Trakya Eğitim Dergisi

ISSN : 2630-6301
Yayımcı : Trakya Eğitim Dergisi

Vol. 16, No. 2, 766-793, 2026
DOI: 10.24315/tred.1665684

Research Article

Sabit ve Anında Bireyselleştirilmiş Çok Aşamalı Testlerin Karşılaştırmalı İncelenmesi: Ölçme Kesinliği ve Madde Güvenliği Açısından Çıkarımlar

Mahmut Sami Yiğiter[1*,**]
Nuri Doğan[2]

[1]Social Sciences University of Ankara, Ankara, Türkiye
mahmutsamiyigiter@gmail.com

[2]Hacettepe University, Educational Sciences Department, Ankara, Türkiye
nuridogan2004@gmail.com

* Sorumlu Yazar

** Bu çalışma, ikinci yazarın danışmanlığında birinci yazarın doktora tezinden türetilmiştir.

Geliş tarihi: 25.03.2025
Kabul tarihi 27.10.2025
Yayım tarihi: 30.04.2026

Özet: Son yıllarda, Bireyselleştirilmiş Bilgisayarlı Testler (BBT) ve Bireyselleştirilmiş Çok Aşamalı Testler (BÇAT) gibi uyarlanabilir test teknikleri geniş ölçekli değerlendirmelerde giderek daha fazla kullanılmaktadır. Bu çalışma, katılımcının yetenek düzeyine göre maddelerin modüller halinde anlık olarak gruplandırıldığı yeni bir yaklaşım olan Anında BÇAT'ı (A-BÇAT) ve Sabit BÇAT'ı (S-BÇAT), farklı simülasyon koşulları altında ölçme kesinliği ve madde güvenliği açısından karşılaştırmayı amaçlamaktadır. Simülasyonlar, TIMSS'te kullanılan 3PL modele dayalı madde parametre dağılımlarından yararlanılarak yürütülmüştür. A-BÇAT ile S-BÇAT'ın karşılaştırılması için toplam 72 farklı koşul analiz edilmiştir. Ölçme kesinliğine ilişkin bulgular, özellikle test uzunluğunun daha kısa olduğu durumlarda A-BÇAT'ın S-BÇAT'a kıyasla daha iyi performans gösterdiğini ve belirgin biçimde daha yüksek ölçme kesinliği sağladığını ortaya koymaktadır. Ayrıca yetenek dağılımları açısından incelendiğinde, özellikle normal olmayan dağılım koşullarında A-BÇAT'ın S-BÇAT'a göre daha üstün ölçme kesinliği sunduğu görülmüştür. Çalışmanın dikkat çekici sonuçlarından biri, A-BÇAT'ta son modül uzunluğu arttıkça ölçme kesinliğinin yükselmesi; buna karşılık S-BÇAT'ta ilk modül uzunluğu arttıkça ölçme kesinliğinin A-BÇAT'a daha fazla yaklaşmasıdır. Madde güvenliği açısından ise A-BÇAT tüm koşullarda daha fazla sayıda madde kullanmış ve S-BÇAT'a kıyasla daha düşük madde maruz kalma oranı göstermiştir. A-BÇAT'ın hem ölçme kesinliği hem de madde güvenliği bakımından ortaya koyduğu olumlu sonuçlar, geniş ölçekli değerlendirmeler ve ilgili alanyazın çerçevesinde tartışılmıştır.

Anahtar Kelime: Madde Güvenliği, Bireyselleştirilmiş Çok Aşamalı Testler, Bilgisayarlı Testler, Madde Güvenliği, Madde Teşhir Oranı.

GİRİŞ

Yüzyıllar boyunca, eğitimsel değerlendirmelerde test katılımcılarının bilgi, beceri ve yeteneklerini ölçmede Doğrusal Testler (DT) yaygın olarak kullanılan temel yöntem olmuştur. Ancak son elli yılda bilgisayar teknolojilerindeki gelişmelerle birlikte Bireyselleştirilmiş Bilgisayarlı Testler (BBT) önemli ölçüde gelişmiş ve giderek daha yaygın hale gelmiştir. BBT, özellikle yetenek düzeyini yüksek doğrulukla kestirebilmesi ve test uzunluğunu azaltabilmesi nedeniyle çok sayıda ulusal ve uluslararası değerlendirmede yaygın biçimde kullanılmaktadır (Khorramdel et al., 2020; Kirsch & Lennon, 2017; Demir & Gelbal, 2025). Buna karşılık, Bireyselleştirilmiş Çok Aşamalı Testler (BÇAT) son dönemde ortaya çıkan yenilikçi yapısıyla dikkat çekmiş ve geniş ölçekli değerlendirmelerde önemli bir yer edinmiştir (van der Linden, 2018).

BÇAT tasarımı, uygulama sürecinde testin güçlük düzeyinin katılımcının yetenek düzeyine göre ayarlanmasına olanak tanır. Bu yönüyle BÇAT, hem BBT hem de DT'nin özelliklerini bünyesinde birleştiren hibrit bir yapı olarak değerlendirilebilir. BBT ile BÇAT, özellikle maddelerin sunulmasının katılımcının önceki performansına dayanması bakımından benzerlik gösterse de temel fark sıralama ve yönlendirme sürecinde ortaya çıkar. BBT'de algoritma her bir maddeden sonra çalışarak katılımcının yeteneğini sürekli günceller. Buna karşılık BÇAT'ta yetenek kestirimi, modül olarak adlandırılan belirli bir madde grubunun tamamlanmasından sonra yapılır. Daha açık ifade etmek gerekirse, katılımcı öncelikle yönlendirme modülü olarak adlandırılan bir madde setiyle karşılaşır. Bu modüldeki performansa göre bireyin yetenek düzeyi kestirilir ve önceden belirlenmiş bir ölçüt puan (kesme puanı) ile karşılaştırılır. Eğer bireyin yeteneği bu ölçütün üzerindeyse daha zor bir modül (daha güç maddelerden oluşan set) uygulanır; altında kalıyorsa daha kolay bir modül (daha basit maddelerden oluşan set) sunulur. Esneklik ve karmaşıklık açısından değerlendirildiğinde, BÇAT, DT ile BBT arasında dengeli bir yapı sunmaktadır. DT'ye kıyasla BÇAT, yetenek ölçeği boyunca daha verimli ve daha doğru ölçümler sağlayarak değerlendirme kesinliğini artırmakta ve test uzunluğunu azaltmaktadır. Bu nedenle ölçme doğruluğu ve test verimliliği açısından BÇAT'ın DT'ye göre genel olarak daha etkili olduğu kabul edilmektedir (Lord, 1971; van der Linden, 2010).

Cite as (APA 7): Yiğiter, M. S. & Doğan, N. (2026). Sabit ve anında bireyselleştirilmiş çok aşamalı testlerin karşılaştırmalı incelenmesi: Ölçme kesinliği ve madde güvenliği açısından çıkarımlar. *Trakya Journal of Education 16*(2), 766-793. https://doi.org/10.24315/tred.1665864

BÇAT'ta kullanılan modüller test uygulanmadan önce tasarlanıp oluşturulduğu ve test katılımcısına bir bütün halinde sunulduğu için, kapsam geçerliği, test yapısının genel niteliği ve uygulama sürecinin yönetimi gibi konularda test geliştiricilere daha fazla kontrol olanağı sağlamaktadır (van der Linden, 2010). Ayrıca BBT'den farklı olarak BÇAT, katılımcıların her modül içinde yanıtlarını değiştirmesine, maddeleri atlamasına, önceki sorulara geri dönmesine ve yeni yanıtlar vermesine olanak tanımaktadır (Chang, 2015). Hambleton ve Xing (2006), geçme-kalma kararlarının verilmesi bağlamında BÇAT, BBT ve DT'yi karşılaştırmıştır. Bulgular, ölçme kesinliği açısından BBT'nin BÇAT'tan bir miktar daha iyi performans gösterdiğini ortaya koymuştur. Bununla birlikte araştırmacılar, daha geniş güçlük aralığına sahip modüller ve içerik açısından daha kapsamlı büyük madde havuzları kullanıldığında BÇAT'ın test geliştiriciler tarafından daha etkili biçimde kullanılabileceği sonucuna ulaşmıştır.

Son yıllarda BÇAT, sahip olduğu avantajlar nedeniyle çok sayıda geniş ölçekli değerlendirmede giderek daha fazla kullanılmaktadır. Örneğin CPA (Certified Public Accountants) sınavı 2004 yılından bu yana BÇAT kullanmaktadır (Breithaupt, Mills & Melican, 2006). 2011 yılında GRE (Graduate Record Examinations) BÇAT tabanlı formata geçmiştir. Aynı yıl OECD, PIAAC (Program for the International Assessment of Adult Competencies) kapsamında BÇAT tasarımını kullanarak uyarlanabilir uluslararası geniş ölçekli bir değerlendirme uygulamıştır (Kirsch & Lennon, 2017). PIAAC'ın ardından PISA'nın 2018 döngüsünde üç temel alanın biri olan okumada BÇAT kullanılmıştır (Khorramdel et al., 2020). PISA 2022'de ise BÇAT tasarımı yalnızca okumada değil matematik okuryazarlığında da kullanılmıştır (OECD, 2023).

BÇAT, test boyunca bireyin yetenek düzeyine uygun maddeler sunarak katılımcıların teste katılımını ve motivasyonunu artırma potansiyeline sahiptir. PISA bulgularına göre bilgisayar temelli değerlendirmelerde, kâğıt temelli değerlendirmelere kıyasla yanıtsız bırakma oranları daha düşüktür (OECD, 2024). Bu nedenle BÇAT'ın değerlendirmelerde yanıtsız bırakma ve rastgele yanıt verme davranışlarını azaltabileceği öngörülmektedir (Yamamoto et al., 2018). Çok sayıda çalışma uyarlanabilir testlerin motivasyon üzerindeki etkisini incelemiştir (Arvey et al., 1990; Bergstrom et al., 1992; Ortner et al., 2013; Pine et al., 1979). Ling ve arkadaşları (2017), uyarlanabilir testlerin sabit maddeli testlere göre daha yüksek katılım ve daha düşük kaygı sağladığını göstermiştir. Ayrıca Martin ve Lazendic (2018), BÇAT'ın bilgisayar temelli testlere göre daha kesin ölçümler sunduğunu ve motivasyon, katılım ile genel test deneyimi açısından olumlu sonuçlar ürettiğini bildirmiştir.

Son dönemde, gerçek yaşam durumlarını mümkün olduğunca yansıtan ortamlarda bilişsel becerileri ölçmeyi amaçlayan etkileşimli değerlendirmelere ilgi artmıştır (Bulut, 2021). Senaryo temelli madde grupları, öğrencilerin madde üzerinde değişiklik yaparak etkileşim kurduğu zengin ortamlar oluşturarak daha dinamik ve özgün bir değerlendirme deneyimi sunmaktadır. Bu nedenle gerçek yaşam durumlarını ve senaryo temelli görevleri etkili biçimde değerlendirebilmek için birden fazla maddeden oluşan daha bütünleşik sorulara ihtiyaç duyulmaktadır. Örneğin bir metne dayalı öneriler yazma ya da gerçek yaşam bağlamında birbirini izleyen bir dizi problemi çözme gibi karmaşık görevler, tek bir sorudan ziyade ilişkili birden fazla maddenin sunulmasını gerektirir. Bu tür durumlarda katılımcı performansının birbiriyle bağlantılı maddeler grubu üzerinden değerlendirilmesi gerekir. BBT uygulamalarında yetenek kestirimi madde bazında yapılmaktadır. Ancak daha önce belirtildiği gibi ortak bir senaryo ya da tema ile birbirine bağlı madde setlerinin tek tek ayrılması, ölçülen yapının doğasını bozabilmektedir. Bu nedenle BBT bu tür senaryolar için uygun değildir. Buna karşılık PISA'da maddelerin yaklaşık %30'u insan puanlayıcılar tarafından puanlanan açık uçlu maddelerden oluşmaktadır. Bu maddelerin puanlaması anında gerçekleşmediği için BBT'nin uygulanması uygun olmayabilir. Modül düzeyinde uyarlanabilir bir test olan BÇAT tasarımı bu bağlamda daha uygun görünmektedir. Ölçülen yapının kuramsal çerçevesiyle uyumlu, iyi tasarlanmış madde gruplarının (modüllerin), olgusal bilgiye dayalı çok sayıda bağımsız maddeden daha doğru değerlendirmeler sağlayacağı düşünülmektedir (Yamamoto et al., 2018; Yigiter & Dogan, 2023).

BBT'nin temel sınırlılıklarından biri, test katılımcısının yetenek düzeyini olduğundan düşük ya da yüksek kestirebilmesidir (Chang & Ying, 2008). Bu durum, yetenek kestiriminin her bir madde yanıtlandıktan sonra Maksimum Fisher Bilgisi (MFI) yöntemi kullanılarak güncellenmesinden kaynaklanmaktadır. Örneğin yüksek yetenek düzeyine sahip bir katılımcı, test kaygısı, motivasyon eksikliği ya da basit hatalar nedeniyle başlangıç maddelerini yanlış yanıtlarsa, sonraki yanıtlar üzerinden gerçek yeteneğinin doğru biçimde belirlenmesi güçleşmektedir. Benzer şekilde düşük yetenek düzeyine sahip bir katılımcı, şans eseri ya da ön bilgi nedeniyle ilk maddeleri doğru yanıtlarsa sistem bireyin yeteneğini olduğundan yüksek kestirebilmektedir. Buna karşılık BÇAT, yetenek düzeyini her aşama tamamlandıktan sonra kestirdiği için BBT'nin bu

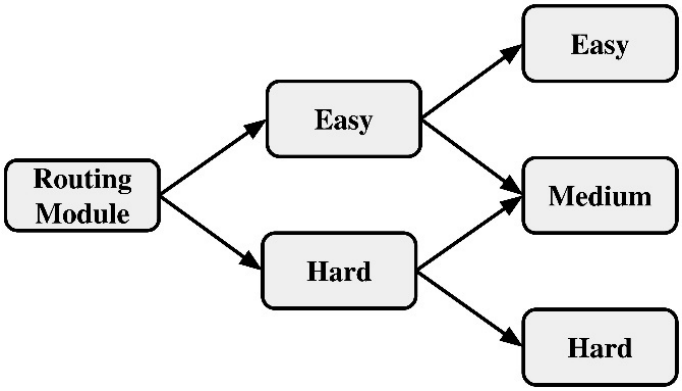
sınırlılıklarını azaltmaya yardımcı olmaktadır. Bu üstünlük, bazı sınav kuruluşlarının geniş ölçekli değerlendirmelerde BBT ile ilgili sorunları belirlemesiyle daha görünür hale gelmiştir. Örneğin 2000 yılında Educational Testing Service (ETS), bilgisayarlı Graduate Record Examinations'ın (GRE-BBT) binlerce katılımcı için puanları hatalı biçimde kestirdiğini belirlemiş ve etkilenen bireylere sınavı ücretsiz yeniden alma hakkı tanımıştır (Carlson, 2000). Benzer bir sorun 2002 yılında Graduate Management Admission Test'te (GMAT) ortaya çıkmış ve yaklaşık bin aday yanlış puan almıştır (Chang, 2004). Bu olayların ardından BÇAT, BBT'nin eksik yönlerini giderebilecek bir çözüm olarak öne çıkmıştır. Sonuç olarak birçok sınav kurumu BBT'den BÇAT'a geçiş yapmıştır (Hendrickson, 2007).

Sabit Bireyselleştirilmiş Çok Aşamalı Testler (S-BÇAT)

S-BÇAT, önceden oluşturulmuş madde gruplarının (modüllerin) algoritma tarafından seçilerek test katılımcılarına aşamalar halinde uygulandığı algoritma temelli bir test yöntemidir. Alanyazında bu yaklaşım farklı adlarla anılmakta olup, en yaygın kullanım "fixed" terimidir; bununla birlikte "standard" ve "conventional" ifadeleri de aynı yöntemi tanımlamak için kullanılmaktadır. Açıklık sağlamak amacıyla bu çalışmada "sabit" terimi tercih edilmiştir. Şekil 1, S-BÇAT tasarımını 1-2-3 formatında göstermektedir. S-BÇAT'ta önceden oluşturulmuş her bir madde grubu modül olarak adlandırılmakta ve test tasarımının temel birimini oluşturmaktadır. İlk aşamada genellikle orta güçlük düzeyinde olan ve çoğunlukla yönlendirme modülü olarak adlandırılan bir modül sunulmaktadır. Test, katılımcının yetenek düzeyine göre birbirini izleyen modüllere yönlendirilerek ilerlemekte ve her modül katılımcının performansına uyum sağlayacak şekilde uygulanmaktadır.

Şekil 1

1-2-3 SBÇAT Tasarımı



Şekil 1, ilk aşamada bir modül, ikinci aşamada iki modül ve üçüncü aşamada üç modülden oluşan 1-2-3 formatındaki bir S-BÇAT tasarımına örnek sunmaktadır. Bu tasarımdaki modüllerin güçlük düzeyleri genellikle kolay, orta ve zor olarak sınıflandırılmaktadır. S-BÇAT'ın ilk aşamasında tüm katılımcılar, genellikle orta güçlükte maddelerden oluşturulan ve yönlendirme modülü olarak adlandırılan ilk modülle başlamaktadır. Katılımcılar bu modüldeki performanslarına göre ikinci aşamada kolay ya da zor modüle yönlendirilmektedir. İkinci aşamada ise verilen yanıtlara bağlı olarak katılımcılar üçüncü aşamada yer alan kolay, orta ya da zor modüllerden birine atanır. Son olarak üçüncü aşamanın tamamlanmasının ardından test sona ermekte ve katılımcının nihai yetenek düzeyi kestirilmektedir. Şekil 1'de gösterilen modül ve aşama yapısı panel olarak adlandırılmakta olup, test uygulamalarında birden fazla panel oluşturulabilmektedir.

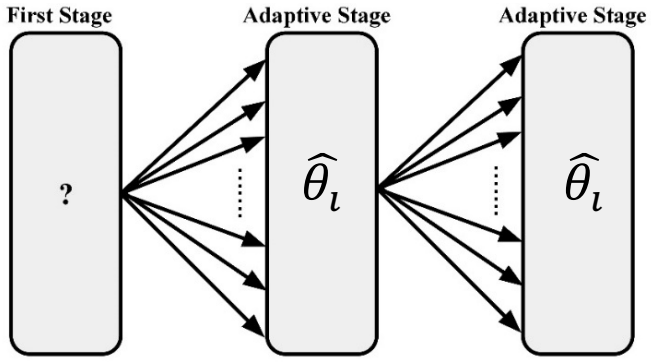
Anında Bireyselleştirilmiş Çok Aşamalı Testler (A-BÇAT)

S-BÇAT'a benzer şekilde, Anında Bireyselleştirilmiş Çok Aşamalı Testler (A-BÇAT) de aşamalar halinde uygulanmaktadır. Ancak S-BÇAT'tan farklı olarak A-BÇAT, modülleri test süreci sırasında bireyin yetenek düzeyine göre anında oluşturmaktadır (Tay, 2015). A-BÇAT'ta ilk aşama genellikle orta güçlük düzeyindeki maddelerden oluşan tam uzunlukta bir testin oluşturulmasını içermektedir. Bu testin ilk modül uzunluğundaki bölümü katılımcıya uygulanmakta ve performansına dayalı olarak ara yetenek kestirimi hesaplanmaktadır. Ardından test, katılımcının ara yetenek kestirimine göre yeniden oluşturulmakta ve ikinci modül uzunluğundaki bölüm uygulanmaktadır. Bu durum, her katılımcının ikinci ve üçüncü aşamalarda, her aşama sonrasındaki ara yetenek parametresine göre uyarlanmış farklı madde setleri alması anlamına

gelmektedir. Özünde A-BÇAT, her katılımcı için kişiselleştirilmiş bir test deneyimi sunmaktadır. Bu uyarlanabilir süreç test tamamlanıncaya kadar devam etmektedir. Şekil 2, gölge testlerin genel çerçevesini ve A-BÇAT'ın uygulanışını göstermektedir.

Şekil 2

A-BÇAT Genel Çerçevesi



Not. $\widehat{\theta}_l$*: kestirilen yetenek düzeyi.*

Şekil 2, üç aşamalı A-BÇAT uygulamasını göstermektedir. İlk aşamada test katılımcısına orta güçlük düzeyinde bir modül sunulmaktadır. İkinci ve sonraki aşamalarda ise katılımcının ara yetenek düzeyine göre oluşturulan modül kombinasyonları uygulanmaktadır. Her ne kadar Şekil 2'de üç aşamalı bir A-BÇAT tasarımı gösterilmiş olsa da, A-BÇAT sisteminde modül sayısı ve her modülün uzunluğu gereksinimlere göre ayarlanabilmektedir. Genel olarak A-BÇAT oldukça esnek bir yapıya sahiptir; her modüldeki madde sayısında ve aşama sayısında farklı düzenlemelere olanak tanımaktadır (Zheng & Chang, 2015).

Madde Güvenliği

Bireyselleştirilmiş Bilgisayarlı Testler (BBT) ve Bireyselleştirilmiş Çok Aşamalı Testler (BÇAT) kapsamında madde güvenliği, bir madde havuzundaki maddelerin aşırı kullanımına veya açığa çıkmasına karşı ne ölçüde korunduğunu ifade etmektedir. Madde güvenliğinin korunması, bir değerlendirmenin adaletini, geçerliğini ve gizliliğini sürdürmenin temel unsurlarından biridir. Belirli maddeler çok sık uygulandığında, bu maddelerin gelecekteki katılımcılar tarafından önceden bilinme riski artmakta ve bu durum sınav programının güvenirliğini zedeleyebilmektedir.

Madde güvenliği çoğunlukla madde maruz kalma oranı ile değerlendirilmektedir. Bu oran, belirli bir maddenin farklı test uygulamaları boyunca katılımcıların yüzde kaçına sunulduğunu göstermektedir. Yüksek maruz kalma oranları, sınırlı sayıdaki bazı maddelerin tekrar tekrar seçildiğini, diğerlerinin ise kullanılmadan kaldığını göstermektedir. Bu tür dengesiz maruz kalma örüntüleri çeşitli istenmeyen sonuçlara yol açabilmektedir: yaygın biçimde kullanılan maddeler katılımcılar arasında dolaşıma girerek bunlara önceden aşina olan bireyler için haksız avantaj oluşturabilir; yetenek kestirimleri yanlı hale gelerek testin psikometrik niteliği düşebilir; ayrıca madde havuzunun operasyonel kullanım ömrü kısalabilir ve bu da sürekli yeni madde geliştirilmesini gerektiren maliyetli ve emek yoğun bir sürece dönüşebilir.

Buna karşılık güvenli bir test sistemi, maddelerin etkili fakat aşırı olmayan biçimde kullanıldığı dengeli bir maruz kalma düzeyi sağlamaktadır. Bu denge hem test kesinliğini hem de madde havuzunun uzun ömürlülüğünü artırmaktadır. Bu çalışmada madde güvenliği, uyarlanabilir test araştırmalarındaki yerleşik yaklaşımlar doğrultusunda ortalama madde maruz kalma oranı kullanılarak değerlendirilmiştir (van der Linden & Veldkamp, 2004; Zheng & Chang, 2015). Anında Bireyselleştirilmiş Çok Aşamalı Testler (A-BÇAT) çerçevesi, güncellenen yetenek kestirimlerine dayalı olarak modülleri dinamik biçimde oluşturduğu için madde kullanımında daha fazla çeşitlilik sağlamakta ve aşırı maruz kalma olasılığını azaltmaktadır. Sonuç olarak A-BÇAT, katılımcılar arasında belirli maddelerin tekrar kullanımına yol açabilen önceden tanımlı modüllere dayalı S-BÇAT tasarımına kıyasla daha güçlü bir madde koruması sunmaktadır.

Literatür Taraması

Alanyazında S-BÇAT ile A-BÇAT'ı karşılaştıran çeşitli çalışmalar bulunmaktadır. Choi ve arkadaşları (2016), test birleştirme sürecinde freeze-fresh mekanizmasını kullanarak BBT, A-BÇAT ve her iki yöntemin birleşimi olan Hibrit-BBT'yi karşılaştırmıştır. Sonuçlar, üç yaklaşımda da oldukça benzer bulunmuştur. Araştırmacılar, freeze-fresh mekanizmasının özellikle ortak kök maddelerin bulunduğu ve test kısıtlarının sağlanmasının gerektiği durumlarda oldukça iyi çalıştığını belirlemiştir. Ayrıca testin ölçme doğruluğunda anlamlı bir azalma görülmemiştir. Zheng ve Chang (2015), modüllerin önceden oluşturulduğu S-BÇAT'tan farklı olarak, her aşamada ara yetenek kestirimlerine dayalı biçimde modülleri anında birleştiren yeni bir BÇAT tasarımı sunmuştur. Bulgular, A-BÇAT ile BBT'nin benzer ölçme doğruluğu sağladığını ve her iki yöntemin de S-BÇAT'a göre daha iyi ölçme doğruluğu ve test güvenliği sunduğunu göstermiştir.

