maksimum Likelihood Weighted Information-MLWI) yöntemlerinin karşılaştırılması yapılacaktır.

Yetenek Kestirim Yöntemleri

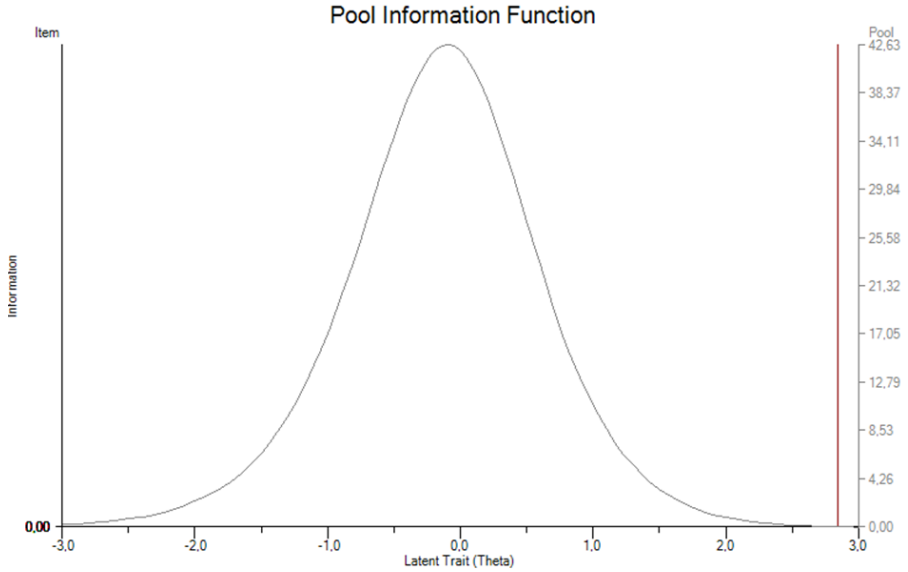
BBT uygulamalarında bireye ait yetenek parametreleri bireyin maddelere verdiği her bir yanıttan sonra güncellenir. Daha sonra bu güncellenen yetenek parametresine bağlı olarak belirlenen madde seçme ölçütü kullanılarak en uygun madde madde havuzundan seçilir. Alan yazına bakıldığında en yaygın olarak kullanılan yetenek kestirim yöntemleri Maksimum Olabilirlik Kestirim Yöntemi (MLE) ve Bayesyan yöntemleridir (Bejar ve Weiss, 1979; Song, 2012). MLE genellikle daha az yanlılık göstermekte olup Bayesyan yöntemine göre daha büyük hata vermektedir. Mevcut çalışmada bu durum göz önüne alınarak Maksimum Olabilirlik Kestirim yöntemi ile Bayesyan yöntemi karşılaştırması yapılacaktır.

Çalışma kapsamında 3PLM oluşturularak kurulan modelin tek boyutlu olması SimulCAT programında belirlenmiştir. Bu durum testin tek boyutluluk ölçümünün ayrıca incelenmesine gerek yoktur. 3PL ve tek boyutlu 500 sorudan oluşan test havuzuna ait *a, b* ve *c* parametre ortalamaları aşağıdaki gibidir.

Tablo 2. *Veri setine ait dağılımlar*

a parametresi	ss	*b* parametresi	ss	*c* parametresi	ss
1,024	0,212	0,027	0,967	0,099	0,060

b parametresi θ ile aynı metrik üzerindedir (DeMars, 2010). *b*=0 olduğunda maddenin ortalama yetenek düzeyindeki bireyler (θ ≈ 0) için %50 doğru cevaplanma olasılığına sahip olduğunu göstermektedir. Ancak bu parametreye ilişkin standart hata (ss = 0,967) oldukça yüksektir; bu durum, zorluk düzeyinin kararlı biçimde tahmin edilemediğini ve yorum yapılırken dikkatli olunması gerektiği anlamına gelir. *a* = 1,024 değerinin 1'in üzerinde olması maddenin bilenle bilmeyeni yeterli düzeyde ayırt ettiğini ortaya koymaktadır. *c* = 0,099, düşük yetenek düzeyine sahip bireylerin maddenin doğru yanıtını yaklaşık %10 oranında tahmin edebileceğini göstermektedir. Bu değer, çoktan seçmeli testlerde beklenen kabul edilebilir bir düzeydedir.



Şekil 4. Madde havuzu bilgi fonksiyonu

Madde Havuzu Bilgi Fonksiyonu (Pool Information Function (PIF)), tüm madde havuzunun belirli bir yetenek düzeyinde (θ) ne kadar bilgi sağladığını gösteren bir grafiktir (Hambleton, Swaminathan, & Rogers, 1991). Testin hangi yetenek düzeyinde daha hassas ölçüm yaptığını gösterir ve soruların bilgi fonksiyonları toplanarak elde edilir. Soru havuzuna ait bilgi fonksiyonu ile teste ait bilgi fonksiyonunun artması standart hatanın azaldığı anlamına gelmektedir.

Test Sonlandırma Kuralı

BBT uygulamalarında testin önceden belirlenmiş sabit bir soru sayısına ulaşıldığında sona erdirilmesini ifade eder (van der Linden & Glas, 2010). Bu yöntemde, her katılımcı aynı sayıda maddeye yanıt verir ve test uzunluğu sabittir. Mevcut çalışma kapsamında test sonlandırma kuralı olarak 20 maddelik sabit soru sayısı kullanılmıştır.

Araştırmanın Etik Izinleri

Simülasyon çalışmaları, insan katılımcıların doğrudan dahil olmadığı ve sadece yapay veriler veya bilgisayar ortamında oluşturulan verilerle yürütüldüğü için, etik kurul onayı gerektirmemektedir (Karasar, 2019). Bu bilgiler ışığında mevcut simülasyon çalışması için herhangi bir etik izne ihtiyaç duyulmamıştır.

Bulgular

MLE yetenek kestiriminde madde seçimlerinde iki yöntem incelenmiştir. Bunlar MFI ve MLWI yöntemleridir. Bu yöntemlere göre 25 replikasyon yapılmış bu replikasyon sonuçlarından elde edilen yetenek ve standart hatalardaki değişimler karşılaştırılmıştır.

Bireylere ait yetenek puanı ile kestirilen yetenek puanlarının tutarlığı yetenek parametrelerine ait RMSD (Root Mean Squared Difference) değerlerinin ortalamaları hesaplanarak belirlenmiştir. RMSD değerlerine ait formül aşağıdaki gibidir.

$$RMSD = \sqrt{\frac{\sum_{j=1}^{n}\left(\hat{\theta} - \theta\right)^2}{n}}$$

$\hat{\theta}$'kestirilen yetenek parametresini, θ gerçek yetenek parametresini ve n ise testi alan birey sayısını temsil etmektedir.

Tablo 3. *MLE yetenek kestiriminde madde seçim yöntemlerinin karşılaştırılması*

	θ Ortalaması	$\hat{\theta}$ Ortalaması	Standart Hata Ortalaması	RMSD
MFI	0,009	0,017	2,980	0,008
MLWI	0,009	0,012	2,457	0,007

Tablo 3'teki bulgulara göre, MLWI madde seçim yöntemi, MFI yöntemine kıyasla daha düşük standart hata değeri üretmiştir. Bu sonuçlar, MLWI yönteminin, test sürecinde bireylerin yetenek düzeylerini daha hassas biçimde kestirebildiğini göstermektedir. MLE yetenek kestirim yönteminde madde seçim yöntemlerinden MLWI standart hata olarak daha düşük çıkmış gerçek yetenek kestirimini daha iyi yapmıştır. Ayrıca kestirilen θ ve gerçek θ ile hesaplanan RMSD değerinin MLWI madde seçim yönteminde daha düşük olduğu hesaplanmıştır. Bu da MLWI yöntemiyle yapılan kestirimlerin gerçeğe daha yakın, yani daha isabetli olduğunu göstermektedir.