Tay (2015), S-BÇAT ile A-BÇAT'ı 12, 18, 24 ve 30 maddelik test uzunluklarında sınıflama doğruluğu ve tutarlılığı açısından karşılaştırmıştır. Sonuçlar, tüm koşullarda A-BÇAT'ın S-BÇAT'a göre daha yüksek sınıflama doğruluğu ve tutarlılığı sağladığını ortaya koymuştur. van der Linden ve Diao (2016), gerçek bir veri seti üzerinden simülasyonlarla beş farklı test yaklaşımını karşılaştırmıştır. Çalışma, LT'nin en düşük verimliliğe sahip olduğunu, onu S-BÇAT'ın izlediğini göstermiştir. Diğer yaklaşımlar ise birbirine benzer verimlilik düzeyleri sergilemiştir.

Han ve Guo (2016), her aşamada test katılımcısının tahmini yetenek düzeyine (θ) dayalı olarak yeni modülleri anında oluşturan bir A-BÇAT tasarımı önermiştir. Birleştirilen modül, test geliştirici tarafından belirlenen hedef test bilgi fonksiyonuna (TIF) ulaşıncaya kadar yinelemeli biçimde yeniden oluşturulmuştur. Çalışmada BBT, S-BÇAT ve A-BÇAT tasarımları karşılaştırılmıştır. Sonuçlar, 1-3-3 BÇAT tasarımının düşük yineleme düzeyinde yeni geliştirilen A-BÇAT tasarımıyla benzer ölçme doğruluğu sağladığını göstermiştir. Ancak yineleme sayısı 100'e çıkarıldığında yeni A-BÇAT tasarımı, ölçme doğruluğu açısından S-BÇAT'tan daha iyi performans göstermiştir. Buna karşın BBT, her iki yöntemden de daha yüksek doğruluk üretmiştir.

Son olarak van der Linden (2021), gerçek bir veri setinden türetilmiş 300 maddelik bir havuz kullanarak anında oluşturulan DT, S-BÇAT, A-BÇAT ve BBT'yi karşılaştırmıştır. Simülasyon sonuçları, A-BÇAT ile BBT'nin benzer ve oldukça yüksek ölçme kesinliği sağladığını göstermiştir. Buna karşılık S-BÇAT, anında oluşturulan DT'den daha doğru sonuçlar üretmekle birlikte hem A-BÇAT hem de BBT'ye kıyasla anlamlı derecede daha düşük ölçme doğruluğu sunmuştur.

Araştırmanın Amacı ve Önemi

Alanyazında S-BÇAT ile A-BÇAT'ı karşılaştıran çeşitli çalışmalar bulunmaktadır. Choi ve arkadaşları (2016), test birleştirme sürecinde freeze-fresh mekanizmasını kullanarak BBT, A-BÇAT ve her iki yöntemi birleştiren Hibrit-BBT'yi karşılaştırmıştır. Sonuçlar, her üç yaklaşımın da benzer çıktılar ürettiğini göstermiştir. Çalışma, özellikle ortak kök maddelerin kullanıldığı ve test kısıtlarına uyulmasının gerektiği durumlarda, freeze-fresh mekanizmasının ölçme doğruluğunda anlamlı bir düşüşe yol açmadan etkili biçimde çalıştığı sonucuna ulaşmıştır. Zheng ve Chang (2015) ise modüllerin önceden oluşturulduğu S-BÇAT'tan farklı olarak, her aşamada elde edilen ara yetenek kestirimlerine göre modülleri anında birleştiren yenilikçi bir BÇAT tasarımı sunmuştur. Araştırma bulguları, A-BÇAT ile BBT'nin ölçme doğruluğunun karşılaştırılabilir düzeyde olduğunu ve her iki yöntemin de S-BÇAT'a göre daha yüksek ölçme doğruluğu ile daha güçlü test güvenliği sağladığını ortaya koymuştur.

Araştırma Soruları

Aşağıda, iki araştırma sorusu ve iki araştırma problemine dayalı olarak geliştirilen altı alt araştırma problemi sunulmuştur.

S1. Farklı simülasyon koşulları altında S-BÇAT ve A-BÇAT yaklaşımlarının ölçme kesinliği ne düzeyde değişmektedir?

S1.1. S-BÇAT ve A-BÇAT yaklaşımlarının RMSE, MAB ve BIAS değerleri test uzunluğuna (20-30-40) göre nasıl değişmektedir?

S1.2. S-BÇAT ve A-BÇAT yaklaşımlarının RMSE, MAB ve BIAS değerleri yetenek dağılımına (normal dağılım, sağa çarpık, sola çarpık, uniform) göre nasıl değişmektedir?

S1.3. S-BÇAT ve A-BÇAT yaklaşımlarının RMSE, MAB ve BIAS değerleri modül/test uzunluğu oranına (B-K-K, O-O-O, K-K-B) göre nasıl değişmektedir?

S2. Farklı simülasyon koşulları altında S-BÇAT ve A-BÇAT yaklaşımlarının madde güvenliği nasıl değişmektedir?

S2.1. S-BÇAT ve A-BÇAT yaklaşımlarının madde güvenliği, farklı test uzunluklarında (20-30-40) madde kullanım sıklığı ve kullanılan madde sayısına göre nasıl değişmektedir?

S2.2. S-BÇAT ve A-BÇAT yaklaşımlarının madde güvenliği, farklı yetenek dağılımlarında (normal dağılım, sağa çarpık, sola çarpık, uniform) madde kullanım sıklığı ve kullanılan madde sayısına göre nasıl değişmektedir?

S2.3. S-BÇAT ve A-BÇAT yaklaşımlarının madde güvenliği, farklı modül/test uzunluğu oranlarında (B-K-K, O-O-O, K-K-B) madde kullanım sıklığı ve kullanılan madde sayısına göre nasıl değişmektedir?

YÖNTEM

Araştırmanın Türü

Bu çalışmanın amacı, farklı simülasyon koşulları altında çeşitli BÇAT yaklaşımlarının etkililiğini incelemektir. Araştırmada kullanılan veriler simülasyon yoluyla üretilmiş ve farklı senaryolar arasında karşılaştırmalar yapılmıştır. Simülasyon çalışmaları, belirli olasılık dağılımlarından rastgele örnekleme yoluyla veri üreten ve ardından elde edilen verileri analiz eden bilgisayar tabanlı deneylerdir. Bu çalışmalar, istatistiksel yöntemlerin kontrollü koşullar altında performansını değerlendirmede önemli bir işleve sahiptir. Psikometri alanında simülasyon çalışmaları, hem yeni hem de alternatif yöntemlerin değerlendirilmesinde kritik bir rol oynamaktadır (Feinberg & Rubright, 2016; Morris, White & Crowther, 2019; Saatçioğlu & Atar, 2022). Bu araştırma, verilerin ilgili olasılık dağılımları ve simülasyon koşullarına göre üretildiği bir Monte Carlo simülasyon çalışmasıdır. Deneysel çalışmalarla benzerlik gösteren Monte Carlo çalışmaları, bilgisayar tarafından üretilen verileri kullanmaktadır (Harwell et al., 1996). Araştırmada ele alınan tüm değişkenlerin gerçek verilerle incelenmesinin karmaşık olması nedeniyle simüle edilmiş veriler kullanılmıştır. Çalışma, farklı BÇAT yöntemlerini karşılaştırmayı amaçladığından, hangi yöntemin en uygun sonuçları ürettiğini belirlemeye yönelik betimsel bir araştırma olarak sınıflandırılabilir (Fraenkel, Wallen & Hyun, 2012).

Veri Üretimi

Veriler, açık kaynaklı ve ücretsiz bir istatistiksel programlama dili olan R kullanılarak üretilmiştir. Veri üretim sürecinde ilk olarak 400 maddeden oluşan madde havuzu parametreleri ile 1000 bireyden oluşan yetenek parametreleri oluşturulmuştur (Şahin & Weiss, 2015). Test birleştirme sürecinde "Rmst" (Luo & Kim, 2018) ve "TestDesign" (Choi, Lim & van der Linden, 2021) paketlerinden yararlanılmıştır. Ardından araştırmacılar tarafından yazılan kodlar kullanılarak, birleştirilen testler üzerinden "mstR" (Magis, Yan & Von Davier, 2017) paketi ile S-BÇAT analizleri, aynı madde havuzu ve yetenek parametreleri kullanılarak "TestDesign" (Choi, Lim & van der Linden, 2021) paketi ile A-BÇAT analizleri gerçekleştirilmiştir.

Madde Havuzu Üretimi

Çalışma kapsamında, 3 parametreli lojistik modele (3PL) dayalı olarak 400 maddeden oluşan bir madde havuzu üretilmiştir (Yiğiter & Boduroğlu, 2024). Madde havuzunun oluşturulmasında, TIMSS 8. sınıf matematik uygulamalarının 2003, 2007, 2011, 2015 ve 2019 döngülerinde kullanılan ve parametreleri 3PL modeline göre kestirilen maddelerin a, b ve c parametre dağılımları incelenmiştir. Bu maddelerin parametre dağılımlarına ilişkin minimum ve maksimum değerler ile çarpıklık ve basıklık katsayıları dikkate alınarak; a parametreleri log-normal dağılımdan $a \sim \ln N(0.2,0.3)$, b parametreleri normal dağılımdan $b \sim N(0,0.7)$ve

c parametreleri beta dağılımından $c \sim Beta(5,16)$üretilmiştir. Madde havuzunda yer alan 400 maddenin parametrelerine ilişkin betimsel istatistikler Tablo 1'de sunulmuştur.

Tablo 1

Madde Parametrelerinin Betimsel İstatistikleri

Parameter	K	Min	Max	Mean	Sd
A	400	0.566	2.287	1.243	0.351
B	400	-1.830	1.764	0.000	0.727
C	400	0.044	0.501	0.236	0.088

Buna ek olarak, madde havuzunun dört farklı içerik alanından oluştuğu varsayılmış ve tüm maddeler rastgele biçimde İçerik 1 (%30 - 120 madde), İçerik 2 (%30 - 120 madde), İçerik 3 (%20 - 80 madde) ve İçerik 4 (%20 - 80 madde) alanlarına atanmıştır.

Yetenek Dağılımı Üretimi

Çalışmada normal, sağa çarpık, sola çarpık ve uniform olmak üzere dört farklı yetenek dağılımı kullanılmıştır. Tüm yetenek dağılımlarında katılımcı sayısı N = 1000 olarak belirlenmiştir. Sağa çarpık ve sola çarpık yetenek dağılımları, Fleishman (1978) tarafından önerilen kuvvet yöntemi ile normal dağılımdan elde edilmiştir. Fleishman'ın kuvvet yöntemi denklemi aşağıda sunulmuştur:

$$Y = a + bX + cX^2 + dX^3 \quad (1)$$

Burada a, b, c ve d, üretilen yetenek dağılımlarının şeklini tanımlayan Fleishman dönüşüm katsayılarını ifade etmektedir (Fleishman, 1978). Bu katsayılar, çalışmada kullanılan hedef normal olmayan dağılımları elde etmek amacıyla simüle edilen veriye yön vermekte ve buna göre hesaplanmaktadır. X, kullanılan normal dağılımdan elde edilen parametreleri ifade etmektedir. Normal, sağa çarpık ve sola çarpık yetenek dağılımlarını üretmek için kullanılan X ile a, b, c, d katsayılarına ilişkin değerler Tablo 2'de sunulmuştur. Uniform yetenek dağılımı ise uniform dağılımdan $U \sim U(-3, +3)$elde edilmiştir.

Veri üretiminde R programlama dili kullanılmıştır. R, açık kaynaklı ve ücretsiz bir istatistiksel programlama dilidir. Veri üretim sürecinde ilk olarak 400 maddeden oluşan madde havuzu parametreleri ile 1000 bireyden oluşan yetenek parametreleri üretilmiştir. Test birleştirme sürecinde "Rmst" (Luo & Kim, 2018) ve "TestDesign" (Choi, Lim & van der Linden, 2021) paketlerinden yararlanılmıştır. Ardından araştırmacılar tarafından yazılan kodlar kullanılarak, birleştirilen testler üzerinden "mstR" (Magis, Yan & Von Davier, 2017) paketi ile S-BÇAT analizleri, aynı madde havuzu ve yetenek parametreleri kullanılarak "TestDesign" (Choi, Lim & van der Linden, 2021) paketi ile A-BÇAT analizleri gerçekleştirilmiştir.

Tablo 2

Yetenek Dağılımı Üretimi

Distribution	N	X	Skewness	Kurtosis	a	b	c	d
Normal	1000	N(0, 1)	0.00	0.00	0.00	1.00	0.00	0.00
Right Skewed	1000	N(0, 1)	1.50	4.00	-0.21	0.85	0.21	0.04
Left Skewed	1000	N(0, 1)	-1.50	4.00	0.21	0.85	-0.21	0.04
Uniform	1000	U(-3, +3)	-	-	-	-	-	-

Araştırma Tasarımı

Bu çalışmada, farklı simülasyon koşulları altında farklı BÇAT yaklaşımları (S-BÇAT ve A-BÇAT) karşılaştırılmıştır. Simülasyonda kullanılan bağımsız ve bağımlı değişkenler Tablo 3'te özetlenmiştir.

Tablo 3

Bağımlı ve Bağımsız Değişkenler

Değişken Türü	Değişken Adı	Açıklama
Bağımsız Değişken	BÇAT Türü	S-BÇAT ve A-BÇAT
Bağımsız Değişken	Test Uzunluğu	20, 30, 40 soru
Bağımsız Değişken	Yetenek Dağılımı	Normal, Sağa Çarpık, Sola Çarpık, Uniform
Bağımsız Değişken	Modül/Test Uzunluğu Oranı	B-K-K, O-O-O, K-K-B
Bağımlı Değişken	Ölçme Kesinliği	RMSE, MAB ve BIAS değerleri
Bağımlı Değişken	Madde Güvenliği	Kullanılan madde sayısı ve madde maruz kalma oranı

Simülasyonda manipüle edilecek koşullar Tablo 4'te sunulmuştur.

Tablo 4

Manipüle Edilen Koşullar

Manipüle Edilen Değişken	Düzey	Düzey Sayısı
BÇAT Yaklaşım Türü	S-BÇAT, A-BÇAT	2
Test Uzunluğu	20, 30, 40	3
Yetenek Dağılımı	Normal, Sağa Çarpık, Sola Çarpık ve Uniform	4
Modül/Test Uzunluğu Oranı	B-K-K [1/2-1/4-1/4], O-O-O [1/3-1/3-1/3], K-K-B [1/4-1/4-1/2]	3

Not. K: Küçük, O: Orta, B: Büyük.

Tablo 4'te görüldüğü üzere, simülasyonlar iki farklı BÇAT türü (S-BÇAT, A-BÇAT), üç farklı test uzunluğu (20, 30, 40), dört farklı yetenek dağılımı (normal, sağa çarpık, sola çarpık ve uniform) ve üç farklı modül/test uzunluğu oranı dağılımı (B-K-K, O-O-O, K-K-B) değiştirilerek yürütülmüştür. Tüm koşullar birbiriyle çaprazlanmıştır. Bu nedenle ilk simülasyon çalışmasında $2 \times 3 \times 4 \times 3 = 72$ koşul incelenmiş ve her koşul için 100 tekrar gerçekleştirilmiştir. Eğitimde ölçme ve psikometri alanındaki önceki simülasyon çalışmalarıyla tutarlı olarak, bu araştırmada her koşul için 100 tekrar kullanılmıştır. Önceki çalışmalar, yaklaşık 100 tekrarın BIAS, RMSE ve madde maruz kalma indeksleri için yeterince kararlı kestirimler sağladığını, aynı zamanda hesaplama verimliliğini koruduğunu göstermektedir (Bulut & Sünbül, 2017; Gür & Gülleroğlu, 2020; Xu et al., 2023). Ayrıca 100 tekrar kullanımı, özellikle karmaşık MTK temelli uyarlanabilir test bağlamlarında örnekleme yanlılığını azaltmak ve ampirik kararlılığı sağlamak için önerilmektedir (Han, 2007; Harwell et al., 1996).

MST Desenlerinin Oluşturulması

Çalışmada S-BÇAT tasarımı olarak 1-2-3 formatı tercih edilmiştir. 1-2-3 S-BÇAT ve A-BÇAT tasarımlarının oluşturulmasında sırasıyla "Rmst" ve "TestDesign" paketleri kullanılmıştır. Bu araştırmada 1-2-3 BÇAT tasarımının tercih edilmesinin temel nedeni, ikinci aşamadaki modüllerin modül bilgi

fonksiyonlarının maksimuma ulaştığı noktaların, diğer aşamalardaki modüllerin maksimum bilgi noktalarıyla çakışmamasıdır. Başka bir ifadeyle, 1-3-3 BÇAT tasarımında ikinci ve üçüncü aşamadaki modüllerin bilgi fonksiyonunu maksimize eden noktaların örtüşmesi, modüllerin kümülatif bilgi fonksiyonu üzerinde azaltıcı bir etki yaratmaktadır. Cetin-Berber, Sarı ve Huggins-Manley (2018), 1-2-3 ve 1-3-3 BÇAT tasarımlarının oldukça benzer ölçme kesinliği sonuçları verdiğini bildirmiştir. Benzer biçimde Yiğiter ve Doğan (2023), 1-3, 1-2-3 ve 1-3-3 BÇAT tasarımlarını inceledikleri çalışmalarında, 1-2-3 ile 1-3-3 tasarımlarının çok benzer ölçme kesinliğine sahip olduğunu, hatta 1-2-3 tasarımının küçük bir farkla daha iyi ölçme kesinliği sunduğunu belirtmiştir. Bu nedenle madde havuzunu daha etkili kullanabilmek amacıyla 1-2-3 BÇAT tasarımı tercih edilmiştir.

Modüllerin oluşturulmasında modül/test uzunluğu oranı dikkate alınmıştır. Örneğin, 40 maddelik bir test uzunluğu ve K-K-B (1/4-1/4-1/2) modül/test oranı için ATA süreci 10-10-20 modül uzunluklarıyla yürütülmüştür. A-BÇAT'ta, bu oranları koruyabilmek amacıyla yeni modül, ilgili madde sayısı tamamlandıktan hemen sonra oluşturularak test katılımcısına sunulmaktadır. Örneğin 10-10-20 modül uzunluğuna sahip bir A-BÇAT tasarımında, yeni modül 1-11-21 madde pozisyonlarındaki kısıtlar dikkate alınarak freeze-fresh mekanizmasıyla (Choi et al., 2021) birleştirilmiş ve katılımcıya sunulmuştur. Freeze-fresh mekanizması (Choi & van der Linden, 2018), A-BÇAT'ta kullanılan dinamik bir modül birleştirme stratejisidir. Bu yaklaşımda her aşamadan sonra daha önce uygulanmış maddeler ya da modüller dondurulmakta (sabit tutulmakta), henüz uygulanmamış kalan maddeler ise güncellenen yetenek kestirimine göre yeniden seçilmektedir. Bu süreç, test boyunca uyarlanabilirliği korurken aynı zamanda içerik dengesi veya madde maruz kalma sınırları gibi tüm test kısıtlarının sağlanmasına olanak tanımaktadır. 1-2-3 BÇAT ve A-BÇAT tasarımlarında modüllerde yer alacak madde sayıları ile madde maruz kalma kontrol yöntemi Tablo 5'te sunulmuştur.

Tablo 5

Test Uzunluğuna Göre Modül Madde Sayıları ve Madde Maruz Kalma Oranı Kontrolü

Tasarım		Test Uzunluğu			Madde Kullanım Sıklığı Oranı	Panel Sayısı
		20	30	40		
		Modül Test Uzunluğu Dağılımı				
1-2-3 BÇAT	B-K-K	10-5-5	15-8-7	20-10-10	0.33	3
	O-O-O	3-4-3	6-7-6	13-14-13		
	K-K-B	5-5-10	7-8-15	10-10-20		
A-BÇAT	B-K-K	10-5-5	15-8-7	20-10-10	0.33	*
	O-O-O	3-4-3	6-7-6	13-14-13		
	K-K-B	5-5-10	7-8-15	10-10-20		

Not. A-BÇAT yönteminde madde maruz kalma oranı, Uygunsuzluk (Ineligibility) yöntemi kullanılarak 0.33 düzeyinde sabitlenerek kontrol edilmiştir (van der Linden & Veldkamp, 2004).

S-BÇAT yaklaşımında, madde maruz kalma oranını kontrol etmek amacıyla panel sayısı kullanılmıştır. Her bir S-BÇAT tasarımında, 0.33'lük madde maruz kalma oranı için 3 panel oluşturulmuştur. A-BÇAT tasarımında ise modüller anlık olarak oluşturulduğundan, madde maruz kalma oranını 0.33 düzeyinde sabitleyerek kontrol etmek için Uygunsuzluk (Ineligibility) yöntemi kullanılmıştır (van der Linden & Veldkamp, 2004). Uygunsuzluk yöntemi, belirli maddelerin aşırı kullanımını önlemek amacıyla geliştirilmiş bir madde maruz kalma kontrol tekniğidir. Bir madde uygulandıktan sonra, maruz kalma oranı belirlenen eşik değerin altına düşene kadar geçici olarak seçime uygun olmayan duruma getirilmektedir. Bu döngüsel mekanizma, test güvenliğinin korunmasına ve katılımcılar arasında madde kullanımının adil biçimde dağıtılmasına yardımcı olmaktadır (van der Linden & Veldkamp, 2004).

Test birleştirme ile panel ve modüllerin oluşturulmasında modül bilgisinin maksimuma ulaştığı noktalar Tablo 5'te sunulmuştur.

Tablo 5

BÇAT Panel ve Modüllerinin Oluşturulması

BÇAT Yaklaşımı	Aşama 1	Aşama 2	Aşama 3
1-2-3 S-BÇAT	$\vartheta = 0$	ϑ = (-0.5, 0.5)	ϑ = (-1, 0, 1)
A-BÇAT	$\vartheta = 0$	$\vartheta = \vartheta^*$	$\vartheta = \vartheta^*$

Not. ϑ*: geçici yetenek düzeyi.

Her iki BÇAT tasarımında da testler başlangıç yetenek düzeyi $\vartheta = 0$olacak şekilde birleştirilmiştir. S-BÇAT yaklaşımında, 1-2-3 tasarımına uygun olarak ikinci ve üçüncü aşama modülleri, Tablo 3'te verilen yetenek düzeylerinde maksimum bilgi değerine ulaşacak biçimde oluşturulmuştur. 1-2-3 S-BÇAT tasarımının test birleştirme sürecinde hibrit yöntemden yararlanılmıştır. Hibrit yöntem, modüllerin aşağıdan yukarı (bottom-up) yaklaşımla maddelerden oluşturulduğu ve test tasarımının yukarıdan aşağı (top-down) yaklaşımla bu modüllerin birleştirilmesiyle kurulduğu yöntemdir. Hibrit test birleştirme yaklaşımında test oluşturma iki aşamada ilerlemektedir: (1) aşağıdan yukarı düzeyde maddeler, bilgi ve içerik kısıtlarını karşılayan paralel modüller oluşturmak üzere birleştirilmekte, (2) yukarıdan aşağı düzeyde ise bu önceden oluşturulmuş modüller paneller halinde birleştirilerek çok aşamalı genel yapı oluşturulmaktadır (Luecht & Nungester, 1998; Breithaupt & Hare, 2007). Bu hibrit strateji her iki yaklaşımın avantajlarını bir araya getirmektedir: bottom-up yaklaşımı madde düzeyinde paralelliği sağlarken, top-down yaklaşımı modül ve panel düzeyindeki üst düzey kısıtların kontrolünü mümkün kılmaktadır. Bu nedenle bu çalışmada hibrit yöntem kullanılmıştır.

Benzer şekilde A-BÇAT'ta, ilk modül $\vartheta = 0$noktasında oluşturulmuş, diğer modüller ise kestirilen geçici ara yetenek düzeyi noktasına göre gölge test yaklaşımıyla birleştirilerek katılımcıya sunulmuştur. Hem S-BÇAT hem de A-BÇAT yaklaşımlarında test birleştirme için Rglpk algoritması kullanılmıştır (Makhorin, 2017). A-BÇAT'ta test birleştirme, Freeze-Refresh Mekanizması altında Gölge Test yaklaşımı kullanılarak yürütülmüştür (Choi & van der Linden, 2018).

Verilerin Analizi

Veri analizinden elde edilen sonuçların değerlendirilmesinde kestirilen ve gerçek yetenek parametreleri, Kök Ortalama Kare Hata (RMSE), Ortalama Mutlak Yanlılık (MAB) ve Yanlılık (BIAS) değerleri kullanılmıştır. RMSE değerleri, aşağıda verilen formül kullanılarak hesaplanmıştır. Bu formülde $n$toplam katılımcı sayısını, $\hat{\theta}$ kestirilen yetenek düzeyini ve θgerçek yetenek düzeyini ifade etmektedir.

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

MAB değerinin hesaplanmasında kullanılan formül aşağıda sunulmuştur.