Tablo 4. *Bayesyan yetenek kestiriminde madde seçim yöntemlerinin karşılaştırılması*

	θ Ortalaması	$\hat{\theta}$ Ortalaması	Standart Hata Ortalaması	RMSD
MFI	0,009	0,013	0,231	0,005
MLWI	0,009	0,013	0,232	0,006

Tablodaki bulgulara göre, Bayesyan yetenek kestirimi için MFI madde seçim yöntemi, MLWI yöntemine kıyasla daha düşük standart hata değeri üretmiştir. Bu sonuçlar, MFI yönteminin, test sürecinde bireylerin yetenek düzeylerini daha hassas biçimde kestirebildiğini göstermektedir. Bayesyan yetenek kestirim yönteminde madde seçim yöntemlerinden MFI standart hata olarak daha düşük çıkmış gerçek yetenek kestirimini daha iyi yapmıştır. Ayrıca kestirilen $\hat{\theta}$ ve gerçek θ ile hesaplanan RMSD değerinin MFI madde seçim yönteminde daha düşük olduğu hesaplanmıştır. Bu da MFI yöntemiyle yapılan kestirimlerin gerçeğe daha yakın, yani daha isabetli olduğunu göstermektedir.

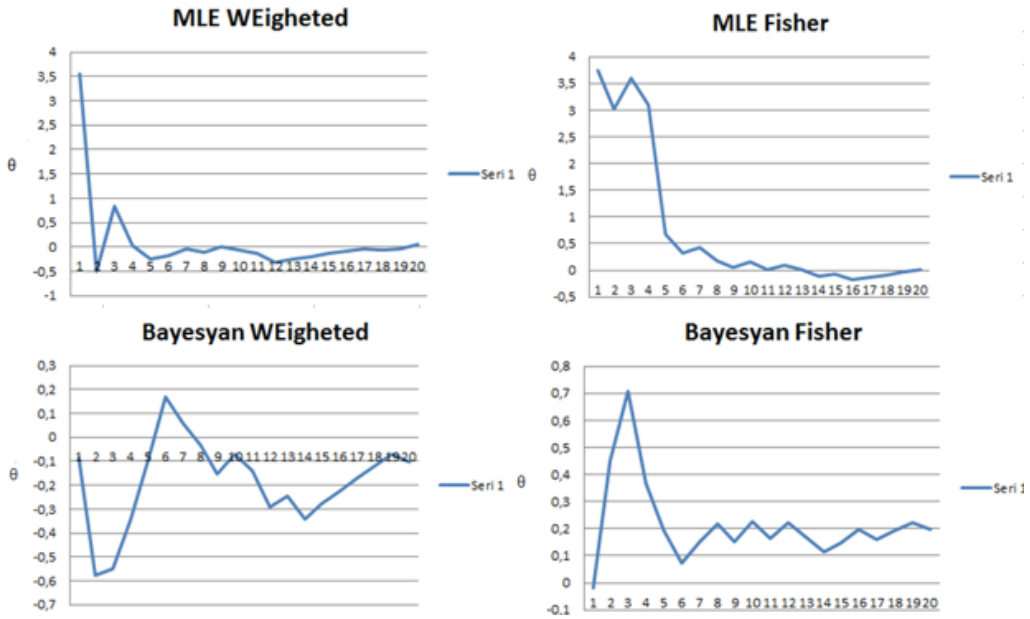
Alan yazında da belirtildiği gibi yetenek kestirimlerinde Bayesyan yönteminde hatanın daha düşük çıktığı mevcut analizde de görülmüştür. Bayesyan yetenek kestirimnde MFI madde seçim yönteminde en az hata ile en yüksek yetenek kestirimi yapıldığı yapılan analizler sonucunda gözlenmiştir. Yapılan analizler model uyumuna bakılması için gerçek θ ve kestirilen $\hat{\theta}$ korelasyonuna bakılmıştır. Bu hesaplamaya ait değerler aşağıdaki tabloda verilmiştir.

Tablo 5. *Modellere ait hesaplanan θ korelasyon değerleri*

Model	θ Ortalaması
MLE MLWI	0,814
MLE MFI	0,767
Bayesyan MLWI	0,973
Bayesyan MFI	0,975

Tabloda en iyi uyumun Bayesyan MFI yöntemine ait olduğu yüksek çıkan korelasyon katsayısıyla görülmektedir. Kestirilen $\hat{\theta}$ için bu yöntemin daha düşük hata verdiği yapılan analizlerde de görülmüştür. Genel olarak bakıldığında ise Bayesyan yönteminin MLE yöntemine göre daha iyi uyum sağladığı söylenebilir.

Çalışmada ayrıca ara θ kestirimleri (Interim theta) kestirimlerine bakılmıştır. Ara θ theta kestirimleri, CAT sürecinde, testin her adımında bireyin θ düzeyine ilişkin yapılan ara tahminlerdir. Yani, test tamamlanmadan önce her madde yanıtlandıkça θ değeri güncellenir ve bu güncel değerlere "interim theta estimates" (ara θ kestirimleri) denir (van der Linden & Glas, 2010; Rudner & Guo, 2009). Orta düzey bir yeteneğe ait bireyin ara θ kestirimleri 25 replikasyon için 20 sorudaki ortalamaları alınmış elde edilen bu değerlerle madde seçim yöntemleri ve yetenek kestirim yöntemlerine göre nasıl değiştiği gözlenmiştir. Bu ara θ kestirimlerine ait grafikler aşağıdaki gibidir.



Şekil 5. Modellere ait θ kestirim grafikleri

Test süreci boyunca bireyin ara θ tahminlerinin durumunu gösteren interim theta grafiği, başta yüksek dalgalanmaya sahip kestirimlerin test ilerledikçe stabil hale geldiğini göstermektedir. Testin başlarındaki dalgalanmalar her yeni soru ile yetenek kestirimini daha iyi yapmakta ve dalgalanma giderek azalmaktadır. Bu durum, testin bireye uyarlanabilir niteliği sayesinde birey hakkında bilgi kazandıkça tahminlerin güvenilirliğinin arttığını θ kestirimini daha güvenilir biçimde tahmin ettiğini göstermektedir.

Tartışma ve Sonuç

Mevcut çalışma kapsamında, BBT ortamlarında çeşitli madde seçme ve yetenek kestirim algoritmalarının performanslarını simülasyon verileri kullanarak karşılaştırmıştır. Analizler, 3PL modele dayalı olarak oluşturulan simülatif veri seti üzerinde SimulCAT yazılımı aracılığıyla yürütülmüştür. Başlangıçta θ değeri rastgele seçilmiş, yetenek kestirimi için MLE ve Bayesyan yaklaşımları kıyaslanmış, madde seçiminde ise MFI ve MLWI yöntemleri kullanılmıştır. Test sonlandırma kriteri olarak da sabit bir madde sayısı (20 madde) belirlenmiş ve 1000 katılımcı için 500 maddelik bir havuzdan seçilen sorularla toplam 25 tekrar (replikasyon) gerçekleştirilmiştir. Elde edilen sonuçlar, yetenek kestirimlerinde MLE yönteminin MLWI madde seçim yöntemi ile, Bayesyan kestirimin ise MFI yöntemi ile birlikte kullanıldığında en düşük kestirim hatası ve standart hata ortalamalarını sağladığını ortaya koymuştur. Ek olarak, ara theta (yetenek) değerleri incelendiğinde de MLE-MLWI ve Bayesyan-MFI kombinasyonlarının daha tutarlı ve doğru kestirimler sunduğu gözlemlenmiştir. Yapılan analizlerde model uyumuna bakılması için gerçek θ ve kestirilen $\hat{\theta}$ korelasyonuna bakılmış en iyi uyumun Bayesyan MFI yöntemine ait olduğu yüksek çıkan korelasyon katsayısıyla görülmektedir. Kestirilen $\hat{\theta}$ için bu yöntemin daha düşük hata verdiği yapılan analizlerde de görülmüştür.