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

BIAS değerinin hesaplanmasında kullanılan formül aşağıda sunulmuştur.

$$\text{BIAS} = \frac{\sum_{i=1}^{n}\widehat{\theta_i}-\theta_i}{n} \quad (4)$$

Yukarıdaki formüllerde $i$katılımcı numarasını, $n$katılımcı sayısını, $\theta_i i$ katılımcısının gerçek yeteneğini ve $\hat{\theta}_i$testten elde edilen i katılımcısına ait kestirilen yetenek düzeyini ifade etmektedir.

Buna ek olarak, farklı simülasyon koşulları altında iki farklı test tasarımının karşılaştırılması amacıyla elde edilen RMSE ve MAB değerleri üzerinden etki büyüklüğü değerleri hesaplanmıştır. Cohen's d değeri ve etki büyüklüğünün hesaplanmasında aşağıdaki formüller kullanılmıştır:

$$Harmonized\ Standard\ Deviation = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2}} \quad (5)$$

$$Cohen\ d = \frac{Mean\ Difference}{Harmonized\ Standard\ Deviation} \quad (6)$$

Yukarıdaki formüllerle hesaplanan Cohen's d değeri, $d < 0.20$olduğunda küçük etki, $0.20 < d < 0.50$olduğunda orta etki ve $0.80 < d$olduğunda büyük etki olarak yorumlanmaktadır (Cohen, 1988).

Madde güvenliğinin incelenmesinde, her bir madde için madde kullanım sıklığı oranı hesaplanmıştır. Madde kullanım sıklığı oranı aşağıdaki formül kullanılarak hesaplanmıştır.

$$Madde\ Kullanım\ Sıklığı\ Oranı = \frac{n_M}{n_T} \quad (7)$$

Formülde n_M, m maddesinin uygulandığı katılımcı sayısını, $n_T$ise toplam katılımcı sayısını ifade etmektedir. Bu oran, ilgili maddenin madde kullanım sıklığını göstermektedir.

BULGULAR

Bu bölümde ölçme kesinliği ve madde güvenliğine ilişkin bulgular başlıklar altında sunulmuştur.

Ölçme Kesinliği Bulguları

Bu bölümde, araştırma probleminin ölçme kesinliği boyutuna yanıt verebilmek amacıyla, aynı madde havuzu ve aynı yetenek dağılımları altında S-BÇAT ve A-BÇAT yöntemlerinden elde edilen yetenek kestirimleri karşılaştırılmıştır. Analiz edilen 72 koşuldan elde edilen RMSE, MAB, d ve BIAS sonuçları Tablo 6'da sunulmuştur.

Tablo 6

Farklı BÇAT Yaklaşımlarına Göre Tüm Koşullara İlişkin Bulgular

Koşul	Test Uzunluğu	Modül/Test Uuzunluğu Oranı	Yetenek Dağılımı	RMSE			MAB			BIAS	
				S-BÇAT	A-BÇAT	d_{RMSE}	A-BÇAT	A-BÇAT	d_{MAB}	S-BÇAT	A-BÇAT
1	20	B-K-K	Normal	0,382	0,371	1,657	0,300	0,291	1,587	0,008	0,007
2	20	B-K-K	Sağa Çarpık	0,423	0,404	2,369	0,315	0,303	1,604	0,050	0,040
3	20	B-K-K	Sola Çarpık	0,431	0,414	1,887	0,311	0,303	1,070	-0,028	-0,029
4	20	B-K-K	Uniform	0,563	0,535	3,305	0,443	0,419	3,411	-0,053	-0,048
5	20	O-O-O	Normal	0,382	0,362	2,438	0,300	0,286	1,990	0,008	0,012
6	20	O-O-O	Sağa Çarpık	0,417	0,399	2,407	0,314	0,302	1,723	0,046	0,037
7	20	O-O-O	Sola Çarpık	0,432	0,407	2,775	0,313	0,296	2,561	-0,023	-0,022
8	20	O-O-O	Uniform	0,552	0,524	2,235	0,432	0,409	2,715	-0,058	-0,047
9	20	K-K-B	Normal	0,398	0,364	4,013	0,314	0,285	4,164	0,008	0,004
10	20	K-K-B	Sağa Çarpık	0,431	0,399	5,287	0,325	0,301	4,368	0,039	0,039
11	20	K-K-B	Sola Çarpık	0,438	0,405	4,146	0,322	0,297	4,928	-0,021	-0,022
12	20	K-K-B	Uniform	0,559	0,519	3,640	0,440	0,406	3,592	-0,047	-0,047
13	30	B-K-K	Normal	0,338	0,323	1,884	0,266	0,253	2,148	0,007	0,009
14	30	B-K-K	Sağa Çarpık	0,380	0,361	2,540	0,281	0,268	2,004	0,044	0,039
15	30	B-K-K	Sola Çarpık	0,388	0,367	2,808	0,277	0,266	1,843	-0,025	-0,018
16	30	B-K-K	Uniform	0,496	0,463	3,486	0,387	0,360	3,366	-0,045	-0,044
17	30	O-O-O	Normal	0,333	0,321	1,604	0,262	0,252	1,675	0,007	0,005
18	30	O-O-O	Sağa Çarpık	0,369	0,349	2,674	0,276	0,263	2,178	0,038	0,032
19	30	O-O-O	Sola Çarpık	0,382	0,356	2,886	0,275	0,260	2,513	-0,023	-0,015
20	30	O-O-O	Uniform	0,476	0,456	2,233	0,370	0,355	1,870	-0,052	-0,044
21	30	K-K-B	Normal	0,34	0,317	2,568	0,268	0,251	2,441	0,006	0,008
22	30	K-K-B	Sağa Çarpık	0,372	0,344	4,317	0,279	0,259	3,640	0,037	0,032

Tablo 6 (Devam)

Koşul	Test Uzunluğu	Modül/Test Uzunluğu Oranı	Yetenek Dağılımı	RMSE			MAB			BIAS	
				S-BÇAT	A-BÇAT	d_{RMSE}	A-BÇAT	A-BÇAT	d_{MAB}	S-BÇAT	A-BÇAT
23	30	K-K-B	Sola Çarpık	0,385	0,356	3,615	0,28	0,258	4,004	-0,021	-0,017
24	40	K-K-B	Uniform	0,477	0,444	3,127	0,372	0,346	2,886	-0,047	-0,043
25	40	B-K-K	Normal	0,306	0,295	1,579	0,24	0,232	1,456	0,005	0,002
26	40	B-K-K	Sağa Çarpık	0,337	0,328	1,638	0,251	0,244	1,554	0,036	0,036
27	40	B-K-K	Sola Çarpık	0,350	0,336	2,158	0,249	0,243	1,206	-0,021	-0,015
28	40	B-K-K	Uniform	0,435	0,420	1,771	0,336	0,325	1,579	-0,047	-0,038
29	40	O-O-O	Normal	0,310	0,293	2,136	0,244	0,230	2,158	0,005	0,005
30	40	O-O-O	Sağa Çarpık	0,339	0,324	2,513	0,254	0,240	2,548	0,034	0,029
31	40	O-O-O	Sola Çarpık	0,353	0,327	3,476	0,254	0,237	3,094	-0,019	-0,018
32	40	O-O-O	Uniform	0,433	0,413	2,361	0,336	0,320	2,297	-0,045	-0,042
33	40	K-K-B	Normal	0,317	0,291	3,733	0,250	0,230	3,350	0,005	0,002
34	40	K-K-B	Sağa Çarpık	0,345	0,317	3,744	0,259	0,239	3,640	0,030	0,029
35	40	K-K-B	Sola Çarpık	0,357	0,330	3,610	0,259	0,240	3,183	-0,018	-0,014
36	40	K-K-B	Uniform	0,433	0,412	2,219	0,337	0,317	2,872	-0,039	-0,044

Not. K: Küçük, O: Orta, B: Büyük.

Tablo 6'da görüldüğü üzere, tüm koşullarda S-BÇAT'tan elde edilen RMSE değerleri [0.306, 0.563] aralığında değişmektedir. A-BÇAT'tan elde edilen RMSE değerleri ise [0.291, 0.535] aralığındadır. Tabloda görüldüğü gibi, S-BÇAT yönteminden elde edilen RMSE değerleri tüm koşullarda A-BÇAT'tan elde edilen RMSE değerlerinden daha yüksektir. Bu durum, A-BÇAT'ın tüm koşullarda S-BÇAT'tan daha etkili olduğunu göstermektedir. Ayrıca ilgili sütunda görüldüğü üzere, Cohen's d etki büyüklüğü tüm koşullarda [d > .80] düzeyindedir ve A-BÇAT yöntemi tüm koşullarda S-BÇAT yönteminden daha büyük bir etki göstermektedir. RMSE sonuçlarına benzer şekilde, S-BÇAT'tan elde edilen MAB değerleri de tüm koşullarda A-BÇAT'tan elde edilen MAB değerlerinden daha yüksektir. MAB değerleri de A-BÇAT'ın S-BÇAT'a göre daha etkili kestirim yaptığını doğrulamaktadır. Benzer biçimde, ilgili sütunda görüldüğü üzere Cohen's d etki büyüklüğü tüm koşullarda [d > .80] düzeyindedir, bu da A-BÇAT yönteminin tüm koşullarda S-BÇAT yönteminden daha büyük etkiye sahip olduğunu göstermektedir.

Hem RMSE hem de MAB sütunlarından görüldüğü üzere, test uzunluğu arttıkça her iki test yaklaşımının ölçme kesinliği belirgin biçimde artmaktadır. Buna karşılık modül/test uzunluğu oranına göre, S-BÇAT'ta K-K-B oranının B-K-K ve O-O-O oranlarına göre daha düşük ölçme kesinliği sunduğu, buna karşın A-BÇAT'ta üç oranın da benzer ölçme kesinliği sağladığı söylenebilir. Yetenek dağılımlarına göre, her iki yöntemde de en yüksek ölçme kesinliği normal dağılımda, ardından sağa ve sola çarpık dağılımlarda elde edilmiştir. Uniform dağılımın ise her iki test yaklaşımı için de en düşük ölçme kesinliğine sahip olduğu ifade edilebilir. Çalışmanın geri kalanında, test uzunluğu, modül/test uzunluğu oranı ve yetenek dağılımına göre S-BÇAT ve A-BÇAT yöntemlerinin ölçme kesinliği sonuçları Tablo 6'dan elde edilen ortalama değerler üzerinden karşılaştırılmıştır.

Test Uzunluğuna Göre Ölçme Kesinliği Bulguları

Test uzunluğuna göre Tablo 6'daki ilgili hücrelerin ortalamalarından elde edilen sonuçlar Tablo 7'de sunulmuştur.

Tablo 7

Test Uzunluğuna Göre RMSE, MAB and d Değerleri

Test Uzunluğu	RMSE			MAB			BIAS	
	S-BÇAT	A-BÇAT	d_{RMSE}	S-BÇAT	A-BÇAT	d_{MAB}	S-BÇAT	A-BÇAT
20	0,451	0,425	3,013	0,344	0,325	2,809	-0,006	-0,006
30	0,387	0,365	2,783	0,293	0,277	2,516	-0,002	-0,001
40	0,360	0,341	2,578	0,272	0,258	2,411	-0,006	-0,006

Tablo 7 incelendiğinde, A-BÇAT yaklaşımının 20, 30 ve 40 test uzunluklarının tamamında yetenek düzeyini S-BÇAT'tan daha iyi kestirdiği söylenebilir. 20 ile 30 test uzunlukları arasındaki RMSE farkı daha yüksekken, 30 ile 40 test uzunlukları arasındaki RMSE farkının azaldığı görülmektedir. Hem RMSE farkı hem de MAB farkı dikkate alındığında, bu durum test uzunluğu artırılsa bile yetenek kestirimindeki etkililiğin test uzunluğu ile orantılı biçimde artmayacağını göstermektedir. Azalan getiriler yasası, üretim faktörleri arttıkça üretimin aynı oranda artmayacağını ve oransal faydanın giderek azalacağını savunmaktadır. Testlerde de test uzunluğunun artırılması belirli bir noktaya kadar ölçme kesinliğini artıracaktır. Ancak belirli bir noktadan sonra test uzunluğunu artırmak, ölçme kesinliğinde beklenen düzeyde iyileşme sağlamayacak; öğrencinin yorgunluk, sıkılma ve tükenme gibi fizyolojik ve psikolojik etkileri nedeniyle beklenen verim elde edilemeyecektir. Bu nedenle test uzunluğunun iyi belirlenmesi gerekmektedir. Bu çalışmada 20 maddelik test uzunluğunun, 30 ve 40 maddelik test uzunluklarına kıyasla daha düşük ölçme kesinliği gösterdiği anlaşılmaktadır. Dolayısıyla bu araştırmanın sonuçlarına göre 30 ya da 40 maddelik test uzunluklarının daha iyi sonuç verdiği söylenebilir. 50, 60 ve 70 maddelik test uzunluklarının ise uyarlanabilir test yaklaşımlarının mantığıyla uyumlu olmadığı, ayrıca azalan getiriler yasasına göre ölçme kesinliğinde orantılı bir artış beklenmemesi gerektiği düşünülmektedir (Yasuda, Mae, Hull & Taniguchi, 2021).

Yetenek Dağılımına Göre Ölçme Kesinliği Sonuçları

Yetenek dağılımına göre Tablo 6'daki ilgili hücrelerin ortalamalarından elde edilen sonuçlar Tablo 8'de sunulmuştur.

Tablo 8

Yetenek Dağılımına Göre RMSE, MAB, BIAS and d Değerleri

Yetenek Dağılımı	RMSE			MAB			BIAS	
	S-BÇAT	A-BÇAT	d_{RMSE}	S-BÇAT	A-BÇAT	d_{MAB}	S-BÇAT	A-BÇAT
Normal	0,345	0,326	2,401	0,272	0,257	2,330	0,007	0,006
Sağa Çarpık	0,379	0,358	3,054	0,284	0,269	2,584	0,039	0,035
Sola Çarpık	0,391	0,366	3,040	0,282	0,267	2,711	-0,022	-0,019
Uniform	0,492	0,465	2,708	0,384	0,362	2,732	-0,048	-0,044

Tablo 8 incelendiğinde, A-BÇAT yaklaşımının dört farklı yetenek dağılımında (normal, sağa çarpık, sola çarpık ve uniform) RMSE, MAB ve d değerlerine göre S-BÇAT yaklaşımından daha etkili yetenek kestirimi sağladığı görülmektedir. Her iki yöntem de en başarılı kestirimi normal dağılımda, ardından sağa çarpık ve sola çarpık dağılımlarda gerçekleştirmektedir. Her iki yöntemin de en düşük ölçme kesinliği uniform

dağılımda elde edilmiştir. Bununla birlikte, etki büyüklüğü değerlerine göre normal dağılımda fark daha küçükken, A-BÇAT yöntemi özellikle sağa çarpık, sola çarpık ve uniform dağılımlarda daha etkili kestirim yapmaktadır. Bu nedenle çarpık ya da uniform yetenek dağılımlarında A-BÇAT yönteminin tercih edilebileceği söylenebilir.

Modül/Test Uzunluğu Oranına Göre Ölçme Kesinliği Sonuçları

Modül/test uzunluğu oranına göre Tablo 6'daki ilgili hücrelerin ortalamalarından elde edilen sonuçlar Tablo 9'da sunulmuştur.

Tablo 9

Modül/Test Uzunluğu Oranına Göre RMSE, MAB, BIAS ve d Değerleri

Modül/Test Uzunluğu Oranı	RMSE			MAB			BIAS	
	S-BÇAT	A-BÇAT	d_{RMSE}	S-BÇAT	A-BÇAT	d_{MAB}	S-BÇAT	A-BÇAT
B-K-K	0,402	0,385	2,257	0,305	0,292	1,902	-0,006	-0,005
O-O-O	0,398	0,378	2,478	0,303	0,288	2,277	-0,007	-0,006
K-K-B	0,404	0,375	3,668	0,309	0,286	3,589	-0,006	-0,006

Not. K: Küçük, O: Orta, B: Büyük

Tablo 9 incelendiğinde, A-BÇAT yaklaşımının tüm modül/test uzunluğu oranı düzeylerinde (B-K-K, O-O-O, K-K-B) RMSE, MAB ve d değerlerine göre S-BÇAT yaklaşımından daha etkili yetenek kestirimi sağladığı görülmektedir. Cohen's d değerlerine göre, A-BÇAT ile S-BÇAT yaklaşımları arasındaki ölçme kesinliği farkı en düşük B-K-K oranında, O-O-O oranında ise benzer düzeydedir. K-K-B oranında ise iki yaklaşım arasındaki ölçme kesinliği farkı en yüksek düzeye ulaşmıştır. Bu durumda, S-BÇAT yaklaşımında son modüldeki madde sayısı arttıkça ölçme kesinliğinin azaldığı şeklinde yorum yapılabilir. Buna karşılık A-BÇAT yaklaşımında modüller sabit olmadığından, modül/test oranı uzunluğundan daha az etkilenmektedir. Ayrıca A-BÇAT yaklaşımında son modül uzunluğunun artırılması bu yaklaşımın ölçme kesinliğini artırmıştır. Bunun yanında, üç modül/test uzunluğu oranının tamamında A-BÇAT'ın S-BÇAT'tan daha büyük etki gösterdiği de özellikle belirtilmelidir.

Madde Güvenliği Sonuçları

Bu bölümde, madde güvenliğine ilişkin araştırma problemlerine yanıt verebilmek amacıyla, tüm koşullara ait kullanılan madde sayısı ve ortalama madde kullanım sıklığı istatistikleri Tablo 10'da sunulmuştur.

Tablo 10

Madde Kullanım Sıklığı Oranları

Koşul	Test Uzunluğu	Modül/Test Uzunluğu Oranı	Yetenek Dağılımı	Toplam Madde Sayısı	Kullanılan Madde Sayısı			Ortalama Madde Kullanım Oranı	
					S-BÇAT	A-BÇAT	Fark	S-BÇAT	A-BÇAT
1	20	B-K-K	Normal	400	105	163	-58	0,190	0,123
2	20	B-K-K	Sağa Çarpık	400	105	160	-55	0,167	0,125
3	20	B-K-K	Sola Çarpık	400	105	156	-51	0,148	0,128

Tablo 10 (Devam)

Koşul	Test Uzunluğu	Modül/Test Uzunluğu Oranı	Yetenek Dağılımı	Toplam Madde Sayısı	Kullanılan Madde Sayısı			Ortalama Madde Kullanım Oranı	
					S-BÇAT	A-BÇAT	Fark	S-BÇAT	A-BÇAT
4	20	B-K-K	Uniform	400	105	174	-69	0,192	0,115
5	20	O-O-O	Normal	400	120	175	-55	0,167	0,114
6	20	O-O-O	Sağa Çarpık	400	120	173	-53	0,147	0,116
7	20	O-O-O	Sola Çarpık	400	120	170	-50	0,190	0,118
8	20	O-O-O	Uniform	400	120	181	-61	0,167	0,110
9	20	K-K-B	Normal	400	135	176	-41	0,148	0,114
10	20	K-K-B	Sağa Çarpık	400	135	176	-41	0,190	0,114
11	20	K-K-B	Sola Çarpık	400	135	173	-38	0,167	0,116
12	20	K-K-B	Uniform	400	135	185	-50	0,148	0,108
13	30	B-K-K	Normal	400	156	221	-65	0,192	0,136
14	30	B-K-K	Sağa Çarpık	400	156	221	-65	0,167	0,136
15	30	B-K-K	Sola Çarpık	400	156	221	-65	0,147	0,136
16	30	B-K-K	Uniform	400	156	232	-76	0,190	0,129
17	30	O-O-O	Normal	400	180	229	-49	0,167	0,131
18	30	O-O-O	Sağa Çarpık	400	180	230	-50	0,148	0,130
19	30	O-O-O	Sola Çarpık	400	180	227	-47	0,190	0,132
20	30	O-O-O	Uniform	400	180	244	-64	0,167	0,123
21	30	K-K-B	Normal	400	204	239	-35	0,148	0,126
22	30	K-K-B	Sağa Çarpık	400	204	236	-32	0,192	0,127
23	30	K-K-B	Sola Çarpık	400	204	235	-31	0,167	0,128
24	40	K-K-B	Uniform	400	204	252	-48	0,147	0,119
25	40	B-K-K	Normal	400	210	267	-57	0,190	0,150
26	40	B-K-K	Sağa Çarpık	400	210	267	-57	0,167	0,150
27	40	B-K-K	Sola Çarpık	400	210	264	-54	0,148	0,152
28	40	B-K-K	Uniform	400	210	281	-71	0,190	0,142
29	40	O-O-O	Normal	400	240	285	-45	0,167	0,14
30	40	O-O-O	Sağa Çarpık	400	240	282	-42	0,148	0,142
31	40	O-O-O	Sola Çarpık	400	240	282	-42	0,192	0,142
32	40	O-O-O	Uniform	400	240	297	-57	0,167	0,135
33	40	K-K-B	Normal	400	270	289	-19	0,147	0,138
34	40	K-K-B	Sağa Çarpık	400	270	288	-18	0,190	0,139
35	40	K-K-B	Sola Çarpık	400	270	286	-16	0,167	0,140
36	40	K-K-B	Uniform	400	270	304	-34	0,148	0,132

Madde havuzundan daha fazla sayıda maddenin kullanılması, maddelerin açığa çıkma olasılığını azaltması bakımından hem madde kullanım sıklığını düşürmekte hem de madde güvenliğini artırmaktadır. Tablo 10 incelendiğinde, A-BÇAT'ın tüm koşullarda S-BÇAT'a göre madde havuzundan daha fazla sayıda madde kullandığı görülmektedir. Özellikle test uzunluğu arttıkça kullanılan madde sayıları birbirine yaklaşsa da, A-BÇAT'ın yine de daha fazla madde kullandığı özellikle belirtilmelidir. Ayrıca Tablo 10'a göre, ortalama madde kullanım sıklığı S-BÇAT için 0.168, A-BÇAT için ise 0.129 olarak bulunmuştur. A-BÇAT'ın ortalama madde kullanım sıklığının S-BÇAT'tan daha düşük olması, bu yaklaşımın madde havuzundan daha fazla sayıda madde kullandığının ve dolayısıyla madde havuzunu daha etkili biçimde kullandığının bir göstergesi olarak yorumlanabilir. Bu çalışmada, S-BÇAT ve A-BÇAT'ta madde kullanım sıklığını sınırlandırmak amacıyla, S-BÇAT'ta 3 panel oluşturulmuş, A-BÇAT'ta ise Uygunsuzluk yöntemi ile madde kullanım sıklığı 0.33 düzeyinde sabitlenmiştir. Bu bulgulara göre, ortalama madde kullanım sıklığı değerleri bakımından A-

BÇAT'ın S-BÇAT'tan daha iyi sonuçlar verdiği görülmektedir. Dolayısıyla madde güvenliği açısından A-BÇAT'ın S-BÇAT'tan daha güvenli olduğu söylenebilir.

Aşağıda, ilgili değişkenlere göre madde kullanım sıklıklarının incelenmesine yönelik bazı örnek grafikler sunulmuştur. Çalışmanın devamında, madde güvenliği alt problemleri doğrultusunda test uzunluğu, yetenek dağılımı ve modül/test uzunluğu açısından madde kullanım sıklığı ile kullanılan madde sayısı, Tablo 10'daki ilgili hücrelerin ortalamaları üzerinden analiz edilmiştir.

Test Uzunluğuna Göre Madde Güvenliği Sonuçları

Farklı test uzunlukları için Tablo 10'daki ilgili hücrelerin ortalamalarından elde edilen sonuçlar Tablo 11'de sunulmuştur.