Analizler sonucunda elde edilen bulgular, özellikle, MLE ile MLWI kombinasyonu ile Bayesyan kestirim ile MFI yönteminin daha düşük kestirim hataları ve standart hatalar yaptığı görülmüştür. Bu sonuçlar, Hambleton ve arkadaşlarının (1991) MTK ve bilgi fonksiyonları üzerine yaptıkları çalışmalarla uyumludur. Ayrıca, Bayesyen-MFI

kombinasyonunun gerçek ve kestirilen yetenek parametreleri arasındaki yüksek korelasyon ile en iyi model uyumunu sağlaması, Bayesyen yaklaşımların özellikle MTK için çok da büyük sayılmayacak örneklemlerde ve başlangıçta yetenek kestirim belirsizliğinde avantajlı olduğunu gösteren önceki araştırmaları (Embretson & Reise, 2000) desteklemektedir. Sonuç olarak, madde seçim ve yetenek kestirim algoritmalarının dikkatli seçimi, BBT uygulamalarında ölçme geçerliliği ve güvenirliği açısından kritik bir rol oynamaktadır.

Öneriler

Elde edilen bulgular, BBT'de madde seçimi ile yetenek kestirimi yöntemlerinin testin ölçüm doğruluğunu ve kestirim güvenilirliğini anlamlı şekilde etkileme potansiyeli olduğunu ortaya koymaktadır. Bu doğrultuda, madde seçim algoritmaları ile yetenek kestirim yöntemlerinin uyumlu bir şekilde tasarlanması, bireysel farklılıkların daha etkili ve doğru bir şekilde değerlendirilmesini mümkün kılmaktadır. Bu alanda ileriye yönelik yapılacak çalışmalarda, simülasyon sonuçlarının gerçek katılımcı verileri üzerinden doğrulanması büyük önem taşımaktadır. Ayrıca, farklı test sonlandırma yöntemlerinin performans üzerindeki etkilerinin sistematik olarak incelenmesi, BBT uygulamalarının etkinliğini artırmak açısından yol gösterici olacaktır.

Bu bağlamda, öneriler şu şekilde sıralanabilir:

1. Simülasyon tabanlı analiz sonuçlarının gerçek katılımcılı sınav verileri ile karşılaştırılması ve yorumlanması gerekmektedir.

2. Test sonlandırma kuralında sabit soru sayısının yanı sıra, ölçme duyarlılığına ve yetenek kestirim güvenilirliğine dayalı dinamik sonlandırma yöntemlerinin etkililiği araştırılmalıdır.

3. Farklı MTK temelli modellerinde ve test yapısına sahip veri setlerinde madde seçim ve yetenek kestirim yöntemlerinin karşılaştırılması yapılmalıdır.

4. BBT uygulamalarında kullanıcı deneyimi ve sınav süresi gibi pratik faktörler de göz önünde bulundurularak buna uygun stratejiler geliştirilmelidir.

Sonuç olarak, madde seçim ve yetenek kestirim yöntemlerinin bir bütün olarak ele alınması, BBT'nin ölçme geçerliliği ve güvenirliğini artırmada kritik bir rol oynamaktadır.

Kaynakça

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.

Bejar, I. I. & Weiss, D. J. (1979). *Computer programs for scoring test data with item characteristic curve models* (Research Rep. No. 79-1). Minneapolis: University of Minnesota, Depeartment of Psychology, Psychometric Methods Program.

Carter, J. E., & Wilkinson, L. (1984). A latent trait analysis of the MMPI. *Multivariate Behavioral Research, 19*(3), 385–407.

DeMars, C., 2010. *Item Response Theory: Understanding Statistics Measurement.* Prof. Dr. Hülya Kelecioğlu (çeviri editörü), Oxford University Press, Oxford, 3-31.

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

Eggen, P. D., & Kauchak, D. (2004). *Educational psychology: Windows on classrooms* (6th ed.). Pearson Prentice Hall.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Psychology Press.

Eroğlu, M. G., & Kelecioğlu, H. (2015). Bireyselleştirilmiş bilgisayarlı test uygulamalarında farklı sonlandırma kurallarının ölçme kesinliği ve test uzunluğu açısından karşılaştırılması. *Uludağ Üniversitesi Eğitim Fakültesi Dergisi, 28*(1), 31–52.

Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society, 22*(5), 700–725.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.

Han, K. T. (2012). *SimulCAT: Computerized adaptive testing simulation program (Version 1.0) [Computer software].* https://www.hantest.net/simulcat/

Ho, T. (2010). A Comparison of item selection procedures using different ability estimation methods in computerized adaptive testing based on the Generalized Partial Credit Model, [Unpublished doctoral dissertation]. University of Texas.