Tablo 11

Test Uzunluğuna Göre Kullanılan Madde Sayısı ve Ortalama Madde Kullanım Sıklığı

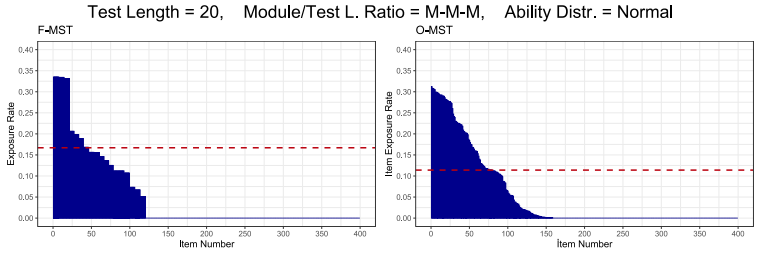
Test Uzunluğu	Kullanılan Madde Sayısı		Ortalama Madde Kullanım Sıklığı Oranı	
	S-BÇAT	A-BÇAT	S-BÇAT	A-BÇAT
20	120	172	0,168	0,117
30	180	232	0,168	0,130
40	240	283	0,168	0,142

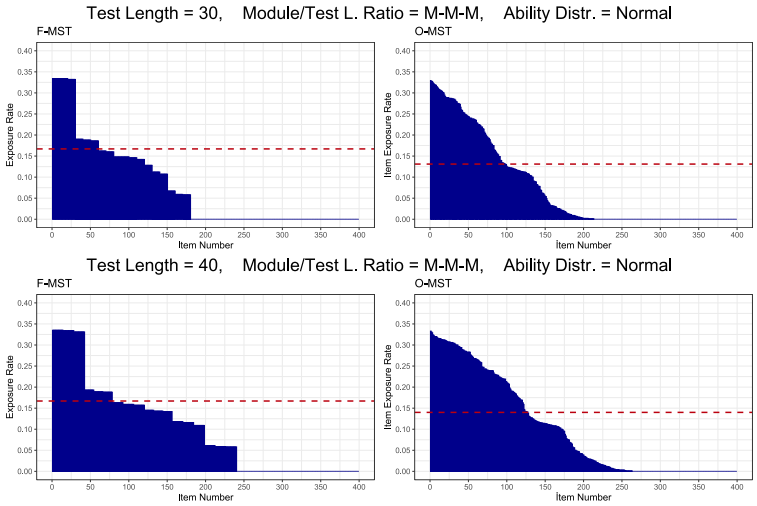
Tablo 11 incelendiğinde, 20 maddelik test uzunluğunda hem kullanılan madde sayısı hem de ortalama madde kullanım sıklığı açısından A-BÇAT yaklaşımının madde güvenliği bakımından daha iyi sonuçlar sağladığı söylenebilir. Benzer şekilde 30 ve 40 maddelik test uzunluklarında da, hem kullanılan madde sayısı hem de ortalama madde kullanım sıklığı açısından A-BÇAT yaklaşımının S-BÇAT'tan daha iyi sonuçlar verdiği ifade edilebilir. Bununla birlikte, test uzunluğu arttıkça S-BÇAT ile A-BÇAT yaklaşımları arasındaki kullanılan madde sayısı farkı azalmaktadır. Bu bulguya benzer biçimde, S-BÇAT'ta tüm test uzunluklarında ortalama madde kullanım sıklığı %16.8 olarak bulunurken, A-BÇAT için bu değerler sırasıyla %11.7, %13.0 ve %14.2 olarak hesaplanmıştır. Bu durumda, kısa test uzunluklarında A-BÇAT ile S-BÇAT arasındaki madde güvenliği farkı A-BÇAT lehine daha belirginken, test uzunluğu arttıkça her iki yaklaşımın madde güvenliği verimliliğinin birbirine yakınsadığı görülmektedir. Dolayısıyla her üç test uzunluğunda da test ve madde güvenliği açısından A-BÇAT'ın S-BÇAT'tan daha verimli olduğu, bununla birlikte özellikle kısa testlerde A-BÇAT'ın madde güvenliği bakımından daha belirgin üstünlük sağladığı söylenebilir.

Şekil 3, aynı yetenek dağılımı ve aynı modül/test uzunluğu oranı altında, 20, 30 ve 40 test uzunluklarına sahip üç farklı koşul için madde kullanım sıklığı grafiklerini göstermektedir.

Şekil 3

Test Uzunluğuna Göre Madde Kullanım Sıklığı Grafiği





Grafiklerdeki mavi bölümler madde kullanım sıklığını göstermektedir. Sütun grafikleri, maddeler en yüksek madde kullanım sıklığından en düşüğe doğru sıralanarak oluşturulmuştur. Grafik üzerindeki kırmızı kesikli çizgiler ise ilgili koşulun ortalama madde kullanım sıklığını göstermektedir. Şekil 3, A-BÇAT yaklaşımının tüm test uzunluğu koşullarında S-BÇAT yaklaşımına göre daha yüksek sayıda madde kullandığını ve daha düşük ortalama madde kullanım sıklığına sahip olduğunu göstermektedir. Buna karşılık S-BÇAT yaklaşımında özellikle yönlendirme modülündeki maddeler nedeniyle bazı maddelerin kullanım sıklığı 0.33 düzeyine ulaşmaktadır ve bu maddelerin açığa çıkma olasılığının daha yüksek olduğu düşünülmektedir. A-BÇAT yaklaşımında ise bazı maddelerin kullanım sıklığı 0.34 düzeyinde olsa da sonraki maddelerin kullanım sıklıkları hızlı biçimde azalmaktadır. Ayrıca A-BÇAT yaklaşımında daha fazla sayıda maddenin kullanılması hem madde güvenliği hem de madde havuzunun daha etkili kullanılması açısından avantajlı görünmektedir.

Yetenek Dağılımlarına Göre Madde Güvenliği Sonuçları

Yetenek dağılımına göre Tablo 10'daki ilgili hücrelerin ortalamalarından elde edilen sonuçlar Tablo 12'de sunulmuştur.

Tablo 12

Yetenek Dağılımına Göre Kullanılan Madde Sayısı ve Ortalama Madde Kullanım Sıklığı

Yetenek Dağılımı	Kullanılan Madde Sayısı		Ortalama Madde Kullanım Sıklığı Oranı	
	S-BÇAT	A-BÇAT	S-BÇAT	A-BÇAT
Normal	180	227	0,168	0,130
Sağa Çarpık	180	226	0,168	0,131
Sola Çarpık	180	224	0,168	0,132
Uniform	180	239	0,168	0,124

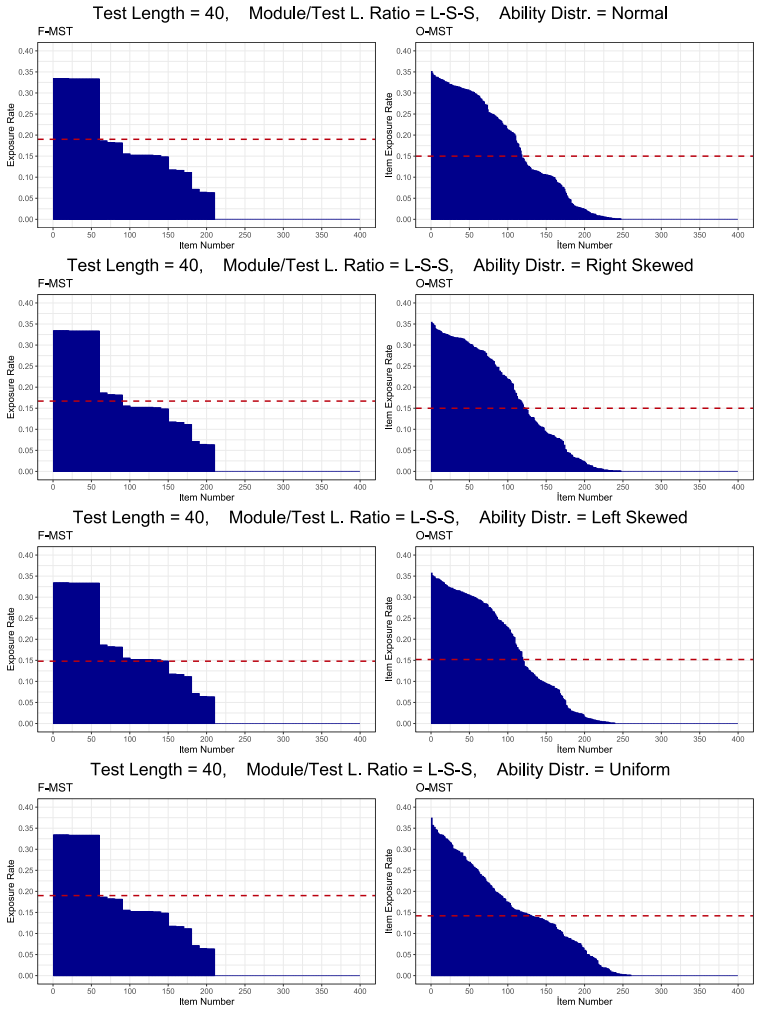
Tablo 12 incelendiğinde, normal, sağa çarpık, sola çarpık ve uniform dağılımların tamamında hem kullanılan madde sayısı hem de ortalama madde kullanım sıklığı açısından A-BÇAT yaklaşımının S-BÇAT'tan daha iyi sonuçlar sunduğu söylenebilir. Normal, sağa çarpık, sola çarpık ve uniform dağılımların tümünde, A-BÇAT yaklaşımının S-BÇAT'a göre daha yüksek kullanılan madde sayısına sahip olduğu görülmektedir. Ayrıca ilk üç yetenek dağılımında S-BÇAT ile A-BÇAT yaklaşımları arasındaki kullanılan madde sayısı farkı benzer düzeydeyken, uniform dağılımda A-BÇAT yaklaşımının S-BÇAT'a göre daha fazla madde kullandığı ve daha düşük madde kullanım sıklığına sahip olduğu görülmektedir. Dolayısıyla uniform

dağılımlarda madde güvenliği açısından A-BÇAT yaklaşımının S-BÇAT'tan daha iyi sonuçlar ürettiği ifade edilebilir.

Şekil 4, 40 maddelik test uzunluğu ve B-K-K modül/test uzunluğu oranı altında, dört farklı yetenek dağılımına (normal, sağa çarpık, sola çarpık ve uniform) göre madde bazında kullanım sıklığı grafiklerini göstermektedir.

Şekil 4

Yetenek Dağılımına Göre Madde Kullanım Sıklığı Grafiği



Şekil 4, tüm farklı yetenek dağılımlarında A-BÇAT yaklaşımının S-BÇAT yaklaşımına göre daha yüksek sayıda madde kullandığını ve daha düşük ortalama madde kullanım sıklığına (kırmızı kesikli çizgiler) sahip olduğunu göstermektedir.

Modül/Test Uzunluğu Oranına Göre Madde Güvenliği Sonuçları

Modül/test uzunluğu oranına göre Tablo 10'daki ilgili hücrelerin ortalamalarından elde edilen sonuçlar Tablo 13'te sunulmuştur.

Tablo 13

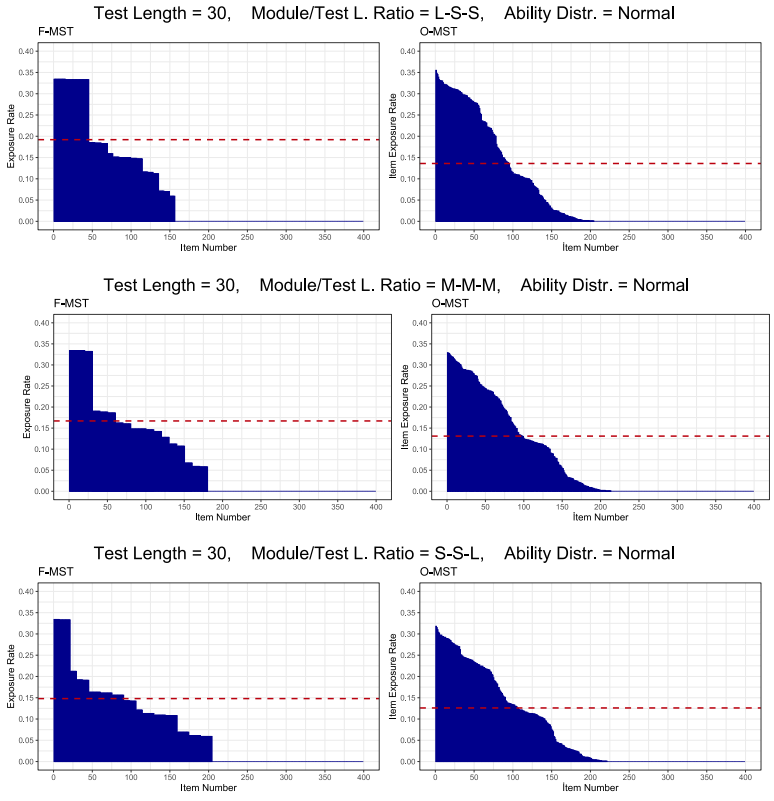
Modül/Test Uzunluğu Oranına Göre Kullanılan Madde Sayısı ve Ortalama Madde Kullanım Sıklıkları

Modül/Test Uzunluğu Oranı	Kullanılan Madde Sayısı		Ortalama Madde Kullanım Sıklığı Oranı	
	S-BÇAT	A-BÇAT	S-BÇAT	A-BÇAT
B-K-K	157	219	0,174	0,135
O-O-O	180	231	0,168	0,128
K-K-B	203	237	0,163	0,125

Tablo 13 incelendiğinde, B-K-K, O-O-O ve K-K-B modül/test uzunluğu oranlarının tamamında hem kullanılan madde sayısı hem de ortalama madde kullanım sıklıkları açısından A-BÇAT yaklaşımının S-BÇAT'tan daha iyi sonuçlar sağladığı görülmektedir. Bununla birlikte, S-BÇAT ile A-BÇAT arasındaki kullanılan madde sayısı farkı, yönlendirme modülündeki madde sayısının daha fazla olması nedeniyle B-K-K oranında daha yüksektir. K-K-B oranına doğru ilerledikçe kullanılan madde sayısı farkı azalmaktadır. K-K-B oranında ise iki yaklaşım arasındaki kullanılan madde sayısı farkı en düşük düzeye inmektedir.

Şekil 5

Modül/Test Uzunluğu Oranına Göre Madde Kullanım Sıklığı Grafiği



Şekil 5, tüm farklı modül/test uzunluğu oranı koşullarında A-BÇAT yaklaşımının S-BÇAT yaklaşımına göre daha yüksek sayıda madde kullandığını ve daha düşük ortalama madde kullanım sıklığına (kırmızı kesikli çizgiler) sahip olduğunu göstermektedir. Bununla birlikte, B-K-K oranında yönlendirme modülünün uzunluğu daha fazla olduğundan, S-BÇAT yaklaşımında yönlendirme modülündeki maddelerin kullanım sıklığı daha yüksektir. Bu durum madde güvenliği açısından sorunlara yol açabilir. K-K-B oranında ise yönlendirme modülünün uzunluğu daha kısa olduğu için madde kullanım sıklığı daha düşüktür. Ancak K-K-B oranının dezavantajı, diğer iki modül/test uzunluğu oranına göre daha düşük ölçme kesinliğine sahip olmasıdır. Bu nedenle modül/test uzunluğu oranına karar verilirken hem madde güvenliği hem de ölçme kesinliği birlikte dikkate alınmalı ve en uygun denge noktası belirlenmelidir.

TARTIŞMA VE SONUÇ

Son yıllarda, özellikle geniş ölçekli değerlendirmelerde test yönetiminde uyarlanabilir testlerin kullanımına yönelik ilgi giderek artmaktadır (Ebenbeck & Gebhardt, 2022; Tomashev et al., 2018; Yamamoto et al., 2018; Zheng & Chang, 2015). S-BÇAT, sabit modül ve panel yapısı sayesinde test yönetiminin kolay olduğu bir uyarlanabilir test yaklaşımıdır. Buna karşılık A-BÇAT, çok sayıda karmaşık test belirtimi ve kısıtı altında başarılı çözümler sunan, BBT ile BÇAT'ın hibrit bir yapısı olarak değerlendirilebilecek yeni bir test yaklaşımıdır (Choi & van der Linden, 2018; van der Linden, 2021). Bu çalışmada S-BÇAT ile A-BÇAT, ölçme kesinliği ve madde kullanım sıklığı açısından karşılaştırılmıştır. Elde edilen bulgular aşağıda başlıklar altında sunulmuştur.

Ölçme Kesinliği

Bu çalışmanın sonuçları, A-BÇAT'ın hem RMSE hem de MAB istatistiklerine göre S-BÇAT'tan daha iyi ölçme kesinliği sunduğunu göstermektedir. Benzer şekilde, ölçme kesinliği açısından tüm koşullarda A-BÇAT'ın S-BÇAT'a göre daha büyük bir etki büyüklüğüne sahip olduğu görülmüştür (Cohen's $d > 0.80$). Başka bir ifadeyle, A-BÇAT yaklaşımı yetenek düzeyini S-BÇAT'a göre daha etkili biçimde kestirmektedir.

Alanyazında bu bulguyla benzer sonuçlara ulaşan çok sayıda çalışma bulunmaktadır. Zheng ve Chang (2015), BBT, A-BÇAT ve S-BÇAT'ı karşılaştırdıkları simülasyon çalışmalarında, BBT ile A-BÇAT'ın çok benzer ölçme kesinliği sağladığını, buna karşılık S-BÇAT'ın daha düşük ölçme kesinliği sonuçları ürettiğini belirtmiştir. Tay (2015), A-BÇAT ile S-BÇAT'ın sınıflama doğruluğu ve sınıflama tutarlılığını karşılaştırdığı doktora tezinde, tüm koşullarda A-BÇAT'ın daha iyi sonuçlar verdiğini ifade etmiştir. van der Linden ve Diao (2016), BBT, A-BÇAT, S-BÇAT ve DT'yi karşılaştırdıkları çalışmalarında, ölçme kesinliği açısından BBT ile A-BÇAT'ın en iyi sonuçları verdiğini, bunları S-BÇAT'ın izlediğini, DT'nin ise son sırada yer aldığını raporlamıştır. Han ve Guo (2016), BBT ile A-BÇAT'ın birbirine çok yakın ölçme kesinliği sunduğunu, buna karşılık S-BÇAT'ın bu iki yaklaşımdan daha düşük ölçme kesinliği sağladığını belirtmiştir. van der Linden (2021) ise DT, BBT, A-BÇAT ve S-BÇAT yaklaşımlarını karşılaştırmış ve A-BÇAT'ın ölçme kesinliği açısından S-BÇAT'tan daha başarılı olduğunu ortaya koymuştur. Sonuç olarak, A-BÇAT ile S-BÇAT'ı karşılaştıran alanyazın bulguları ile mevcut çalışmanın sonuçlarının oldukça benzer olduğu görülmektedir (Tay, 2015; van der Linden, 2021; Zheng & Chang, 2015).

Bu çalışmada, 20, 30 ve 40 maddelik test uzunluklarının tamamında A-BÇAT'ın S-BÇAT'tan daha iyi ölçme kesinliği sağladığı sonucuna ulaşılmıştır. Bununla birlikte, 20 maddelik test uzunluğunda A-BÇAT yaklaşımı S-BÇAT'a göre daha belirgin biçimde üstünken, test uzunluğu arttıkça iki yaklaşımın ölçme kesinliği sonuçlarının birbirine yakınsadığı görülmüştür. Yasuda, Mae, Hull ve Taniguchi (2021), BBT'nin test uzunluğu üzerine yaptıkları çalışmada, test uzunluğu arttıkça ölçme kesinliği ve yanlılığın anlamlı düzeyde değişmediği sonucuna ulaşmıştır. Dolayısıyla mevcut çalışmada da her iki yaklaşımın ölçme kesinliğinin test uzunluğu arttıkça orantılı biçimde azalmadığı görülmektedir. van der Linden (2021), bu durumu azalan getiriler yasası ile açıklamaktadır. Ayrıca çok uzun testlerin, uyarlanabilir test yaklaşımının temel mantığıyla çeliştiği de belirtilmektedir (van der Linden, 2021).

Çalışmada iki BÇAT yaklaşımı normal, sağa çarpık, sola çarpık ve uniform olmak üzere dört farklı yetenek dağılımı altında karşılaştırılmıştır. Bulgular, A-BÇAT yaklaşımının dört yetenek dağılımının tamamında S-BÇAT'tan daha iyi ölçme kesinliği sunduğunu göstermiştir. Her iki yaklaşımda da ölçme kesinliği en yüksek normal dağılımda, ardından sağa çarpık ve sola çarpık dağılımlarda elde edilmiştir. Uniform dağılımda ise her iki yaklaşım için ölçme kesinliği düşüktür. Bununla birlikte, ölçülen yapının doğası ve evrenin dağılım özellikleri dikkate alındığında, eğitim alanında normal olmayan örneklemlerle sık karşılaşılmaktadır (MEB, 2021). Bu noktada özellikle sağa ve sola çarpık dağılımlarda A-BÇAT'ın ölçme kesinliğinin S-BÇAT'tan belirgin biçimde daha iyi olduğu sonucuna ulaşılmıştır. Sonuç olarak A-BÇAT yaklaşımı değişen yetenek dağılımlarından daha az etkilenirken, S-BÇAT yaklaşımı daha fazla etkilenmektedir.

Araştırmada iki BÇAT yaklaşımı K-K-B, O-O-O ve B-K-K olmak üzere üç farklı modül/test uzunluğu oranı altında karşılaştırılmıştır. Modül/test uzunluğu oranına göre A-BÇAT'ın tüm oranlarda S-BÇAT'tan daha yüksek ölçme kesinliği sunduğu belirlenmiştir. Alanyazında, yönlendirme modülünün uzun olduğu

durumlarda S-BÇAT'ın daha iyi ölçme kesinliği sağladığını gösteren çalışmalar bulunmaktadır (Boztunç, 2019; Cai, Anthony, Albano, & Roussos, 2021; Kim & Plake, 1993; Zheng, 2016). Bu çalışmada da alanyazınla benzer biçimde, S-BÇAT'ta ilk modülün uzunluğu arttıkça ölçme kesinliğinin yükseldiği görülmüştür. Yine literatürdeki sonuçlarla tutarlı olarak, ikinci ve üçüncü aşamadaki modül uzunlukları arttığında, bu tasarımlarda çok sayıda madde kullanım gereksinimi nedeniyle S-BÇAT'ın ölçme kesinliği azalmaktadır. Buna karşılık A-BÇAT modül/test uzunluğu oranından daha az etkilenmekte ve özellikle son aşamadaki madde sayısı arttığında diğer oranlara göre daha yüksek ölçme kesinliği sağlamaktadır. Başka bir ifadeyle, mevcut çalışmanın en önemli bulgularından biri, S-BÇAT'ta ilk modül uzunluğu arttıkça ölçme kesinliğinin artması; A-BÇAT'ta ise son modül uzunluğu arttıkça ölçme kesinliğinin artmasıdır. Bununla birlikte, genel modül/test uzunluğu oranı açısından A-BÇAT tüm koşullarda ölçme kesinliği bakımından S-BÇAT'tan daha üstündür.

Madde Güvenliği

Madde güvenliği açısından, tüm koşullarda A-BÇAT'ın S-BÇAT'a göre daha düşük madde kullanım sıklığına ve daha yüksek kullanılan madde sayısına sahip olduğu sonucuna ulaşılmıştır. Bu sonuç, A-BÇAT'ın madde havuzunu S-BÇAT'a göre daha etkili kullandığını göstermektedir. Bununla birlikte A-BÇAT'ta maddelerin kullanım sıklıkları belirli bir ivmeyle azalırken, S-BÇAT'ta yönlendirme modülündeki maddelerin kullanım sıklıkları yüksek olmakta, ikinci ve üçüncü aşamalarda ise bu sıklıklar azalmaktadır. Mevcut çalışmanın bulgularına benzer biçimde Zheng ve Chang (2015), A-BÇAT'ta test örtüşme oranlarının daha düşük olduğunu ve S-BÇAT'ta benzer rotaları izleyen katılımcıların madde güvenliği açısından sorun oluşturabileceğini belirtmiştir. S-BÇAT'ta panel ve modüller sabit olduğundan, katılımcının izleyeceği rotadaki modüllerde yer alan maddeler test yöneticisi tarafından önceden bilinmektedir. Buna karşılık A-BÇAT, her katılımcı için kendi yetenek düzeyine uygun testi anlık olarak oluşturduğu için, katılımcıya uygulanacak maddeler test yöneticisi dâhil hiç kimse tarafından önceden bilinmemekte veya belirlenememektedir. Bu nedenle A-BÇAT bireye özgü, benzersiz testler oluşturmaktadır. Bu bağlamda A-BÇAT, madde ve test güvenliği açısından S-BÇAT'a göre daha avantajlıdır (Zheng & Chang, 2015).

Tüm test uzunluklarında, A-BÇAT yaklaşımının S-BÇAT'a göre daha yüksek kullanılan madde sayısı ve daha düşük madde kullanım sıklığına sahip olduğu sonucuna ulaşılmıştır. Bu durum, A-BÇAT'ın tüm test uzunluklarında madde güvenliği bakımından S-BÇAT'tan daha verimli sonuçlar ürettiğini göstermektedir.

Farklı yetenek dağılımlarında da A-BÇAT'ın S-BÇAT'a göre daha fazla sayıda madde kullandığı ve daha düşük madde kullanım sıklığına sahip olduğu belirlenmiştir. Dolayısıyla madde güvenliği açısından A-BÇAT'ın S-BÇAT'tan daha iyi sonuçlar verdiği söylenebilir. Ayrıca normal, sağa çarpık ve sola çarpık dağılımlarda iki yaklaşım arasındaki kullanılan madde sayısı ve ortalama madde kullanım sıklığı farkı benzer düzeydeyken, uniform dağılımda bu farkın A-BÇAT lehine arttığı görülmüştür. Bunun nedeni, uniform dağılımda uç yetenek düzeylerindeki katılımcı sayısının artmasıdır. Sonuç olarak uniform dağılımlarda da A-BÇAT yaklaşımının madde güvenliği bakımından S-BÇAT'tan daha etkili olduğu ifade edilebilir.

Tüm modül/test uzunluğu oranlarında, A-BÇAT S-BÇAT'a göre daha yüksek kullanılan madde sayısı ve daha düşük madde kullanım sıklığı göstermektedir. Bu durum, A-BÇAT'ın farklı modül/test uzunluğu oranlarında da madde güvenliği açısından S-BÇAT'tan daha iyi sonuçlar verdiğini göstermektedir. Ayrıca S-BÇAT yaklaşımında B-K-K oranında yönlendirme modülü daha uzun olduğu için, kullanılan madde sayısı daha düşük ve madde kullanım sıklığı daha yüksektir. K-K-B oranında ise üçüncü aşamadaki modüller daha uzun olduğundan, kullanılan madde sayısı daha yüksek ve madde kullanım sıklığı daha düşüktür. Bu durumda K-K-B oranının, madde güvenliği açısından B-K-K oranına göre daha verimli olduğu söylenebilir.