Hunter, G. (2020). *The Sigmoid Function.* 5 Eylül 2023 tarihinde https://blog.mbedded.ninja/programming/artificial-intelligence/the-sigmoid-function/?utm_source=chatgpt.com uzantısından erişilmiştir.

Kalender, İ. (2004). Bilgisayar ortamında bireyselleştirilmiş testlerin eğitimde kullanımı. XIII. Ulusal Eğitim Bilimleri Kurultayı. *İnönü Üniversitesi, Eğitim Fakültesi,* Temmuz 6-9, Malatya.

Kalender, I. (2009). CITO. *Egitim Kuram ve Uygulama, 5*, 39-48.

Karasar, N. (2019). *Bilimsel araştırma yöntemi: Kavramlar, ilkeler, teknikler* (23. baskı). Nobel Akademik Yayıncılık.

Kezer, F., & Koç, N. (2014). Bilgisayar ortamında bireye uyarlanmış test stratejilerinin karşılaştırılması [A comparison of computerized adaptive testing strategies]. *Eğitim Bilimleri Araştırmaları Dergisi, 4*(1), 145–174.

Meijer, R. R., & Nering, M. L. (1999). Multidimensional item response models [Monograph]. Lawrence Erlbaum Associates.

Overstall, A. M. (2020). *Properties of using Fisher information gain for Bayesian design of experiments*. https://arxiv.org/abs/2003.07315 uzantısından 10 Eylül 2022 tarihinde erişilmiştir.

Rudner, L. (1998). *An On-line, Interactive, Computer Adaptive Testing Mini Tutorial*. ERIC Clearinghouse on Assessment and Evaluation.

Rudner, L. M., & Guo, F. (2009). Computer adaptive testing for small scale programs and instructional systems. *Journal of Applied Testing Technology, 10*(1), 1–19.

Song, T. (2012). *The effect of fitting a tridimensional IRT model to multidimensional data in content-balanced computerized adaptive testing*. Unpublished Doctoral Dissertation. Michigan State University.

Tabachnick, B.G. & Fidell, L. S. (2013). *Using multivariate statistics*, 6th edition. Boston:Pearson.

Thissen, D., & Mislevy, R.J. (2000). *Testing algorithms.* In H. Wainer (Ed.). Computerized adaptive testing, (101-135). Lawrence Erlbaum Assc.

van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice*. Dordrecht: Kluwer Academic Publishers.

van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing.* Springer.

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22*(3), 203–226.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Routledge.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*(4), 473–492. https://doi.org/10.1177/014662168200600405

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

Extended Summary

Introduction

Technological advancements have enabled measurement and evaluation processes in education to become more flexible, responsive, and individualized. This transformation has increased interest in next-generation testing applications that consider individual differences. In this context, Computerized Adaptive Testing (CAT) has emerged as an alternative to traditional paper-and-pencil tests, offering shorter administration times and higher measurement precision by selecting items based on the individual's ability level (Meijer & Nering, 1999; Embretson & Reise, 2000).

Structured on the foundations of Item Response Theory (IRT), these tests dynamically select items in response to the examinee's answers, with ability (θ) estimates being updated after each response. The operation of CAT is guided by several key components: the test initialization rule, item selection algorithm, and termination criteria (van der Linden & Glas, 2000; Weiss, 1982). The adaptive nature of the test, which is based on the interaction between the individual and the items, provides significant advantages in terms of measurement validity and efficiency.

In this study, the accuracy of ability estimates and associated standard errors were compared across different item selection methods and test lengths using simulation data. The measurement performance of CAT systems was evaluated accordingly.

Method

In this study, a simulation-based research design was employed, in which a unidimensional item pool consisting of 500 items was created. The test termination rule was defined as responding to a fixed number of 20 items per individual. For each condition in the simulation scenarios, data were generated based on the Three-Parameter Logistic Model (3PLM), which estimates the probability of a correct response by considering an individual's ability level and item parameters—discrimination (a), difficulty (b), and guessing (c) (Baker, 2001; Embretson & Reise, 2000).