Bu çalışmanın sonuçları, A-BÇAT yaklaşımının hem ölçme kesinliği hem de madde ve test güvenliği açısından S-BÇAT yaklaşımına göre daha avantajlı olduğunu göstermektedir. Bununla birlikte S-BÇAT yaklaşımının A-BÇAT'a göre iki önemli üstünlüğü bulunmaktadır: (1) sabit panel ve modül yapısının test katılımcılarına kolay açıklanabilmesi ve (2) test uygulamasının yönetiminin daha kolay olması (Yan, Von Davier, & Lewis, 2016). Bu iki üstünlük, S-BÇAT'ın tercih edilmesindeki en önemli gerekçeler arasında yer almaktadır. Test yapısının ve puanlama yönteminin test katılımcıları ile diğer ilgili paydaşlar tarafından kolay anlaşılabilmesi, testin adil biçimde yürütüldüğüne ilişkin algıyı güçlendirmesi açısından önemlidir. Bu nedenle uyarlanabilir test yaklaşımının seçiminde, ölçme kesinliği ile madde-test güvenliğinin yanı sıra testin risk

durumu, değerlendirme türü (norm ya da ölçüt bağımlı), test uygulamasının katılımcılara açıklanabilirliği ve toplumun ilgili teste ilişkin sosyolojik bakış açısının da dikkate alınması gerektiği düşünülmektedir.

Sınırlılıklar

Bu çalışmanın, sonuçların yorumlanmasında dikkate alınması gereken bazı sınırlılıkları bulunmaktadır. İlk olarak, analizlerin tamamı simüle edilmiş verilere dayanmaktadır; bu durum gerçek test ortamlarının karmaşıklığını tam olarak yansıtmayabilir. İkinci olarak, çalışma yalnızca iki kategorili puanlanan çoktan seçmeli maddeler bağlamında simüle edilmiştir. Gelecek araştırmalar, çok kategorili ya da karma formatlı testleri de kapsayarak bu sınırlılığı giderebilir. Üçüncü olarak, çalışmada operasyonel test maddeleri yerine varsayımsal madde havuzları kullanılmıştır. Bu havuzlar psikometrik özellikler ve içerik dengesi açısından gerçek test maddelerinden farklılık gösterebilir. Gelecek çalışmalar, gerçek madde havuzları, daha büyük ve daha çeşitli örneklemler ile operasyonel bilgisayarlı test sistemlerinden elde edilen ampirik verileri kullanarak bu araştırmayı genişletebilir.

Öneriler

Bu başlık altında uygulayıcılara ve araştırmacılara yönelik öneriler sunulmuştur.

Uygulayıcılara Yönelik Öneriler

Bu çalışmanın sonuçları doğrultusunda, uygulayıcılara yönelik yedi öneri aşağıda sunulmuştur. (1) Kısa test uzunluklarında daha etkili yetenek kestirimleri sağladığı için A-BÇAT yönteminin S-BÇAT'a tercih edilmesi önerilmektedir. (2) Eğer A-BÇAT tasarımı tercih edilecekse, ölçme kesinliğini artıracağı için son modül uzunluğunun artırılması önerilmektedir. (3) BÇAT tasarımında son aşamadaki madde sayısının yüksek olması isteniyorsa, S-BÇAT yerine A-BÇAT yaklaşımının tercih edilmesi önerilmektedir. (4) Eğer yönlendirme modülünde yer alacak madde sayısının yüksek olması isteniyorsa, S-BÇAT yöntemi ile A-BÇAT yönteminin etkililik düzeyleri birbirine yakın olduğundan S-BÇAT yönteminin de tercih edilebileceği ifade edilebilir. (5) Uygulanacak testin kesme puanı yetenek ölçeğinin orta noktalarına yakınsa, A-BÇAT ve S-BÇAT yöntemleri oldukça benzer ölçme kesinliği sunmaktadır. Ancak kesme puanı uç yetenek düzeylerinde yer alıyorsa, A-BÇAT yönteminin tercih edilmesi önerilmektedir. (6) Test uygulamasında madde ve test güvenliği sorunlarının ortaya çıkabileceği düşünülüyorsa, S-BÇAT yerine A-BÇAT yönteminin tercih edilmesi önerilmektedir. (7) Bu çalışmanın bulguları, ölçme kesinliği ile test güvenliğini dengelemeyi amaçlayan PISA, TIMSS ve ulusal sınavlar gibi geniş ölçekli değerlendirmelerin tasarım ve uygulama süreçlerine katkı sağlayabilir. Önerilen simülasyon çerçevesi, gelecekteki uyarlanabilir test uygulamalarında madde seçimi ve madde kullanım sıklığı kontrol stratejilerinin optimize edilmesinde test geliştiricilere rehberlik edebilir.

Araştırmacılara Yönelik Öneriler

Gelecekte yapılacak çalışmalara yön vermek amacıyla araştırmacılara yönelik on öneri aşağıda sunulmuştur. (1) Bu çalışmada, TIMSS parametre dağılımlarına göre simülasyon ortamında bir madde havuzu üretilmiştir. Benzer bir çalışma gerçek bir veri havuzu kullanılarak tasarlanabilir. (2) Bu araştırmada S-BÇAT panel tasarımı olarak "1-2-3" formatı kullanılmıştır. Benzer çalışmalar "1-3", "1-4", "1-2-4" gibi farklı tasarımlarla yürütülebilir. (3) Bu çalışmada yetenek kestirim yöntemi olarak EAP yöntemi tercih edilmiştir. Yetenek kestirim yöntemi olarak MLE, MLEF ve MAP yöntemlerinin kullanıldığı farklı araştırmalar tasarlanabilir. (4) Bu çalışmada yalnızca dört farklı içerik alanına göre kısıt eklenerek test birleştirme işlemleri gerçekleştirilmiştir. Gerçek madde havuzları kullanılarak ortak kök maddeler, düşman maddeler, sözcük sayısı ve ortalama süre gibi farklı test özellikleri ve kısıtları altında benzer çalışmalar yürütülebilir. Ayrıca kısıt sayısı artırılarak BBT ile A-BÇAT yaklaşımlarının etkililiği karşılaştırılabilir. (5) S-BÇAT'ta yönlendirme yöntemi olarak MFI yöntemi kullanılmıştır. Yönlendirme yöntemi olarak AMI, doğru sayısı ve diğer yöntemlerin kullanıldığı farklı çalışmalar tasarlanabilir. (6) Bu çalışmada A-BÇAT ile S-BÇAT karşılaştırılmıştır. Gelecek araştırmalarda farklı S-BÇAT tasarımları hibrit BBT yaklaşımı ile karşılaştırılabilir. (7) Bu çalışmada test uzunlukları olarak 20, 30 ve 40 dikkate alınmıştır. Farklı test uzunlukları altında benzer çalışmalar

yürütülebilir. Ayrıca BBT yaklaşımları için optimal test uzunluğunu belirlemeye yönelik araştırmalar yapılabilir. (8) Bu çalışmada B-K-K, O-O-O ve K-K-B modül/test uzunluğu oranları dikkate alınmıştır. Farklı modül/test uzunluğu oranları altında yeni çalışmalar gerçekleştirilebilir. (9) Bu çalışmada madde havuzu büyüklüğü 400 olarak belirlenmiştir. Madde havuzu büyüklüğü değiştirilerek benzer araştırmalar tasarlanabilir. (10) Bu çalışmada ölçme kesinliği ve madde güvenliği incelenmiştir. Gelecek araştırmalarda A-BÇAT ile S-BÇAT'ın sınıflama doğruluğu karşılaştırılabilir.

KAYNAKÇA

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology*, *43*(4), 695–716. https://doi.org/10.1111/j.1744-6570.1990.tb00679.x

Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). Altering the level of difficulty in computer adaptive testing. *Applied Measurement in Education*, *5*(2), 137–149. https://doi.org/10.1207/s15324818ame0502_4

Boztunç Öztürk, N. (2019). How the length and characteristics of routing module affect ability estimation in ca-MST? *Universal Journal of Educational Research*, *7*(1), 164–170. https://doi.org/10.13189/ujer.2019.070121

Breithaupt, K. J., Mills, C. N., & Melican, G. J. (2006). Facing the opportunities of the future. *Computer-based testing and the Internet: Issues and advances*, 219-251.

Bulut, O. (2021). Beyond multiple-choice with digital assessments. *ELearn*, *2021*(Special Issue), 1–10. https://doi.org/10.1145/3472394

Bulut, O., & Sünbül, Ö. (2017). R Programlama Dili ile Madde Tepki Kuramında Monte Carlo Simülasyon Çalışmaları. *Egitimde ve Psikolojide Olcme ve Degerlendirme Dergisi, 8*(3), 266–287. https://doi.org/10.21031/epod.305821

Cai, L., Albano, A. D., & Roussos, L. A. (2021). An investigation of item calibration methods in multistage testing. *Measurement: Interdisciplinary Research and Perspectives*, *19*(3), 163–178. https://doi.org/10.1080/15366367.2021.1878778

Carlson, S. (2000). ETS finds flaws in the way online GRE rates some students. *Chronicle of Higher Education, 47*(8), A47.

Cetin-Berber, D. D., Sari, H. I., & Huggins-Manley, A. C. (2019). Imputation methods to deal with missing responses in computerized adaptive multistage testing. *Educational and psychological measurement, 79*(3), 495-511.

Chang, H.-H. (2004). Understanding computerized adaptive testing: From Robbins-Monro to Lord and beyond. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 117-133). Thousand Oaks, CA: Sage.

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. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika, 73*(3), 441–450.

Choi, S. W., & van der Linden, W. J. (2018). Ensuring content validity of patient-reported outcomes: a shadow-test approach to their adaptive measurement. *Quality of Life Research, 27*(7), 1683-1693.

Choi, S. W., Lim, S., & van der Linden, W. J. (2021). TestDesign: an optimal test design approach to constructing fixed and adaptive tests in R. *Behaviormetrika,* 1-39.

Choi, S. W., Moellering, K. T., Li, J., & van der Linden, W. J. (2016). Optimal reassembly of shadow tests in CAT. *Applied psychological measurement*, *40*(7), 469-485. https://doi.org/10.1177/0146621616654597.

Cohen, J. (1988). Statistical power analysis fort he behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. *Educational measurement, 4*, 471-515.

Demir, H., & Gelbal, S. (2025). A systematic review on Computerized Adaptive Testing. *Journal of Education Faculty*, *27*(1), 137–150. https://doi.org/10.17556/erziefd.1577880

Ebenbeck, N., & Gebhardt, M. (2022). Simulating computerized adaptive testing in special education based on inclusive progress monitoring data. *Frontiers in Education, 7*. https://doi.org/10.3389/feduc.2022.945733

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice, 35*(2), 36-49.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*(4), 521-532.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education*. McGram-Hill Publishing.

Gür, R., & Gülleroğlu, H. (2020). The effect of item exposure control methods on measurement precision and test security under different measurement conditions in computerized adaptive testing. *TED EĞİTİM VE BİLİM*, *45*(202), 113–139. https://doi.org/10.15390/eb.2020.8256

Hambleton, R. K., & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass–fail decisions. *Applied Measurement in Education, 19*(3), 221-239.

Han, K. C. T., & Guo, F. (2016). Multistage testing by shaping modules on the fly. In Computerized Multistage Testing (pp. 157-172). Chapman and Hall/CRC.

Han, K. T. (2007). WinGen: Windows software that generates item response theory parameters and item responses. *Applied Psychological Measurement*, *31*(5), 457–459. https://doi.org/10.1177/0146621607299271

Harwell, M., Stone, C. A., Hsu, T. C. & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101-125. doi: 10.1177/014662169602000201

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*(2), 101–125. https://doi.org/10.1177/014662169602000201

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement Issues and Practice*, *26*(2), 44–52. https://doi.org/10.1111/j.1745-3992.2007.00093.x

Khorramdel, L., Pokropek, A., Joo, S. H., Kirsch, I., & Halderman, L. (2020). Examining gender DIF and gender differences in the PISA 2018 reading literacy scale: A partial invariance approach. *Psychological Test and Assessment Modeling, 62*(2), 179-231.

Kim, H., & Plake, B. (1993). Monte Carlo simulation comparison of two-stage testing and computer adaptive testing. Unpublished doctoral dissertation, University of Nebraska, Lincoln.

Kirsch, I., & Lennon, M. L. (2017). PIAAC: a new design for a new era. *Large-scale Assessments in Education, 5*(1), 1-22.

Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a computerized adaptive test more motivating than a fixed-item test? *Applied Psychological Measurement*, *41*(7), 495–511. https://doi.org/10.1177/0146621617707556

Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika, 36*(3), 227-242. https://doi.org/10.1007/BF02297844

Luo, X., & Kim, D. (2018). A top-down approach to designing the computerized adaptive multistage test. *Journal of Educational Measurement, 55*(2), 243-263.

Magis, D., Yan, D., & Von Davier, A. A. (2017). Computerized adaptive and multistage testing with R: Using packages catr and mstr. Springer.

Makhorin A (2017). GNU Linear Programming Kit. Version 4.61, URL http://www.gnu. org/software/glpk/glpk.html.

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology*, *110*(1), 27–45. https://doi.org/10.1037/edu0000205

Mead, A. D. (2006). An introduction to multistage testing. *Applied Measurement in Education*, *19*(3), 185–187. https://doi.org/10.1207/s15324818ame1903_1

MEB (2021). 2021 Ortaöğretim Kurumlarına İlişkin Merkezi Sınav Raporu. Milli Eğitim Bakanlığı.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine, 38*(11), 2074-2102.

OECD (2023). *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*, OECD Publishing, Paris, https://doi.org/10.1787/53f23881-en.

OECD (2024), *PISA 2022 Technical Report,* PISA, OECD Publishing, Paris, https://doi.org/10.1787/01820d6d-en.

Ortner, T. M., Weißkopf, E., & Koch, T. (2014). I will probably fail: Higher ability students' motivational experiences during adaptive achievement testing. *European Journal of Psychological Assessment: Official Organ of the European Association of Psychological Assessment*, *30*(1), 48–56. https://doi.org/10.1027/1015-5759/a000168

Patsula, L. N., & Hambleton, R. K. (1999). A comparative study of ability estimates obtained from computer-adaptive and multi-stage testing. In annual meeting of the National Council on Measurement in Education, Montreal, Quebec.

Pine, S. M., Church, A. T., Gialluca, K. A., & Weiss, D. J. (1979). *Effects of Computerized Adaptive Testing on Black and White Students*. Minnesota Univ Minneapolis Dept Of Psychology.

Saatçioğlu, F. M., & Atar, H. Y. (2022). Investigation of the effect of parameter estimation and classification accuracy in mixture IRT models under different conditions. *International Journal of Assessment Tools in Education*, *9*(4), 1013–1029. https://doi.org/10.21449/ijate.1164590

Stark, S., & Chernyshenko, O. S. (2006). Multistage testing: Widely or narrowly applicable?. *Applied Measurement in Education, 19*(3), 257-260.

Şahin, A., & Weiss, D. J. (2015). Effects of calibration sample size and item bank size on ability estimation in computerized adaptive testing. *Educational Sciences Theory & Practice, 15*(6). https://doi.org/10.12738/estp.2015.6.0102

Tay, P. H. (2015). On-the-fly assembled multistage adaptive testing. University of Illinois at Urbana-Champaign.

Tomashev, M. V., Avdeev, A. S., & Krasnova, M. V. (2018). Adaptive testing as a tool for managing quality of education. *Informatics and Education, 9*, 27–33. https://doi.org/10.32517/0234-0453-2018-33-9-27-33

van der Linden, W. J. (2009). Constrained adaptive testing with shadow tests. In Elements of adaptive testing (pp. 31-55). Springer, New York, NY.

van der Linden, W. J. (2010). *Elements of adaptive testing*. C. A. Glas (Ed.). New York, NY: Springer.

van der Linden, W. J. (2018). Optimal test design. Handbook of item response theory: Vol. 3. Applications, 167-195.

van der Linden, W. J. (2021). Review of the shadow-test approach to adaptive testing. Behaviormetrika, 1-22.

van der Linden, W. J., & Diao, Q. (2016). *Using a universal shadow-test assembler with multistage testing*. Computerized multistage testing: Theory and applications, 101-118.

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics: A Quarterly Publication Sponsored by the American Educational Research Association and the American Statistical Association*, *29*(3), 273–291. https://doi.org/10.3102/10769986029003273

van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, *44*(2), 117–130. https://doi.org/10.1111/j.1745-3984.2007.00030.x

Xu, L., Jiang, Z., Han, Y., Liang, H., & Ouyang, J. (2023). Developing computerized Adaptive Testing for a national health professionals exam: An attempt from psychometric simulations. *Perspectives on Medical Education*, *12*(1), 462–471. https://doi.org/10.5334/pme.855

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. *Educational Measurement Issues and Practice*, *37*(4), 16–27. https://doi.org/10.1111/emip.12226

Yan, D., Von Davier, A. A., & Lewis, C. (Eds.). (2016). Computerized multistage testing: Theory and applications. CRC Press.

Yasuda, J. I., Mae, N., Hull, M. M., & Taniguchi, M. A. (2021). Optimizing the length of computerized adaptive testing for the force concept inventory. *Physical review physics education research, 17*(1), 1-15.

Yasuda, J.-I., Mae, N., Hull, M. M., & Taniguchi, M.-A. (2021). Optimizing the length of computerized adaptive testing for the Force Concept Inventory. *Physical Review Physics Education Research*, *17*(1). https://doi.org/10.1103/physrevphyseducres.17.010115

Yigiter, M. S., & Dogan, N. (2023). Computerized multistage testing: Principles, designs and practices with R. *Measurement: Interdisciplinary Research and Perspectives*, *21*(4), 254–277. https://doi.org/10.1080/15366367.2022.2158017

Yiğiter, M. S., & Boduroğlu, E. (2024). Item Response Theory assumptions: A comprehensive review of studies with document analysis. *International Journal of Educational Studies and Policy, 5*(2), 119-138. https://doi.org/10.5281/ZENODO.14016086

Yiğiter, M. S., & Doğan, N. (2023). The effect of test design on misrouting in computerized multistage testing. *International Journal of Turkish Education Sciences*, *2023*(21), 549–587. https://doi.org/10.46778/goputeb.1267319

Zheng, W. (2016). Making test batteries adaptive by using multistage testing techniques (Doctoral dissertation, University of North Carolina, Greensboro, NC).

Zheng, Y., & Chang, H.-H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*, *39*(2), 104–118. https://doi.org/10.1177/0146621614544519

Trakya Journal of Education

ISSN : 2630-6301
Publisher : Trakya University

Vol. 16, No. 2, 794-820, 2026
DOI: 10.24315/tred.1665684

Research Article

Comparative Study of Fixed and On-the-Fly Computerized Multistage Testing: Implications for Measurement Accuracy and Item Security

Mahmut Sami Yiğiter[1*,**]
Nuri Doğan[2]

[1]Social Sciences University of Ankara, Ankara, Türkiye
mahmutsamiyigiter@gmail.com

[2] Hacettepe University, Educational Sciences Department, Ankara, Türkiye
nuridogan2004@gmail.com

[*] Corresponding Author

[**] This study was produced from the doctoral dissertation conducted by the first author under the supervision of the second author.

Received: 25.03.2025
Accepted: 27.10.2025
Available Online: 30.04.2026

Abstract: In recent years, adaptive testing techniques such as Computerized Adaptive Testing (CAT) and Computerized Multistage Testing (MST) have been increasingly incorporated into large-scale evaluations. This study aims to compare Fixed-MST (F-MST) and On-the-Fly MST (O-MST), a novel approach in which items are grouped into modules based on the participant's ability level, in terms of measurement precision and item security across various simulation scenarios. The simulations were carried out using item parameter distributions derived from the 3PL model applied in TIMSS. A total of 72 different conditions were analyzed to compare O-MST with F-MST. The findings on measurement precision reveal that O-MST performs better than F-MST, especially when the test lengths are shorter, where O-MST shows substantially higher measurement precision. Moreover, when examining ability distributions, O-MST demonstrates better measurement precision compared to F-MST, particularly in cases of non-normal distributions. A significant result from this study is that the measurement precision of O-MST improves as the length of the final module increases, whereas the measurement precision of F-MST becomes more similar to O-MST as the length of the initial module increases. Regarding item security, O-MST employed a greater number of items and exhibited a lower item exposure rate compared to F-MST in all conditions. The favorable results in terms of measurement precision and item security for O-MST are discussed within the framework of large-scale assessments and relevant literature.

Keywords: Computerized Multistage Testing, Adaptive Testing, Item Security, Item Exposure Rate.

INTRODUCTION

For centuries, Linear Tests (LT) have been the popular method for assessing the knowledge, skills, and abilities of test takers in educational evaluations. However, in the past fifty years, with advancements in computer technology, Computerized Adaptive Testing (CAT) has undergone significant development and has become increasingly popular. CAT has been widely adopted in numerous national and international assessments, particularly due to its ability to provide accurate ability estimates and reduce test length (Khorramdel et al., 2020; Kirsch & Lennon, 2017; Demir & Gelbal, 2025). In contrast, Computerized Multistage Testing (MST) has gained attention for its recent innovations and has secured a place in large-scale assessments (van der Linden, 2018).

MST design allows the difficulty level of the test to be adjusted according to the test taker's ability during the assessment. MST can be considered a hybrid of Computerized Adaptive Testing (CAT) and Linear Testing (LT), as it incorporates features from both models. While CAT and MST share similarities, particularly in how the administration of test items is based on the test taker's performance on preceding items, the key distinction lies in the sequencing process. In CAT, the algorithm adjusts after every item, continuously estimating the test taker's ability. In contrast, MST estimates the participant's ability only after a series of items, referred to as modules, have been completed. More specifically, the test taker first receives a set of items known as the routing module. Based on the test taker's performance in this module, their ability is estimated and compared to a predetermined criterion score (cut-off score). If the individual's ability exceeds the criterion, a more difficult module (set of challenging items) is administered; if the ability falls below the criterion, an easier module (set of simpler items) is provided. In terms of flexibility and complexity, MST can be seen as a balance between LT and CAT. Compared to LT, MST offers more efficient and accurate measurements along the ability scale, leading to improved precision in assessment and reduced test length. As a result, MST is generally considered more effective than LT regarding measurement accuracy and test efficiency (Lord, 1971; van der Linden, 2010).

Cite as (APA 7): Yiğiter, M. S. & Doğan, N. (2026). Comparative study of fixed and on-the-fly computerized multistage testing: implications for measurement accuracy and item security. *Trakya Journal of Education 16*(2), 794-820. https://doi.org/10.24315/tred.1665684

Since the modules used in MST are designed and assembled before the test is administered and are presented as a unit to the test taker, it provides test developers with greater control over aspects such as test coverage, the overall quality of the test structure, and the management of the testing process itself (van der Linden, 2010). Additionally, unlike CAT, MST allows test takers to modify their responses within each module, skip items, revisit previous questions, and provide new answers (Chang, 2015). Hambleton and Xing (2006) conducted a comparison between MST, CAT, and LT in the context of making pass-fail decisions. Their findings indicated that CAT outperformed MST slightly in terms of measurement precision. However, the researchers concluded that MST could be more effectively utilized by test developers if it involved modules with a broader difficulty range and larger item banks that were more comprehensive in content.

In recent years, MST has been increasingly adopted in numerous large-scale assessments due to its advantages. For instance, the CPA (Certified Public Accountants) examination has been utilizing MST since 2004 (Breithaupt, Mills & Melican, 2006). In 2011, the GRE (Graduate Record Examinations) transitioned to an MST-based format. That same year, the OECD introduced an adaptive international large-scale assessment using the MST design as part of the PIAAC (Program for the International Assessment of Adult Competencies) (Kirsch & Lennon, 2017). Following PIAAC, the 2018 round of the PISA (Programme for International Student Assessment) employed MST in reading, one of the three main assessment domains (Khorramdel et al., 2020). In PISA 2022, the MST design was used not only in reading but also in mathematics literacy (OECD, 2023).

MST has the potential to enhance test takers' engagement and motivation by offering items tailored to their individual ability levels throughout the test. According to PISA, computer-based assessments (CBA) exhibit lower non-response rates compared to paper-based assessments (PBA) (OECD, 2024). As a result, it is anticipated that MST can help mitigate non-response and random answering behaviors in assessments (Yamamoto et al., 2018). Numerous studies have explored the impact of adaptive tests on motivation (Arvey et al., 1990; Bergstrom et al., 1992; Ortner et al., 2013; Pine et al., 1979). Ling et al. (2017) found that adaptive testing led to greater engagement and lower anxiety compared to fixed-item tests. Furthermore, Martin and Lazendic (2018) reported that MST provides more precise measurement than computer-based testing and produces positive outcomes in terms of motivation, engagement, and overall test experience.