The item parameters were generated as follows: discrimination parameters (a) followed a log-normal distribution (mean = 0, SD = 0.2), difficulty parameters (b) were drawn from a normal distribution (mean = 0, SD = 1), and guessing parameters (c) were uniformly distributed between 0 and 0.20. Each test session included 200 participants, and two sessions were conducted per day. For each method, 25 replications were performed, and the mean values from these replications were used in the analyses.

The study aimed to compare the Maximum Fisher Information (MFI) method developed by Veerkamp and Berger (1997) and the Maximum Likelihood Weighted Information (MLWI) method proposed by Eggen (1999). Additionally, Maximum Likelihood (ML) and Bayesian estimation methods were compared in terms of measurement accuracy and standard error. The test termination criterion was defined as a fixed-length test consisting of 20 items.

Findings

The findings indicate that, when employing the Maximum Likelihood Estimation (MLE) method, the Maximum Likelihood Weighted Information (MLWI) item selection strategy resulted in lower standard errors and Root Mean Square Deviation (RMSD) values compared to the Maximum Fisher Information (MFI) method. This outcome suggests that, within the MLE framework, the MLWI approach is more effective at producing precise and reliable ability estimates. In contrast, when using the Bayesian estimation method, the MFI item selection algorithm demonstrated superior performance relative to MLWI, yielding the lowest standard error and RMSD values among the evaluated combinations. Additionally, the highest correlation between true ability parameters (θ) and their corresponding estimates (θ̂) was observed in the Bayesian-MFI pairing, indicating that this combination achieved the best overall model fit and estimation accuracy. Furthermore, analysis of interim theta estimates revealed that ability estimations became increasingly stable as the adaptive test progressed. This trend reflects the fundamental adaptive characteristic of the testing process, wherein the growing interaction between test items and individual test-takers enhances the precision and reliability of ability estimations over the course of the assessment. Taken together, these results underscore the importance of selecting appropriate item selection algorithms in conjunction with the estimation methods to optimize the accuracy and stability of ability measurement in computerized adaptive testing environments.

Conclusion and Discussion

This study compared the performance of various item selection and ability estimation methods within Computerized Adaptive Testing (CAT) environments using simulation data. Analyses were conducted on a simulated dataset generated based on the 3PL model via the SimulCAT software. Maximum Likelihood Estimation (MLE) and Bayesian approaches were evaluated for ability estimation, while Maximum Fisher Information (MFI) and Maximum Likelihood Weighted Information (MLWI) methods were employed for item selection. The test termination criterion was set to a fixed length of 20 items, and simulations were repeated 25 times with a pool of 500 items administered to 1,000 examinees. Results indicated that the MLE-MLWI and Bayesian-MFI combinations yielded lower estimation errors and standard errors compared to other methods. Furthermore, the Bayesian-MFI combination demonstrated the highest correlation between true and estimated ability parameters, indicating superior model fit. These findings are consistent with prior research by Hambleton et al. (1991) and support earlier evidence from Embretson and Reise (2000) that Bayesian

methods offer advantages in CAT, particularly in situations with moderate sample sizes and initial uncertainty in ability estimation. Overall, the results underscore the critical role of careful selection of item selection and ability estimation algorithms in ensuring measurement validity and reliability in CAT applications.

Recommendations

The findings indicate that item selection and ability estimation methods in Computerized Adaptive Testing (CAT) significantly impact measurement accuracy and estimation reliability. Therefore, the harmonious and integrated design of item selection algorithms and ability estimation procedures is essential to effectively and accurately assess individual differences. Future research should prioritize validating simulation results with data from actual test-takers to ensure practical applicability. Moreover, test termination criteria should be expanded beyond fixed-length rules to include dynamic methods based on measurement precision and estimation reliability. Comparative studies examining item selection and ability estimation methods across various Item Response Theory (IRT) models and test structures are also warranted. Additionally, practical factors such as test duration and user experience should be considered when developing CAT strategies to enhance overall test effectiveness. Consequently, approaching item selection and ability estimation as an integrated system plays a critical role in improving the validity and reliability of CAT applications. These recommended directions are expected to contribute significantly to advancing both the theoretical framework and practical implementation of CAT systems.