Recently, there has been a growing interest in interactive assessments that aim to measure cognitive skills in settings that closely resemble real-world scenarios or as accurately as possible (Bulut, 2021). Scenario-based item groups create rich environments where students engage by modifying the item, providing a more dynamic and authentic assessment experience. Consequently, there is a need for more integrative questions that consist of multiple items to effectively assess real-life situations and scenario-based tasks. For instance, completing complex tasks such as writing recommendations based on a text or addressing a series of problems in a real-life context would require the presentation of multiple related items rather than a single question. In such cases, it is necessary to assess the test taker's performance through a group of interconnected items. In CAT applications, ability estimation is performed on an item-by-item basis. However, as mentioned earlier, splitting a set of items, which are linked by a common scenario or theme, into individual items distorts the measurement of the construct. Therefore, CAT is not suitable for this type of scenario. In contrast, in PISA, around 30% of the items are constructed-response items that are scored by human coders. Since the scoring of these items is not immediate, it may not be appropriate to apply CAT to these items. The MST design, as a module-level adaptive test, appears to be a more fitting approach in this context. Well-designed groups of items (modules) that align with the framework of the construct being measured are likely to provide more accurate assessments than a large number of independent items based on factual knowledge (Yamamoto et al., 2018, Yigiter and Dogan, 2023).

One of the key limitations of CAT is that it may either underestimate or overestimate a test taker's ability level (Chang & Ying, 2008). This occurs because the ability estimate is updated after each item is answered using the Maximum Fisher Information (MFI) method. For instance, if a participant with a high ability level answers the initial items incorrectly due to test anxiety, lack of motivation, or simple errors, it becomes challenging to accurately assess their true ability based on subsequent responses. Likewise, if a participant with a low ability level answers early items correctly by chance or due to prior knowledge, the system may overestimate their ability level. In contrast, MST estimates the ability level after completing each stage, which helps to mitigate the limitations of CAT. This advantage was highlighted when several testing organizations identified issues with CAT in large-scale assessments. For example, in 2000, the Educational Testing Service

(ETS) found that the computerized Graduate Record Examinations (GRE-CAT) was inaccurately predicting scores for thousands of test takers and offered affected individuals the opportunity to retake the test at no cost (Carlson, 2000). A similar problem occurred in 2002 with the Graduate Management Admission Test (GMAT), where approximately one thousand candidates received incorrect scores (Chang, 2004). Following these incidents, MST gained traction as a solution to address the shortcomings of CAT. As a result, many testing organizations shifted from CAT to MST (Hendrickson, 2007).

Fixed Computerized Multistage Testing (F-MST)

F-MST is an algorithm-driven testing method where pre-assembled groups of items (modules) are selected by the algorithm and administered to test takers in stages. The F-MST approach is known by various terms in the literature, with "fixed" commonly used, though "standard" and "conventional" are also used to describe this method. For clarity, this study uses the term "fixed." Figure 1 illustrates the F-MST design in a 1-2-3 format. In F-MST, each pre-assembled group of items is referred to as a module, which serves as the fundamental unit of the testing design. During the first stage, a module, typically called the routing module, is presented at a medium difficulty level. The test progresses by directing the participant through successive modules based on their ability level, with each module adapting to the participant's performance.

Figure 1

1-2-3 F-MST Desing

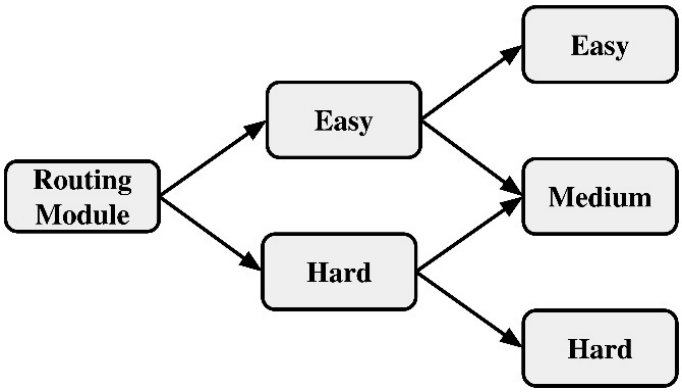


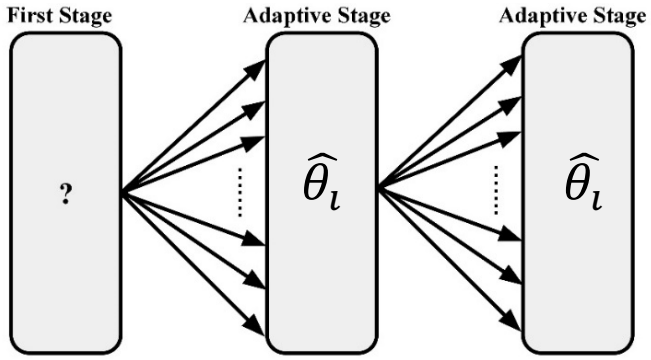
Figure 1 presents an example of a 1-2-3 F-MST design, which consists of one module in the first stage, two modules in the second stage, and three modules in the third stage. The difficulty levels of the modules in this design are typically categorized as easy, medium, or difficult. In the initial stage of the F-MST, all test takers begin with the first module, known as the routing module, which is typically designed with moderate item difficulty. Based on their performance in the routing module, participants are then directed to either an easy or difficult module in the second stage. In the second stage, depending on their responses, participants are assigned to one of the easy, medium, or difficult modules in the third stage. Finally, after the third stage, the test concludes, and the participant's final ability level is estimated. The structure of modules and stages depicted in Figure 1 is referred to as a panel, and multiple panels can be created in test implementations.

On-the-fly Computerized Multistage Testing (O-MST)

Similar to the F-MST, the On-the-Fly Computerized Multistage Testing (O-MST) is administered in phases. However, unlike F-MST, O-MST assembles modules "on the fly" during the test based on the individual's ability level (Tay, 2015). In O-MST, the first stage typically involves creating a full-length test with items of moderate difficulty. The initial module-length portion of the test is administered to the test taker, and an interim ability estimate is calculated based on their performance. Subsequently, the test is re-assembled according to the participant's interim ability estimate, and the second module-length portion of the test is administered. This means that each test taker receives a different set of items in the second and third phases, which are tailored according to their interim ability parameter after each stage. In essence, O-MST offers a personalized test experience for each participant. This adaptive process continues until the test is completed. Figure 2 illustrates the general framework of shadow tests and the implementation of O-MST.

Figure 2

O-MST General Framework



Note. $\widehat{\theta}_l$*: interim ability level.*

Figure 2 illustrates the three-stage O-MST application. In the first stage, the test taker is presented with a module of medium difficulty. In the second and subsequent stages, a combination of modules is administered based on the test taker's interim ability level. Although a three-stage O-MST design is depicted in Figure 2, the number of modules and the length of each module can be adjusted as needed in the O-MST system. Overall, O-MST is highly flexible, allowing for variations in the number of items within each module and the number of stages (Zheng & Chang, 2015).

Item Security

Within computerized adaptive testing (CAT) and multistage testing (MST), item security refers to how effectively the items in an item pool are protected from being overused or disclosed. Preserving item security is a crucial element of maintaining the fairness, validity, and confidentiality of an assessment. When certain items are administered too frequently, they risk becoming known to future participants, which can undermine the credibility of the testing program.

Item security is often assessed through the item exposure rate, indicating the percentage of examinees who receive a specific item across test administrations. High exposure rates suggest that a limited subset of items is being selected repeatedly, while others remain unused. Such uneven exposure patterns can lead to several undesirable outcomes: widely exposed items may circulate among participants, creating an unfair advantage for those familiar with them; the psychometric quality of the test may decline due to biased ability estimates; and the operational life span of the item pool may shorten, requiring continuous development of new items—a costly and labor-intensive process.

A secure testing system, by contrast, achieves balanced exposure where items are utilized efficiently yet not excessively. This balance enhances both test precision and the longevity of the item bank. In this study, item security was evaluated using the average item exposure rate, following the conventions established in adaptive testing research (van der Linden & Veldkamp, 2004; Zheng & Chang, 2015). The On-the-Fly Multistage Testing (O-MST) framework dynamically assembles modules based on updated ability estimates, which promotes greater diversity in item use and reduces the likelihood of overexposure. As a result, O-MST provides stronger item protection compared with the Fixed MST (F-MST) design, where predefined modules may lead to the repetitive use of certain items across participants.

Literature Review

The literature contains various studies comparing F-MST and O-MST. Choi et al. (2016) compared CAT, O-MST, and Hybrid-CAT, a combination of the two methods, using the freeze-fresh mechanism for test combination. The results are quite similar in all three approaches. The authors found that the freeze-fresh mechanism worked particularly well when there were common-root items and test constraints had to be met. Furthermore, there was no significant decrease in the measurement accuracy of the test. A novel MST design that combines modules on-the-fly based on intermediate ability predictions at each stage was presented by

Zheng and Chang (2015), in contrast to the F-MST where modules are assembled in advance. Their results showed that O-MST and CAT provide comparable measurement accuracy and both methods offer better measurement accuracy and test security than F-MST.

Tay (2015) compared F-MST and O-MST regarding classification accuracy and consistency across test lengths of 12, 18, 24, and 30 items. The results indicated that O-MST provided higher classification accuracy and consistency than F-MST in all conditions. van der Linden and Diao (2016) compared five different test approachs using a real dataset through simulations. The study found that LT was the least efficiency, followed by F-MST. The other approaches showed similar efficiency levels.

Han and Guo (2016) proposed an O-MST design that assembles new modules on-the-fly at each stage based on a predicted ability level (θ) of the test taker. The merged module is iteratively reassembled until it reaches the desired test information function (TIF) specified by the test developer. The study compared CAT, F-MST, and O-MST design. Results showed 1-3-3 MST similar measurement accuracy to the newly developed O-MST design with low iteration shaping. However, as the number of iterations increased to 100, the new O-MST design outperformed F-MST in terms of measurement accuracy. CAT, however, produced better accuracy than both methods. van der Linden (2021) compared on-the-fly LT, F-MST, O-MST, and CAT using an item pool of 300 items derived from a real dataset. In the simulation study, both O-MST and CAT achieved similar and very high measurement precision. In contrast, F-MST, although more accurate than the on-the-fly LT, produced significantly lower measurement accuracy than both O-MST and CAT.

Finally, van der Linden (2021) compared on-the-fly LT, F-MST, O-MST, and CAT using a 300-item pool derived from a real dataset. The simulation results showed that both O-MST and CAT achieved similar and excellent measurement precision. F-MST, while more accurate than on-the-fly LT, produced significantly lower measurement accuracy than both O-MST and CAT.

Aim and Importance of the Research

Several studies in the literature have compared F-MST and O-MST. Choi et al. (2016) compared CAT, O-MST, and Hybrid-CAT, which combines both methods, using the freeze-fresh mechanism in test combination. The results indicated that all three approaches produced comparable outcomes. The study concluded that the freeze-fresh mechanism is effective, as it does not lead to a significant decrease in measurement accuracy, particularly when common-root items are used and test constraints need to be adhered to. Zheng and Chang (2015) introduced an innovative MST design that merges modules on-the-fly based on interim ability estimations at each stage, unlike F-MST, where modules are pre-assembled. The study's findings revealed that the measurement accuracy of O-MST and CAT were comparable, with both methods offering better measurement accuracy and test security compared to F-MST.

Research Questions

Two research questions and two research problems and six sub-research problems based on these two research problems are given below.

Q1. To what extent does the measurement accuracy of F-MST and O-MST approaches change under different simulation conditions?

Q1.1. How do the RMSE, MAB and BIAS values of the F-MST and O-MST approaches change according to the test length (20-30-40)?

Q1.2. How do the RMSE, MAB and BIAS values of F-MST and O-MST approaches vary according to ability distribution (normal distribution, right skewed, left skewed, uniform)?

Q1.3. How do the RMSE, MAB and BIAS values of the F-MST and O-MST approaches vary according to the module/test length ratio (L-S-S, M-M-M, S-S-L)?

Q2. How does the item security of the F-MST and O-MST approaches change under different simulation conditions?

Q2.1. How does the item security of the F-MST and O-MST approaches change according to the frequency of item use and the number of items used at different test lengths (20-30-40)?

Q2.2. How does the item security of the F-MST and O-MST approaches vary according to the frequency of item use and the number of items used in different ability distributions (normal distribution, right-skewed, left-skewed, left-skewed, uniform)?

Q2.3. How does the item security of the F-MST and O-MST approaches change according to the frequency of item use and the number of items used in different module/test length ratios (L-S-S, M-M-M, S-S-L)?

METHOD

Type of Research

The aim of this study was to examine the effectiveness of various MST approaches under different simulation conditions. The data used in the study were generated through simulation, with comparisons made across different scenarios. Simulation studies are computer-based experiments that generate data by random sampling from specified probability distributions and then analyze the resulting data. These studies are instrumental in assessing the performance of statistical methods under controlled conditions. In psychometrics, simulation studies play a crucial role in evaluating both new and alternative methodologies (Feinberg & Rubright, 2016; Morris, White & Crowther, 2019; Saatçioğlu & Atar, 2022). This research is a Monte Carlo simulation study, where data are simulated according to relevant probability distributions and simulation conditions. Monte Carlo studies, which share similarities with experimental studies, use computer-generated data (Harwell et al., 1996). Given the complexity of addressing all the variables considered in the study using real data, simulated data were utilized. As the study seeks to compare different MST methods, it can be classified as descriptive research, as it aims to identify which method produces the most suitable results (Fraenkel, Wallen & Hyun, 2012).

Data Generation

The data was generated using R, an open-source and free statistical programming language. In generating the data, firstly, item pool parameters consisting of 400 items and ability parameters consisting of 1000 individuals were generated (Şahin and Weiss, 2015). "Rmst" (Luo & Kim, 2018) and "TestDesign" (Choi, Lim & van der Linden, 2021) packages were used for test combination. Then, with the codes written by the researchers, F-MST analyses were conducted with the combined tests using the "mstR" (Magis, Yan & Von Davier, 2017) package, and O-MST analyses were conducted with the "TestDesign" (Choi, Lim & van der Linden, 2021) package using the same item pool and ability parameters.

Item Pool Generation

Within the scope of the study, an item pool consisting of 400 items was produced based on the 3-parameter logistic model (3PL) (Yiğiter and Boduroğlu, 2024). In the generation of the item pool, the distributions of the a, b and c parameters of the items used in the 2003, 2007, 2011, 2015 and 2019 TIMSS 8th grade mathematics applications and whose parameters were estimated based on the 3PL were analyzed. Considering the distributions of the parameters of these items under the minimum and maximum values, skewness and kurtosis coefficients; a parameters $a \sim lnN(0.2, 0.3)$ from log-normal distribution, b parameters $b \sim N(0, 0.7)$ from normal distribution, c parameters $c \sim Beta(5, 16)$ from beta distribution. The descriptive statistics of the parameters of the 400 items in the item pool are presented in Table 1.

Table 1

Descriptive Statistics of Item Parameters

Parameter	K	Min	Max	Mean	Sd

A	400	0.566	2.287	1.243	0.351
B	400	-1.830	1.764	0.000	0.727
C	400	0.044	0.501	0.236	0.088

In addition, assuming that the item pool consists of four different content areas, all items were randomly assigned to Content 1 (30% - 120 items), Content 2 (30% - 120 items), Content 3 (20% - 80 items) and Content 4 (20% - 80 items).

Ability Distributions Generation

In the study, four different normal, right skewed, left skewed and uniform ability distributions were used. The number of participants in all ability distributions was set as N=1000. Right skewed and left skewed ability distributions were obtained from the normal distribution with the power method proposed by Fleishman (1978). Fleishman's power method equation is presented in the equation below:

$$Y = a + bX + cX^2 + dX^3 \quad (1)$$

Where a, b, c, and d are the Fleishman transformation coefficients that define the shape of the generated ability distributions (Fleishman, 1978). These coefficients determine to the simulated data and are computed to produce the desired non-normal distributions used in the study. X refers to the parameters obtained with the normal distribution used. The X distributions and a, b, c, d distributions used to generate the normal, right skewed and left skewed ability distributions are given in Table 2. The uniform ability distribution, U~U(-3, +3), was obtained from the uniform distribution.

R programming language was used to generate the data. R is an open source and free statistical programming language. In generating the data, firstly, item pool parameters consisting of 400 items and ability parameters consisting of 1000 individuals were generated. “Rmst” (Luo & Kim, 2018) and “TestDesign” (Choi, Lim & van der Linden, 2021) packages were used for test combination. Then, with the codes written by the researchers, F-MST analyses were conducted with the combined tests using the “mstR” (Magis, Yan & Von Davier, 2017) package, and O-MST analyses were conducted with the “TestDesign” (Choi, Lim & van der Linden, 2021) package using the same item pool and ability parameters.

Table 2

Ability Distributions Generation

Distribution	N	X	Skewness	Kurtosis	a	b	c	d
Normal	1000	N(0, 1)	0.00	0.00	0.00	1.00	0.00	0.00
Right Skewed	1000	N(0, 1)	1.50	4.00	-0.21	0.85	0.21	0.04
Left Skewed	1000	N(0, 1)	-1.50	4.00	0.21	0.85	-0.21	0.04
Uniform	1000	U(-3, +3)	-	-	-	-	-	-

Research Design

In this study, different MST approaches (F-MST and O-MST) were compared under different simulation conditions. The independent and dependent variables used in the simulation are summarised in Table 3.

Table 3

Independent and Dependent Variables of the Study

Variable Type	Variable Name	Description
Independent Variable	MST Type	Fixed-MST (F-MST) and On-the-Fly MST (O-MST)
Independent Variable	Test Length	20, 30, 40 items
Independent Variable	Ability Distribution	Normal, Right Skewed, Left Skewed, Uniform
Independent Variable	Module/Test Length Ratio	L-S-S, M-M-M, S-S-L
Dependent Variable	Measurement Accuracy	RMSE, MAB, and BIAS values
Dependent Variable	Item Security	Number of items used and item exposure rate

The conditions to be manipulated in the simulation are presented in Table 4.

Table 4

Manipulated Conditions

Manipulated Variable	Level	Number of Levels
MST Type Approach	F-MST, O-MST	2
Test Length	20, 30, 40	3
Ability Distribution	Normal, Right Skewed, Left Skewed and Uniform	4
Module/Test Length Ratio	L-S-S [1/2-1/4-1/4], M-M-M [1/3-1/3-1/3], S-S-L [1/4-1/4-1/2]	3

Note. S: Short, M=Medium, L=Long.

As can be seen in Table 4, simulations were performed by varying two different MST types (F-MST, O-MST), three different test lengths (20, 30, 40), four different ability distributions (normal, right skewed, left skewed and uniform) and three different module/test length ratio distributions (L-S-S, M-M-M, S-S-L). All conditions were crossed with each other. Therefore, 2x3x4x3=72 conditions were examined in the first simulation study and 100 replications were performed for each condition. Consistent with previous simulation studies in educational measurement and psychometrics, this study employed 100 replications per condition. Prior research has shown that around 100 replications provide sufficiently stable estimates of bias, RMSE, and item exposure indices while maintaining computational efficiency (Bulut & Sünbül, 2017; Gür & Gülleroğlu, 2020; Xu et al., 2023). Using 100 replications has also been recommended to minimize sampling bias and ensure empirical stability in complex IRT-based adaptive testing contexts (Han, 2007; Harwell et al., 1996).

Creating MST Designs

In the study, 1-2-3 F-MST design was preferred as the F-MST design. “Rmst” and “TestDesign” packages were used to create 1-2-3 F-MST and O-MST designs, respectively. In this study, 1-2-3 MST design was preferred. The reason for choosing this design is that the points where the module information functions of the modules in the second stage of the 1-2-3 MST design are maximized do not overlap with the points where the module information functions of the other stages are maximized. In other words, the overlap of the points that maximize the module information function of the modules in the second and third stages in the 1-3-3 MST design has a decreasing effect on the cumulative information function of the modules. Cetin-Berber, Sari, and Huggins-Manley (2018) report that 1-2-3 and 1-3-3 MST designs have very similar measurement

accuracy results. Similarly, Yiğiter and Doğan (2023), in their study examining 1-3, 1-2-3, and 1-3-3 MST designs, state that 1-2-3 and 1-3-3 MST designs have very similar measurement accuracy, and even 1-2-3 MST design offers better measurement accuracy with a slight difference. Therefore, the 1-2-3 MST design was preferred to utilize the item pool more effectively.

The module/test length ratio was taken into account in the creation of the modules. For example, for a test length of 40 items and module/test length ratios of S-S-L (1/4-1/4-1/2), the ATA process was performed with module lengths of 10-10-20. For O-MST, in order to ensure these ratios, the new module will be combined immediately after these items and presented to the test taker. For example, for an O-MST design with module lengths of 10-10-20, the new module was merged and presented to the test taker with the freeze-fresh mechanism (Choi et al., 2021), taking into account the constraints in item positions 1-11-21. The freeze-refresh mechanism (Choi & van der Linden, 2018) is a dynamic module assembly strategy used in On-the-Fly Multistage Testing (O-MST). In this approach, after each stage, items or modules that have already been administered are frozen (kept fixed), while the remaining unadministered items are refreshed (reselected) based on the updated ability estimate. This process ensures that all test constraints (such as content balance or item exposure limits) remain satisfied while maintaining adaptivity throughout the test. The number of items to be included in the modules in 1-2-3 MST and O-MST designs and the item exposure control method are presented in Table 5.

Table 5

Module Item Counts and Item Exposure Rate Control by Test Length

Design		Test Length			Item Exposure Rate	Number of Panels
		20	30	40		
		Module Length Distribution				
1-2-3 F-MST	L-S-S	10-5-5	15-8-7	20-10-10	0.33	3
	M-M-M	3-4-3	6-7-6	13-14-13		
	S-S-L	5-5-10	7-8-15	10-10-20		
O-MST	L-S-S	10-5-5	15-8-7	20-10-10	0.33	*
	M-M-M	3-4-3	6-7-6	13-14-13		
	S-S-L	5-5-10	7-8-15	10-10-20		

Note. *In the O-MST method, the item exposure rate was controlled by fixing it at 0.33 with the Ineligibility method (van der Linden & Weldkamp, 2004).

In the F-MST approach, the number of panels was used to control the item exposure rate. For each F-MST design, 3 panels were formed for a substance exposure rate of 0.33. In the O-MST design, since the modules were created on the fly, the ineligibility method (van der Linden & Weldkamp, 2004) was used to control the item exposure rate by fixing it at 0.33. The ineligibility method is an item exposure control technique designed to prevent overuse of certain items. After an item is administered, it becomes temporarily ineligible for selection until its exposure rate falls below a specified threshold. This rotation mechanism helps maintain test security and ensures a fair distribution of item usage across examinees (van der Linden & Weldkamp, 2004).

The points where the module information was maximized for the creation of panels and modules with test assembly are presented in Table 5.

Table 5

Assembling MST Panels and Modules

MST Approach	Stage 1	Stage 2	Stage 3
1-2-3 F-MST	$\vartheta = 0$	$\vartheta = (-0.5, 0.5)$	$\vartheta = (-1, 0, 1)$

O-MST	$\vartheta = 0$	$\vartheta = \vartheta*$	$\vartheta = \vartheta*$

Note. ϑ* interim ability level.

In both MST designs, the tests were combined so that the initial ability level $\vartheta = 0$. In the F-MST approach, with the 1-2-3 design, the modules of the second and third stages were created to reach the maximum information value at the ability levels given in Table 3. In the test assembly process of the 1-2-3 F-MST design, the hybrid method (the method in which the modules are created from the items with the bottom-up method and the test design is created by assembling the modules with the top-down method) was utilized. In the hybrid test assembly approach, test construction proceeds in two stages: (1) At the bottom-up level, items are combined to create parallel modules that meet information and content constraints. (2) At the top-down level, these pre-assembled modules are then combined into panels to form the overall multistage structure (Luecht & Nungester, 1998; Breithaupt & Hare, 2007). This hybrid strategy benefits from both approaches—bottom-up ensures item-level parallelism, while top-down provides control over higher-level constraints across modules and panels.So, this study was used to hybrid method.

Similarly, in O-MST, the first module was created at $\vartheta = 0$ and the other modules were assembled and presented to the participant with the shadow test approach according to the point of the estimated temporary ability level. In both F-MST and O-MST approaches, the Rglpk (Makhorin, 2017) algorithm was used for test combining. In O-MST, test coupling is performed under the Shadow Test approach using the Freeze-Refresh Mechanism (Choi & van der Linden, 2018).

Data Analysis

The estimated and true ability parameters, Root Mean Square Error (RMSE), Mean Absolute Bias (MAB) and Bias (BIAS) values were used to evaluate the results obtained from the data analysis. RMSE values were calculated with the formula given below, where n is the total number of participants, n is the estimated ability level and n is the true ability level.

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

The formula used to calculate the MAB value is given below.

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

The formula used to calculate the BIAS value is given below.

$$\text{BIAS} = \frac{\sum_{i=1}^{n}\widehat{\theta_i}-\theta_i}{n} \quad (4)$$

In the above formulas, where i represents a participant's number, n represents the number of participants, θ_i represents the true ability of participant *i*, and $\widehat{\theta_i}$ represents the estimated ability of participant obtained from the test.

In addition, effect size values were calculated with the RMSE and MAB values obtained for the comparison of two different test designs under different simulation conditions. The following formulas were used to calculate the Cohen's d value and effect size:

$$Harmonized\ Standard\ Deviation = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2}} \quad (5)$$

$$Cohen\ d = \frac{Mean\ Difference}{Harmonized\ Standard\ Deviation} \quad (6)$$

Cohen's d value calculated with the above formulas is interpreted as small effect if d < 0.20, medium effect if 0.20 < d < 0.50, and large effect if 0.80 < d (Cohen, 1988).

In the examination of item security, the item exposure rate was calculated for each item. The item exposure rate was calculated with the following formula.

$$Item\ Exposure\ Rate = \frac{n_M}{n_T} \quad (7)$$

While n_M in the formula refers to the number of participants to whom item m was applied, n_T refers to the total number of participants.

RESULTS

In this section, the findings on measurement accuracy and item security are given under headings.

Measurement Accuracy Results

In this section, in order to answer the measurement accuracy part of this research problem, the ability estimates obtained from F-MST and O-MST methods were compared under the same item pool and the same ability distributions. The RMSE, MAB, d and BIAS results obtained from the 72 conditions analyzed are presented in Table 6.

Table 6

Findings from All Conditions According to Different MST Approaches

Condition	Test Length	Module/Test Length Ratio	Ability Distribution	RMSE			MAB			BIAS	
				F-MST	O-MST	d_{RMSE}	F-MST	O-MST	d_{MAB}	F-MST	O-MST
1	20	L-S-S	Normal	0,382	0,371	1,657	0,300	0,291	1,587	0,008	0,007
2	20	L-S-S	Right Skewed	0,423	0,404	2,369	0,315	0,303	1,604	0,050	0,040
3	20	L-S-S	Left Skewed	0,431	0,414	1,887	0,311	0,303	1,070	-0,028	-0,029
4	20	L-S-S	Uniform	0,563	0,535	3,305	0,443	0,419	3,411	-0,053	-0,048
5	20	M-M-M	Normal	0,382	0,362	2,438	0,300	0,286	1,990	0,008	0,012
6	20	M-M-M	Right Skewed	0,417	0,399	2,407	0,314	0,302	1,723	0,046	0,037
7	20	M-M-M	Left Skewed	0,432	0,407	2,775	0,313	0,296	2,561	-0,023	-0,022
8	20	M-M-M	Uniform	0,552	0,524	2,235	0,432	0,409	2,715	-0,058	-0,047
9	20	S-S-L	Normal	0,398	0,364	4,013	0,314	0,285	4,164	0,008	0,004
10	20	S-S-L	Right Skewed	0,431	0,399	5,287	0,325	0,301	4,368	0,039	0,039
11	20	S-S-L	Left Skewed	0,438	0,405	4,146	0,322	0,297	4,928	-0,021	-0,022
12	20	S-S-L	Uniform	0,559	0,519	3,640	0,440	0,406	3,592	-0,047	-0,047
13	30	L-S-S	Normal	0,338	0,323	1,884	0,266	0,253	2,148	0,007	0,009
14	30	L-S-S	Right Skewed	0,380	0,361	2,540	0,281	0,268	2,004	0,044	0,039
15	30	L-S-S	Left Skewed	0,388	0,367	2,808	0,277	0,266	1,843	-0,025	-0,018
16	30	L-S-S	Uniform	0,496	0,463	3,486	0,387	0,360	3,366	-0,045	-0,044
17	30	M-M-M	Normal	0,333	0,321	1,604	0,262	0,252	1,675	0,007	0,005
18	30	M-M-M	Right Skewed	0,369	0,349	2,674	0,276	0,263	2,178	0,038	0,032
19	30	M-M-M	Left Skewed	0,382	0,356	2,886	0,275	0,260	2,513	-0,023	-0,015
20	30	M-M-M	Uniform	0,476	0,456	2,233	0,370	0,355	1,870	-0,052	-0,044
21	30	S-S-L	Normal	0,34	0,317	2,568	0,268	0,251	2,441	0,006	0,008
22	30	S-S-L	Right Skewed	0,372	0,344	4,317	0,279	0,259	3,640	0,037	0,032
23	30	S-S-L	Left Skewed	0,385	0,356	3,615	0,28	0,258	4,004	-0,021	-0,017

24	40	S-S-L	Uniform	0,477	0,444	3,127	0,372	0,346	2,886	-0,047	-0,043
25	40	L-S-S	Normal	0,306	0,295	1,579	0,24	0,232	1,456	0,005	0,002
26	40	L-S-S	Right Skewed	0,337	0,328	1,638	0,251	0,244	1,554	0,036	0,036
27	40	L-S-S	Left Skewed	0,350	0,336	2,158	0,249	0,243	1,206	-0,021	-0,015
28	40	L-S-S	Uniform	0,435	0,420	1,771	0,336	0,325	1,579	-0,047	-0,038
29	40	M-M-M	Normal	0,310	0,293	2,136	0,244	0,230	2,158	0,005	0,005
30	40	M-M-M	Right Skewed	0,339	0,324	2,513	0,254	0,240	2,548	0,034	0,029
31	40	M-M-M	Left Skewed	0,353	0,327	3,476	0,254	0,237	3,094	-0,019	-0,018
32	40	M-M-M	Uniform	0,433	0,413	2,361	0,336	0,320	2,297	-0,045	-0,042
33	40	S-S-L	Normal	0,317	0,291	3,733	0,250	0,230	3,350	0,005	0,002
34	40	S-S-L	Right Skewed	0,345	0,317	3,744	0,259	0,239	3,640	0,030	0,029
35	40	S-S-L	Left Skewed	0,357	0,330	3,610	0,259	0,240	3,183	-0,018	-0,014
36	40	S-S-L	Uniform	0,433	0,412	2,219	0,337	0,317	2,872	-0,039	-0,044

Note. S: Short, M=Medium, L=Long

As seen in Table 6, the RMSE values obtained from F-MST in all conditions vary between [0.306, 0.563]. The RMSE values obtained from O-MST are in the range of [0.291, 0.535]. As seen in the table, the RMSE values obtained from the F-MST method are higher than the RMSE values obtained from the O-MST in all conditions. This shows that O-MST is more effective than F-MST in all conditions. Also, as seen in the column, the Cohen's d effect size value is [Cohen's d > .80] in all conditions and the O-MST method has a larger effect than the F-MST method in all conditions. Similar to the RMSE results, the MAB value obtained from F-MST is higher than the MAB value obtained from O-MST in all conditions. The MAB values also confirm that O-MST estimates more effectively than F-MST. Similarly, as seen in the column, the Cohen's d eth size value is [d > .80] in all conditions, indicating that the O-MST method has a larger effect than the F-MST method.

As can be seen from both the RMSE and MAB columns, the measurement accuracy of both test approaches increases significantly as the test length increases. On the other hand, according to the module/test length ratio, it can be said that the S-S-L ratio in the F-MST has lower measurement accuracy than the L-S-S and M-M-M ratios, while all three ratios have similar measurement accuracy in the O-MST. According to ability distributions, both methods have the highest measurement accuracy in normal distributions, followed by right and left skewed distributions. It can be stated that Uniform distribution has the lowest measurement accuracy in terms of ability distribution of both test approaches. In the rest of the study, the measurement accuracy results of F-MST and O-MST methods according to test length, module/test length ratio and ability distribution were compared by obtaining the average value from Table 6.

Measurement Accuracy Results by Test Length

The results obtained from the averages of the related cells in Table 6 according to the test length are presented in Table 7.

Table 7

RMSE, MAB and d values by Test Length

Test Length	RMSE			MAB			BIAS	
	F-MST	O-MST	d_{RMSE}	F-MST	O-MST	d_{MAB}	F-MST	O-MST
20	0,451	0,425	3,013	0,344	0,325	2,809	-0,006	-0,006
30	0,387	0,365	2,783	0,293	0,277	2,516	-0,002	-0,001

40	0,360	0,341	2,578	0,272	0,258	2,411	-0,006	-0,006

When Table 7 is analyzed, it can be said that the O-MST approach estimates ability better than the F-MST at test lengths of 20, 30 and 40. While the difference in RMSE between 20 and 30 test lengths was higher, the difference in RMSE between 30 and 40 test lengths decreased. According to both the RMSE difference and the MAB difference, this indicates that even if the test length is increased, the effectiveness of ability estimation will not increase in proportion to the test length. The law of diminishing returns argues that production will not increase at the same rate as the factors of production increase, and that the proportional benefit will gradually decrease. Increasing the test length in tests will improve measurement accuracy up to a certain point. Increasing the test length beyond a certain point will not make the desired improvement in measurement accuracy, and the expected efficiency will not be achieved due to the physiological and psychological effects of the student such as fatigue, boredom and exhaustion. Therefore, the test length should be determined well. In this study, it is understood that 20 test lengths show low measurement accuracy compared to 30 and 40 test lengths. Therefore, it can be said that 30 or 40 test lengths give better results according to the results of this study. Test lengths of 50, 60, and 70 items are thought to be incompatible with the logic of adaptive testing approaches, and according to the law of diminishing returns, measurement accuracy should not be expected to decrease proportionally (Yasuda, Mae, Hull & Taniguchi, 2021).

Measurement Accuracy Results by Ability Distribution

The results obtained from the averages of the related cells in Table 6 according to ability distribution are presented in Table 8.

Table 8

RMSE, MAB, BIAS and d values according to Ability Distribution

Ability Distribution	RMSE			MAB			BIAS	
	F-MST	O-MST	d_{RMSE}	F-MST	O-MST	d_{MAB}	F-MST	O-MST
Normal	0,345	0,326	2,401	0,272	0,257	2,330	0,007	0,006
Right Skewed	0,379	0,358	3,054	0,284	0,269	2,584	0,039	0,035
Left Skewed	0,391	0,366	3,040	0,282	0,267	2,711	-0,022	-0,019
Uniform	0,492	0,465	2,708	0,384	0,362	2,732	-0,048	-0,044

When Table 8 is analyzed, it is seen that the O-MST approach provides more effective ability estimation than the F-MST approach according to the RMSE, MAB and d values in four different ability distributions (normal, Right Skewed, Left Skewed and Uniform). Both methods perform the most successful estimation in the normal distribution, followed by right skewed and left skewed distributions. Both methods have the lowest measurement accuracy in the uniform distribution. On the other hand, according to the effect size values, while the effect size difference in the normal distribution is small, the O-MST method estimates more effectively especially in right skewed, left skewed and uniform distributions. Therefore, O-MST method can be preferred in skewed or uniform ability distributions.

Measurement Accuracy Results by Module/Test Length Ratio

The results obtained from the averages of the related cells in Table 6 according to the module/test ratio are presented in Table 9.

Table 9

RMSE, MAB, BIAS and d values according to Module/Test Length Ratio

Module/Test Length Ratio	RMSE			MAB			BIAS	
	F-MST	O-MST	d_{RMSE}	F-MST	O-MST	d_{MAB}	F-MST	O-MST
L-S-S	0,402	0,385	2,257	0,305	0,292	1,902	-0,006	-0,005
M-M-M	0,398	0,378	2,478	0,303	0,288	2,277	-0,007	-0,006
S-S-L	0,404	0,375	3,668	0,309	0,286	3,589	-0,006	-0,006

Note. S: Short, M=Medium, L=Long

When Table 9 is analyzed, it is seen that the O-MST approach provides more effective ability estimation than the F-MST approach according to RMSE, MAB and d values at all module/test length ratio levels (L-S-S, M-M-M, S-S-L). According to Cohen's d values, the difference in measurement accuracy between the O-MST and F-MST approaches is the smallest in the L-S-S ratio, while it is similar in the M-M-M ratio. In the S-S-L ratio, the difference in measurement accuracy between the two approaches reached its highest value. In this case, it can be interpreted that in the F-MST approach in the last module, measurement accuracy decreases as the number of items in the last module increases. On the other hand, since the modules are not fixed in the O-MST approach, the O-MST is less affected by the length of the module/test ratio. Moreover, increasing the length of the last module in the O-MST approach increased the measurement precision of this approach. It should also be noted that in all three module/test length ratios, O-MST showed a larger effect than F-MST.

Item Security Results

In this section, in order to answer the research problems related to item security, the number of items used and average item exposure rate statistics for all conditions are presented in Table 10.

Table 10

Item Exposure Rates

Condition	Test Length	Module/Test Length Ratio	Ability Distribution	Total Item Number	Number of Items Used			Average Item Exposure Rate	
					F-MST	O-MST	Diff	F-MST	O-MST
1	20	L-S-S	Normal	400	105	163	-58	0,190	0,123
2	20	L-S-S	Right Skewed	400	105	160	-55	0,167	0,125
3	20	L-S-S	Left Skewed	400	105	156	-51	0,148	0,128
4	20	L-S-S	Uniform	400	105	174	-69	0,192	0,115
5	20	M-M-M	Normal	400	120	175	-55	0,167	0,114
6	20	M-M-M	Right Skewed	400	120	173	-53	0,147	0,116
7	20	M-M-M	Left Skewed	400	120	170	-50	0,190	0,118

Table 10 (Continued)

Condition	Test Length	Module/Test Length Ratio	Ability Distribution	Total Item Number	Number of Items Used			Average Item Exposure Rate	
					F-MST	O-MST	Diff	F-MST	O-MST

8	20	M-M-M	Uniform	400	120	181	-61	0,167	0,110
9	20	S-S-L	Normal	400	135	176	-41	0,148	0,114
10	20	S-S-L	Right Skewed	400	135	176	-41	0,190	0,114
11	20	S-S-L	Left Skewed	400	135	173	-38	0,167	0,116
12	20	S-S-L	Uniform	400	135	185	-50	0,148	0,108
13	30	L-S-S	Normal	400	156	221	-65	0,192	0,136
14	30	L-S-S	Right Skewed	400	156	221	-65	0,167	0,136
15	30	L-S-S	Left Skewed	400	156	221	-65	0,147	0,136
16	30	L-S-S	Uniform	400	156	232	-76	0,190	0,129
17	30	M-M-M	Normal	400	180	229	-49	0,167	0,131
18	30	M-M-M	Right Skewed	400	180	230	-50	0,148	0,130
19	30	M-M-M	Left Skewed	400	180	227	-47	0,190	0,132
20	30	M-M-M	Uniform	400	180	244	-64	0,167	0,123
21	30	S-S-L	Normal	400	204	239	-35	0,148	0,126
22	30	S-S-L	Right Skewed	400	204	236	-32	0,192	0,127
23	30	S-S-L	Left Skewed	400	204	235	-31	0,167	0,128
24	40	S-S-L	Uniform	400	204	252	-48	0,147	0,119
25	40	L-S-S	Normal	400	210	267	-57	0,190	0,150
26	40	L-S-S	Right Skewed	400	210	267	-57	0,167	0,150
27	40	L-S-S	Left Skewed	400	210	264	-54	0,148	0,152
28	40	L-S-S	Uniform	400	210	281	-71	0,190	0,142
29	40	M-M-M	Normal	400	240	285	-45	0,167	0,14
30	40	M-M-M	Right Skewed	400	240	282	-42	0,148	0,142
31	40	M-M-M	Left Skewed	400	240	282	-42	0,192	0,142
32	40	M-M-M	Uniform	400	240	297	-57	0,167	0,135
33	40	S-S-L	Normal	400	270	289	-19	0,147	0,138
34	40	S-S-L	Right Skewed	400	270	288	-18	0,190	0,139
35	40	S-S-L	Left Skewed	400	270	286	-16	0,167	0,140
36	40	S-S-L	Uniform	400	270	304	-34	0,148	0,132

The use of more items from the item pool both decreases the exposure rate and increases item security in terms of reducing the probability of disclosure of the items. When Table 10 is analyzed, it is seen that O-MST used more items from the item pool than F-MST in all conditions. It should be noted that O-MST used more items, even though the number of item use approached each other, especially as the test length increased. In addition, according to Table 10, the mean frequency of item use was 0.168 for the F-MST and 0.129 for the O-MST. The fact that the average item exposure rate of O-MST is lower than that of F-MST can be interpreted as an indicator that O-MST utilizes more items from the item pool and therefore uses the item pool more effectively. In this study, in order to limit the frequency of item use in F-MST and O-MST, 3 panels were formed in F-MST and the item exposure rate was fixed at 0.33 in O-MST with the ineligibility method. According to these findings, O-MST gives better results than F-MST in terms of average item exposure rate values. Therefore, it can be said that O-MST is safer than F-MST in terms of item security.

Some graphs containing item exposure rates according are shown as examples for examining item exposure rates according to the relevant variables below. In the continuation of the study, the item exposure rate and the number of items used in terms of test length, ability distribution and module/test length according to the item security sub-problems were analyzed with the averages of the related cells in Table 10.

Item Security Results by Test Length

The results obtained from the averages of the related cells in Table 10 for different test lengths are presented in Table 11.

Table 11

Number of Items Used and Average Frequency of Item Used by Test Length

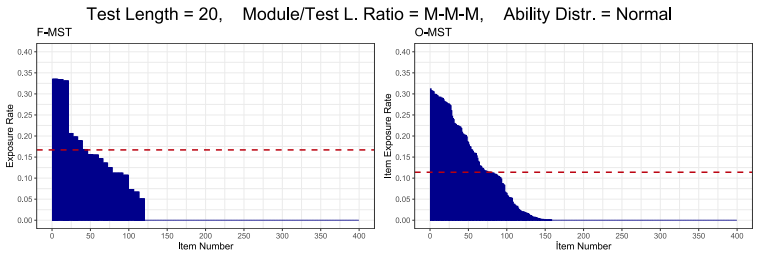
Test Length	Number of Items Used		Average Item Exposure Rate	
	F-MST	O-MST	F-MST	O-MST
20	120	172	0,168	0,117
30	180	232	0,168	0,130
40	240	283	0,168	0,142

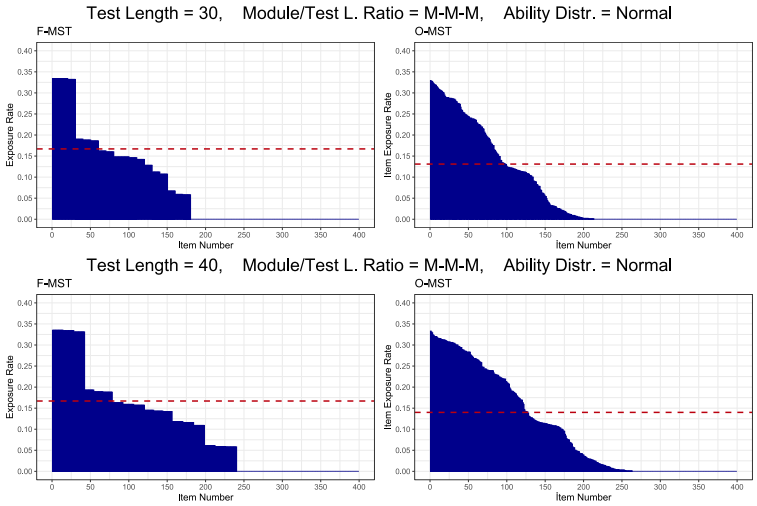
When Table 11 is analyzed, it can be stated that the O-MST approach provides better results in terms of item security according to both the number of items used and the average item use frequencies at 20 test lengths. Similarly, at 30 and 40 test lengths, it can be said that the O-MST approach provides better results in terms of item security than the F-MST in terms of both the number of items used and the average item use frequency. On the other hand, the difference in the number of items used between F-MST and O-MST approaches decreases as the test length increases. Similar to this finding, while the average item exposure rate was 16.8% for F-MST at all test lengths, it was calculated as 11.7%, 13.0% and 14.2% for O-MST, respectively. In this case, while the item security between O-MST and F-MST was better in favor of O-MST at short test lengths, the item security efficiency of both approaches converged as the test length increased. Therefore, it can be said that O-MST is more efficient than F-MST in terms of test and item security in all three test lengths, while O-MST is more efficient in terms of item security in short tests.

Figure 3 shows the item exposure rate plots for three different conditions with the same ability distribution and the same module/test length ratio, with test lengths of 20, 30 and 40.

Figure 3

Graph of Item Exposure Rates by Test Length





The blue sections in the graphs show the item exposure rate. Column graphs were created by sorting the items from the maximum to the minimum item exposure rate. The red dashed lines on the graph show the average item exposure rate of the related condition. Figure 3 shows that the O-MST approach uses a higher number of items and has a lower average item exposure rate than the F-MST approach in all test length conditions. On the other hand, in the F-MST approach, the item exposure rate of some items is at the level of 0.33, especially due to the items in the routing module, and it is thought that these items are more likely to be exposed. In the O-MST approach, while the item exposure rate of some items is at the level of 0.34, the frequency of item use of the following items decreases. In addition, it can be said that the use of more items in the O-MST approach is beneficial for both item security and more effective use of the item pool.

Item Security Results by Ability Distributions

The results obtained from the averages of the related cells in Table 10 according to ability distribution are presented in Table 12.

Table 12

Number of Items Used and Average Item Exposure Rate by Ability Distribution

Ability Distribution	Number of Items Used		Average Item Exposure Rate	
	F-MST	O-MST	F-MST	O-MST
Normal	180	227	0,168	0,130
Right Skewed	180	226	0,168	0,131
Left Skewed	180	224	0,168	0,132
Uniform	180	239	0,168	0,124

When Table 12 is examined, it can be stated that the O-MST approach offers better results than the F-MST in terms of item security according to both the number of items used and the average frequency of item use in the normal, right skewed, left skewed and uniform distributions. In normal, right skewed, left skewed and uniform distributions, it is seen that the O-MST approach has a higher number of items used than the F-MST approach. In addition, while the difference in the number of items used between the F-MST and O-MST approaches is similar for the first three ability distributions, it is seen that the O-MST approach has a higher number of items used and a lower item exposure rate than the F-MST approach in the uniform distribution. Therefore, it can be stated that the O-MST approach produces better results than the F-MST in terms of item security in Uniform distributions.

Figure 4 shows the item-by-item item exposure rate graphs of four different conditions according to four different ability distributions (Normal, right skewed, left skewed and uniform) with a test length of 40 and a module/test length ratio of L-S-S.

Figure 4

Graph of Item Exposure Rates According to Ability Distribution

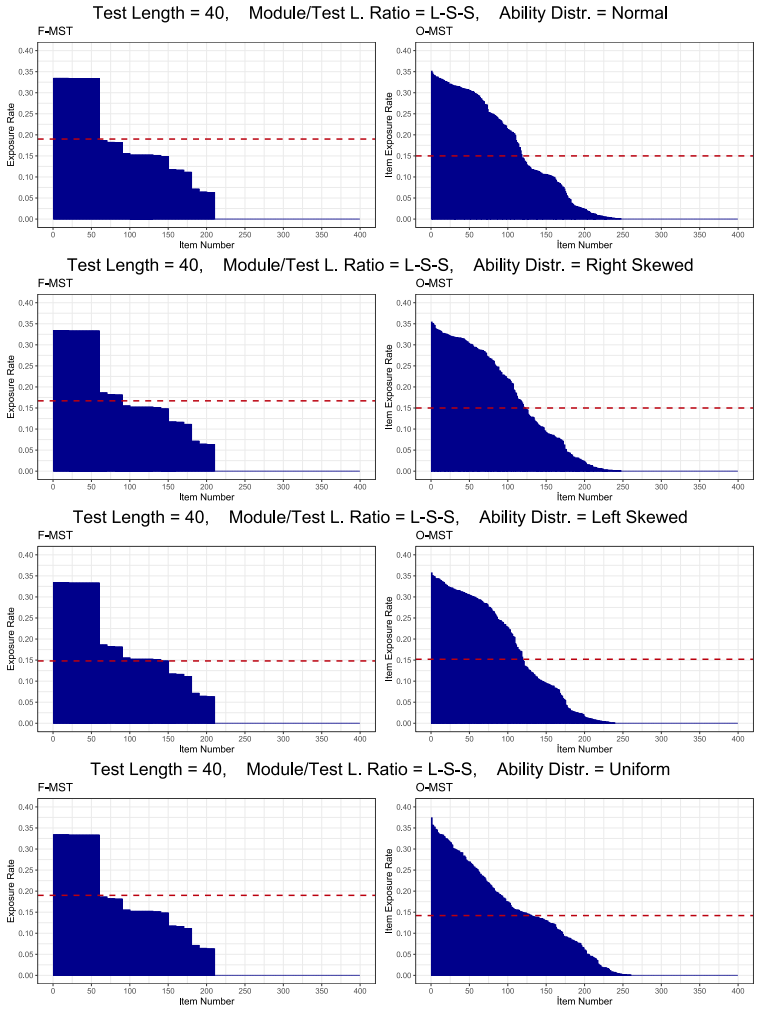


Figure 4 shows that the O-MST approach uses a higher number of items and has a lower average item exposure rate (red dashed lines) than the F-MST approach under all different ability distributions.

Item Security Results by Module/Test Length Ratio

The results obtained from the averages of the related cells in Table 10 according to the module/test length ratio are presented in Table 13.

Table 13

Number of Items Used and Average Item Exposure Rates by Module/Test Length Ratio

Module/Test Length Ratio	Number of Items Used		Average Item Exposure Rate	
	F-MST	O-MST	F-MST	O-MST

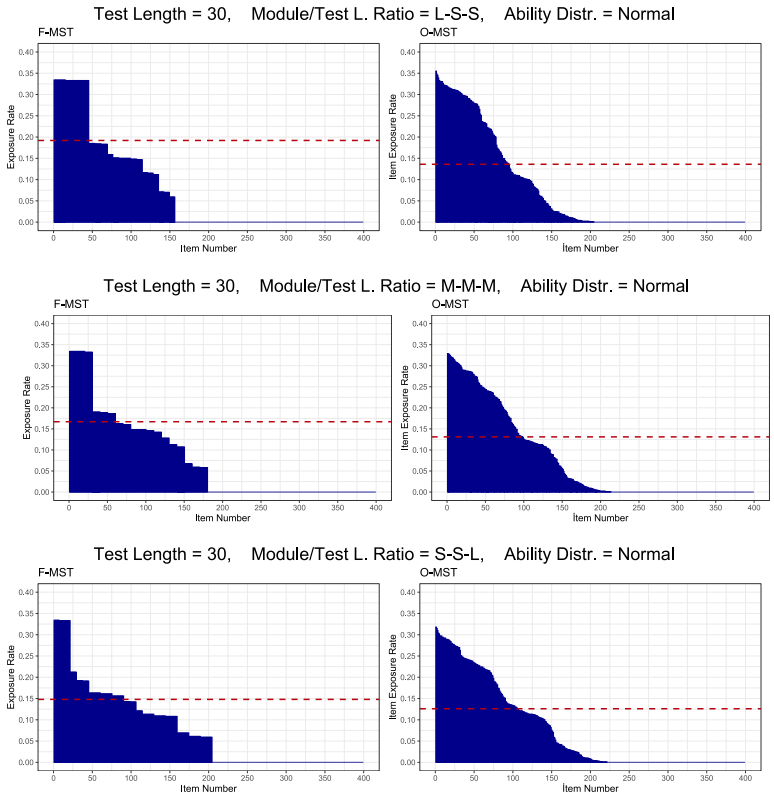
L-S-S	157	219	0,174	0,135
M-M-M	180	231	0,168	0,128
S-S-L	203	237	0,163	0,125

Note. S: Short, M=Medium, L=Long.

When Table 13 is analyzed, it is seen that O-MST approach provides better results than F-MST in terms of item security according to both the number of items used and average item exposure rates in L-S-S, M-M-M and S-S-L module/test length ratios. On the other hand, the difference between the number of items used in F-MST and O-MST is higher at the L-S-S ratio because of the higher number of items in the routing module, while the difference between the number of items used decreases as we move towards the S-S-L ratio. In the S-S-L ratio, the difference in the number of items used between the two approaches is minimized.

Figure 5

Graph of Item Exposure Rates by Module/Test Length Ratio



Note. S: Short, M=Medium, L=Long

Figure 5 shows that the O-MST approach uses a higher number of items and has a lower average item exposure rate (red dashed lines) than the F-MST approach under all different module/test ratio length conditions. On the other hand, since the length of the routing module is longer in the L-S-S ratio, the item exposure rate of the items in the routing module is higher in the F-MST approach. This may lead to problems in terms of item security. In the S-S-L ratio, on the other hand, since the length of the routing module is shorter, the frequency of item use is lower. However, the disadvantage of the S-S-L ratio is that it has lower measurement precision than the other two module/test length ratios. Therefore, when deciding on the module/test length ratio, both item security and measurement accuracy should be taken into account to determine the most appropriate point.

DISCUSSION AND CONCLUSION

In recent years, there has been a growing interest in the use of adaptive testing for test management, especially in large-scale assessments (Ebenbeck & Gebhardt, 2022; Tomashev et al., 2018; Yamamoto et al., 2018; Zheng & Chang, 2015). F-MST is an adaptive testing approach where test management is easy due to the fixed modules and panels structure. O-MST, on the other hand, is a new testing approach that can be called

a hybrid of CAT and MST that provides successful solutions under a large number of complex test specifications and constraints (Choi, van der Linden, 2018; van der Linden, 2021). In this study, F-MST and O-MST were compared in terms of measurement accuracy and item security. The results are presented below under headings.

Measurement Accuracy

The results of this study show that O-MST offers better measurement accuracy than F-MST according to both RMSE and MAB statistics. Similarly, O-MST shows a larger effect than F-MST in terms of measurement accuracy in all conditions (Cohen's d > 0.8). In other words, the O-MST approach estimates ability level more effectively than the F-MST. There are many studies in the literature that reach similar conclusions to this finding. Zheng and Chang (2015), in a simulation study comparing CAT, O-MST and F-MST, state that CAT and O-MST provide very similar measurement accuracy, while F-MST produces lower measurement accuracy results. Tay (2015), in his doctoral thesis comparing the classification accuracy and classification consistency of O-MST and F-MST, states that O-MST provides better results than F-MST in all conditions. van der Linden and Diao (2016), in their study comparing CAT, O-MST, F-MST and LT, report that CAT and O-MST provide the best results in terms of measurement accuracy, followed by F-MST, while LT ranks last. Han and Guo (2016) state that CAT and O-MST offer very close measurement accuracy, while F-MST offers less measurement accuracy than these two approaches. van der Linden (2021) compares LT, CAT, O-MST and F-MST approaches and states that O-MST is more successful than F-MST in terms of measurement accuracy. It is seen that the results of the studies in the literature comparing O-MST and F-MST and the current study are quite similar (Tay, 2015; van der Linden, 2021; Zheng & Chang, 2015).

In this study, it was concluded that O-MST provided better measurement accuracy than F-MST at test lengths of 20, 30 and 40. On the other hand, while the O-MST approach was more effective than the F-MST at 20 test lengths, the measurement accuracy of the two approaches converged as the test length increased. Yasuda, Mae, Hull, and Taniguchi (2021), in their study on the test length of CAT, concluded that measurement accuracy and bias did not decrease significantly as test length increased. Therefore, in the current study, it is seen that the measurement accuracy of both approaches does not decrease in proportion to the test length as the test length increases. Van der Linden (2021) explains this situation with the law of diminishing returns. It is also stated that tests with long test lengths contradict the logic of the adaptive testing approach (van der Linden, 2021).

In the study, the two MST approaches were compared under four different ability distributions: normal, right skewed, left skewed and uniform. It was concluded that the O-MST approach provided better measurement accuracy than the F-MST in all four different ability distributions. In both approaches, the measurement accuracy is highest in the normal distribution, followed by right skewed and left skewed distributions. In the uniform distribution, measurement accuracy is low for both approaches. On the other hand, considering both the structure of the measured variable and the distribution of the population, non-normally distributed samples are frequently encountered in education (MoNE, 2021). At this point, it was concluded that the measurement accuracy of O-MST is considerably better than F-MST, especially in right and left skewed distributions. As a result, the O-MST approach is less affected by changing ability distributions, while the F-MST approach is more affected.

In the study, the two MST approaches were compared under three different module/test length ratios: S-S-L, M-M-M, L-S-S. According to the module/test length ratio, O-MST was found to offer higher measurement accuracy than F-MST in all ratios. There are studies indicating that F-MST provides better measurement accuracy when the length of the routing module is longer (Boztunç, 2019; Cai, Anthony, Albano, & Roussos, 2021; Kim & Plake, 1993; Zheng, 2016). In this study, similar to the studies in the literature, F-MST gives better measurement accuracy results as the length of the routing module increases. Again, similar to the results in the literature, as the module lengths of the second and third stages increase, the measurement accuracy of the F-MST decreases due to the need for a large number of items used in these designs. The O-MST, on the other hand, is less affected by the module/test length ratio and provides better measurement precision than other module/test length ratios when the number of items in the last stage increases. In other words, one of the most significant findings of the current study is that in the F-MST, measurement accuracy increases as the length of the first module increases, whereas in the O-MST, measurement accuracy increases

as the length of the last module increases. On the other hand, O-MST is better than F-MST in terms of measurement accuracy in all conditions according to the overall module/test length ratio.

Item Security

In terms of item security, it was concluded that O-MST had a lower item exposure rate and a higher number of items used than F-MST in all conditions. This result shows that O-MST utilizes the item pool more effectively than F-MST. On the other hand, in O-MST, the item exposure rates of the items decrease with a certain acceleration, while in F-MST, the item exposure rates of the items in the routing module are high, and the item exposure rates decrease in the second and third stages. Similar to the results of the current study, Zheng and Chang (2015) reported that O-MST had lower test overlap rates and that participants following similar routes in F-MST may pose item security problems. In the F-MST, the panel and modules are fixed, so the items in the modules in the participant's route are known by the test administrator. Since the O-MST assembles a test for each participant at his/her own ability level, the items to be administered to the participant are not known or specified beforehand by anyone, including the test administrator. Therefore, O-MST creates unique tests customized to the individual. In this context, O-MST has an advantage over F-MST in terms of item and test security (Zheng and Chang, 2015).

It was concluded that the O-MST approach had higher item counts and lower item exposure rates than the F-MST approach for all test lengths. This shows that O-MST produces more efficient results in terms of item safety than F-MST in all test lengths.

It was concluded that O-MST had a higher number of items used and a lower item exposure rate than F-MST in different ability distributions. Therefore, it can be stated that O-MST gives better results than F-MST in terms of item security. In addition, while the difference in the number of items used and average item exposure rate between O-MST and F-MST approaches was similar in normal, right skewed and left skewed distributions, it was observed that the difference between the two approaches increased in favor of O-MST in uniform distributions. This is due to the increase in the number of participants at extreme ability levels in the uniform distributions. As a result, it can be stated that the O-MST approach is more effective than the F-MST in terms of item security in uniform distributions.

At all module/test length ratios, O-MST has a higher number of items used and a lower item exposure rate than F-MST. This shows that O-MST gives better results than F-MST at different module/test length ratios in terms of item security. In addition, since the routing module is longer in the L-S-S ratio in the F-MST approach, the number of items used is lower and the item exposure rate is higher. In the S-S-L ratio, since the modules in the third stage are longer, the number of items used is higher and the item exposure rate is lower. In this case, it can be said that the L-S-S-S ratio is more efficient in terms of item security than the S-S-L ratio.

The results of this study show that the O-MST approach is more advantageous than the F-MST approach in terms of both measurement accuracy and item and test security. On the other hand, the F-MST approach has two important advantages over the O-MST approach: (I) easy explanability of the fixed panel and module structure to test takers and (II) easy administration of the test implementation (Yan, Von Davier, & Lewis, 2016). These two advantages are among the most important reasons for choosing the F-MST. The fact that the structure and scoring method of the test can be easily understood by test takers and other relevant stakeholders is important in terms of making them feel that the test is conducted in a fair manner. Therefore, it is thought that the choice of an adaptive testing approach should be made by examining the risk status of the test, the type of assessment (norm or criterion-referenced), the explainability of the test administration to the participants, and the sociological perspective of the society on the relevant test, in addition to measurement accuracy and item-test security.

Limitations

This study has several limitations that should be considered when interpreting its results. First, the analyses were entirely based on simulated data, which may not fully reflect the complexity of real testing environments. Second, the study was simulated only within the context of dichotomously scored multiple-

choice items. Future research could address this limitation by including polytomous or mixed-format tests. Third, the study used hypothetical item pools rather than operational test items, which may differ in terms of psychometric properties and content balance. Future studies may extend this work by employing real item pools, larger and more diverse samples, and empirical data obtained from operational computerized testing systems.

Suggestions

Under this heading, there are suggestions for practitioners and researchers.

Suggestions for Practitioners

In line with the results of this study, seven recommendations for practitioners are presented below. (1) It is recommended that the O-MST method should be preferred over the F-MST for short test lengths because it provides more effective ability estimations. (2) If the O-MST design is to be preferred, it is recommended to increase the last module length because it will improve measurement accuracy. (3) If the number of items in the last stage of the MST design is desired to be high, it is recommended to prefer the O-MST approach instead of F-MST. (4) If the number of items to be included in the orientation module is desired to be high, it can be stated that the F-MST method is also preferable since the effectiveness level of the F-MST method and the O-MST method are close to each other. (5) If the cut-off score of the test to be administered is around the midpoints of the ability scale, the O-MST and F-MST methods offer very similar measurement accuracy. However, if the cut-off score is at the extreme ability levels, it is recommended to prefer the O-MST method. (6) If it is thought that there will be item and test security problems in the test administration, it is recommended to prefer O-MST method instead of F-MST. (7) The findings of this study may inform the design and implementation of large-scale assessments such as PISA, TIMSS, and national examinations that seek to balance measurement accuracy and test security. The proposed simulation framework can guide test developers in optimizing item selection and exposure control strategies in future adaptive testing applications.

Suggestions for Researchers

Ten suggestions for researchers are presented below to guide researchers in future studies. (1) In this study, an item pool was generated in a simulation environment according to TIMSS parameter distributions. A similar study can be designed with a real data pool. (2) In this study, “1-2-3” design was used as F-MST panel design. Similar studies can be conducted with different designs such as “1-3”, “1-4”, “1-2-4”, etc. (3) In this study, the EAP method was preferred as the ability estimation method. Different studies can be designed by using MLE, MLEF, MAP methods as ability estimation methods. (4) In this study, only test assembly procedures were performed by adding constraints according to four different content areas. Using real item pools, similar studies can be conducted under different test features and constraints such as common-root items, enemy items, number of words, average time. In addition, the effectiveness of CAT and O-MST approaches can be compared by increasing the number of constraints. (5) In F-MST, MFI method was used as a routing method. Different studies can be designed by using AMI, number of correct and other methods as routing methods. (6) In this study, O-MST and F-MST were compared. Research can be designed comparing different F-MST designs with the hybrid CAT approach. (7) In this study, test lengths of 20, 30 and 40 are considered as test lengths. Similar studies can be conducted under different test lengths. Different studies can be conducted to determine the optimal test length for CAT approaches. (8) In this study, L-S-S, M-M-M and S-S-L module/test length ratios were considered. Different studies can be conducted under different module/test length ratios. (9) In this study, the item pool size was set as 400. Similar studies can be designed by varying the item pool size. (10) In this study, measurement accuracy and item security were investigated. In future studies, the classification accuracy of O-MST and F-MST can be compared.

REFERENCES

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology*, *43*(4), 695–716. https://doi.org/10.1111/j.1744-6570.1990.tb00679.x

Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). Altering the level of difficulty in computer adaptive testing. *Applied Measurement in Education*, *5*(2), 137–149. https://doi.org/10.1207/s15324818ame0502_4

Boztunç Öztürk, N. (2019). How the length and characteristics of routing module affect ability estimation in ca-MST? *Universal Journal of Educational Research*, *7*(1), 164–170. https://doi.org/10.13189/ujer.2019.070121

Breithaupt, K. J., Mills, C. N., & Melican, G. J. (2006). Facing the opportunities of the future. *Computer-based testing and the Internet: Issues and advances*, 219-251.

Bulut, O. (2021). Beyond multiple-choice with digital assessments. *ELearn*, *2021*(Special Issue), 1–10. https://doi.org/10.1145/3472394

Bulut, O., & Sünbül, Ö. (2017). R Programlama Dili ile Madde Tepki Kuramında Monte Carlo Simülasyon Çalışmaları. *Egitimde ve Psikolojide Olcme ve Degerlendirme Dergisi, 8*(3), 266–287. https://doi.org/10.21031/epod.305821

Cai, L., Albano, A. D., & Roussos, L. A. (2021). An investigation of item calibration methods in multistage testing. *Measurement: Interdisciplinary Research and Perspectives*, *19*(3), 163–178. https://doi.org/10.1080/15366367.2021.1878778

Carlson, S. (2000). ETS finds flaws in the way online GRE rates some students. *Chronicle of Higher Education, 47*(8), A47.

Cetin-Berber, D. D., Sari, H. I., & Huggins-Manley, A. C. (2019). Imputation methods to deal with missing responses in computerized adaptive multistage testing. *Educational and psychological measurement, 79*(3), 495-511.

Chang, H.-H. (2004). Understanding computerized adaptive testing: From Robbins-Monro to Lord and beyond. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 117-133). Thousand Oaks, CA: Sage.

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. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika, 73*(3), 441–450.

Choi, S. W., & van der Linden, W. J. (2018). Ensuring content validity of patient-reported outcomes: a shadow-test approach to their adaptive measurement. *Quality of Life Research, 27*(7), 1683-1693.

Choi, S. W., Lim, S., & van der Linden, W. J. (2021). TestDesign: an optimal test design approach to constructing fixed and adaptive tests in R. *Behaviormetrika,* 1-39.

Choi, S. W., Moellering, K. T., Li, J., & van der Linden, W. J. (2016). Optimal reassembly of shadow tests in CAT. *Applied psychological measurement*, *40*(7), 469-485. https://doi.org/10.1177/0146621616654597.

Cohen, J. (1988). Statistical power analysis fort he behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Demir, H., & Gelbal, S. (2025). A systematic review on Computerized Adaptive Testing. *Journal of Education Faculty*, *27*(1), 137–150. https://doi.org/10.17556/erziefd.1577880

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. *Educational measurement, 4*, 471-515.

Ebenbeck, N., & Gebhardt, M. (2022). Simulating computerized adaptive testing in special education based on inclusive progress monitoring data. *Frontiers in Education, 7*. https://doi.org/10.3389/feduc.2022.945733

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice, 35*(2), 36-49.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*(4), 521-532.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education*. McGram-Hill Publishing.

Gür, R., & Gülleroğlu, H. (2020). The effect of item exposure control methods on measurement precision and test security under different measurement conditions in computerized adaptive testing. *TED EĞİTİM VE BİLİM*, *45*(202), 113–139. https://doi.org/10.15390/eb.2020.8256

Hambleton, R. K., & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass–fail decisions. *Applied Measurement in Education, 19*(3), 221-239.

Han, K. C. T., & Guo, F. (2016). Multistage testing by shaping modules on the fly. In Computerized Multistage Testing (pp. 157-172). Chapman and Hall/CRC.

Han, K. T. (2007). WinGen: Windows software that generates item response theory parameters and item responses. *Applied Psychological Measurement*, *31*(5), 457–459. https://doi.org/10.1177/0146621607299271

Harwell, M., Stone, C. A., Hsu, T. C. & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101-125. doi: 10.1177/014662169602000201

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*(2), 101–125. https://doi.org/10.1177/014662169602000201

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement Issues and Practice*, *26*(2), 44–52. https://doi.org/10.1111/j.1745-3992.2007.00093.x

Khorramdel, L., Pokropek, A., Joo, S. H., Kirsch, I., & Halderman, L. (2020). Examining gender DIF and gender differences in the PISA 2018 reading literacy scale: A partial invariance approach. *Psychological Test and Assessment Modeling, 62*(2), 179-231.

Kim, H., & Plake, B. (1993). Monte Carlo simulation comparison of two-stage testing and computer adaptive testing. Unpublished doctoral dissertation, University of Nebraska, Lincoln.

Kirsch, I., & Lennon, M. L. (2017). PIAAC: a new design for a new era. *Large-scale Assessments in Education, 5*(1), 1-22.

Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a computerized adaptive test more motivating than a fixed-item test? *Applied Psychological Measurement*, *41*(7), 495–511. https://doi.org/10.1177/0146621617707556

Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika, 36*(3), 227-242. https://doi.org/10.1007/BF02297844

Luo, X., & Kim, D. (2018). A top-down approach to designing the computerized adaptive multistage test. *Journal of Educational Measurement, 55*(2), 243-263.

Magis, D., Yan, D., & Von Davier, A. A. (2017). Computerized adaptive and multistage testing with R: Using packages catr and mstr. Springer.

Makhorin A (2017). GNU Linear Programming Kit. Version 4.61, URL http://www.gnu.org/software/glpk/glpk.html.

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology*, *110*(1), 27–45. https://doi.org/10.1037/edu0000205

Mead, A. D. (2006). An introduction to multistage testing. *Applied Measurement in Education*, *19*(3), 185–187. https://doi.org/10.1207/s15324818ame1903_1

MEB (2021). 2021 Ortaöğretim Kurumlarına İlişkin Merkezi Sınav Raporu. Milli Eğitim Bakanlığı.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine, 38*(11), 2074-2102.

OECD (2023). *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*, OECD Publishing, Paris, https://doi.org/10.1787/53f23881-en.

OECD (2024), *PISA 2022 Technical Report,* PISA, OECD Publishing, Paris, https://doi.org/10.1787/01820d6d-en.

Ortner, T. M., Weißkopf, E., & Koch, T. (2014). I will probably fail: Higher ability students' motivational experiences during adaptive achievement testing. *European Journal of Psychological Assessment: Official Organ of the European Association of Psychological Assessment*, *30*(1), 48–56. https://doi.org/10.1027/1015-5759/a000168

Patsula, L. N., & Hambleton, R. K. (1999). A comparative study of ability estimates obtained from computer-adaptive and multi-stage testing. In annual meeting of the National Council on Measurement in Education, Montreal, Quebec.

Pine, S. M., Church, A. T., Gialluca, K. A., & Weiss, D. J. (1979). *Effects of Computerized Adaptive Testing on Black and White Students*. Minnesota Univ Minneapolis Dept Of Psychology.

Saatçioğlu, F. M., & Atar, H. Y. (2022). Investigation of the effect of parameter estimation and classification accuracy in mixture IRT models under different conditions. *International Journal of Assessment Tools in Education*, *9*(4), 1013–1029. https://doi.org/10.21449/ijate.1164590

Stark, S., & Chernyshenko, O. S. (2006). Multistage testing: Widely or narrowly applicable?. *Applied Measurement in Education, 19*(3), 257-260.

Şahin, A., & Weiss, D. J. (2015). Effects of calibration sample size and item bank size on ability estimation in computerized adaptive testing. *Educational Sciences Theory & Practice, 15*(6). https://doi.org/10.12738/estp.2015.6.0102

Tay, P. H. (2015). On-the-fly assembled multistage adaptive testing. University of Illinois at Urbana-Champaign.

Tomashev, M. V., Avdeev, A. S., & Krasnova, M. V. (2018). Adaptive testing as a tool for managing quality of education. *Informatics and Education, 9*, 27–33. https://doi.org/10.32517/0234-0453-2018-33-9-27-33

van der Linden, W. J. (2009). Constrained adaptive testing with shadow tests. In Elements of adaptive testing (pp. 31-55). Springer, New York, NY.

van der Linden, W. J. (2010). *Elements of adaptive testing*. C. A. Glas (Ed.). New York, NY: Springer.

van der Linden, W. J. (2018). Optimal test design. Handbook of item response theory: Vol. 3. Applications, 167-195.

van der Linden, W. J. (2021). Review of the shadow-test approach to adaptive testing. Behaviormetrika, 1-22.

van der Linden, W. J., & Diao, Q. (2016). *Using a universal shadow-test assembler with multistage testing*. Computerized multistage testing: Theory and applications, 101-118.

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics: A Quarterly Publication Sponsored by the American Educational Research Association and the American Statistical Association*, *29*(3), 273–291. https://doi.org/10.3102/10769986029003273

van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, *44*(2), 117–130. https://doi.org/10.1111/j.1745-3984.2007.00030.x

Xu, L., Jiang, Z., Han, Y., Liang, H., & Ouyang, J. (2023). Developing computerized Adaptive Testing for a national health professionals exam: An attempt from psychometric simulations. *Perspectives on Medical Education*, *12*(1), 462–471. https://doi.org/10.5334/pme.855

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. *Educational Measurement Issues and Practice*, *37*(4), 16–27. https://doi.org/10.1111/emip.12226

Yan, D., Von Davier, A. A., & Lewis, C. (Eds.). (2016). Computerized multistage testing: Theory and applications. CRC Press.

Yasuda, J. I., Mae, N., Hull, M. M., & Taniguchi, M. A. (2021). Optimizing the length of computerized adaptive testing for the force concept inventory. *Physical review physics education research, 17*(1), 1-15.

Yasuda, J.-I., Mae, N., Hull, M. M., & Taniguchi, M.-A. (2021). Optimizing the length of computerized adaptive testing for the Force Concept Inventory. *Physical Review Physics Education Research*, *17*(1). https://doi.org/10.1103/physrevphyseducres.17.010115

Yigiter, M. S., & Dogan, N. (2023). Computerized multistage testing: Principles, designs and practices with R. *Measurement: Interdisciplinary Research and Perspectives*, *21*(4), 254–277. https://doi.org/10.1080/15366367.2022.2158017

Yiğiter, M. S., & Boduroğlu, E. (2024). Item Response Theory assumptions: A comprehensive review of studies with document analysis. *International Journal of Educational Studies and Policy, 5*(2), 119-138. https://doi.org/10.5281/ZENODO.14016086

Yiğiter, M. S., & Doğan, N. (2023). The effect of test design on misrouting in computerized multistage testing. *International Journal of Turkish Education Sciences*, *2023*(21), 549–587. https://doi.org/10.46778/goputeb.1267319

Zheng, W. (2016). Making test batteries adaptive by using multistage testing techniques (Doctoral dissertation, University of North Carolina, Greensboro, NC).

Zheng, Y., & Chang, H.-H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*, *39*(2), 104–118. https://doi.org/10.1177/0146621614544519