BOĞAZİÇİ ÜNİVERSİTESİ EĞİTİM DERGİSİ
2024, Cilt 41-3, Sayı 2
https://doi.org/10.52597/buje.1329338

Çok Aşamalı Testlerin Panel Deseni, Modül Uzunluğu, Örneklem Büyüklüğü ve Yetenek Parametresi Kestirim Yöntemleri Açısından Farklı Koşullar Altında Karşılaştırılması[a]

Serap Büyükkıdık[b] ve Fatma Gökçen Ayva Yörü[c]

Öz

Bu araştırmada çeşitli simülasyon koşullarında çok aşamalı testlerin performansları, hata kareler ortalamasının karekökü (Root Mean Square Error-RMSE), tahminin standart hatası (Standard Error of Estimate-SEE), yanlılık (BIAS) ve ortalama mutlak hata (Mean Absolute Error-MAE) değerlendirme kriterleri açısından karşılaştırılmıştır. Test simülasyonunda panel deseni (1-3, 1-2-3, 1-3-3), modül uzunluğu (6, 12, 18), örneklem büyüklüğü (300, 1000, 3000), yetenek parametresi kestirim yöntemi (beklenen sonsal dağılım [Expected a Posteriori-EAP], maksimum sonsal dağılım [Maximum a Posteriori-MAP] ve sınırlı en çok olabilirlik kestirimi [Maximum Likelihood Estimation with Fences-MLEF]) olmak üzere 81 koşul (3x3x3x3) belirlenmiştir. Araştırma sonucunda RMSE ile MAE değerlerinin genellikle benzer sonuçlar verdiği ve modül uzunluğu arttıkça ölçme doğruluğunun da arttığı bulunmuştur. Ayrıca RMSE, SEE ve MAE'nin 1-3 panel deseninde en yüksek, 1-3-3 deseninde ise en düşük değerleri aldığı saptanmıştır. Araştırmacılara 1-3-3 panel deseninde, en az 12 modül uzunluğunda ve EAP yöntemi kullanarak çalışma yapmaları önerilmektedir.

Anahtar Kelimeler: çok aşamalı test, panel desen, modül uzunluğu, örneklem büyüklüğü, yetenek parametresi kestirim yöntemi

Makale Hakkında

Gönderim tarihi: 18.07.2023
Düzeltme tarihi: 08.11.2023
Kabul tarihi: 17.11.2023
Elektronik Yayın Tarihi: 30.08.2024

Giriş

Eğitim değerlendirmelerinde sınava girenlerin bilgi, beceri ve yeteneklerini ölçmede uzun zamandır yaygın olarak geleneksel kağıt-kalem testleri (doğrusal testler) kullanılmaktadır (Magis vd., 2017; Yan vd., 2014a). Hızlı gelişen bilgisayar ve internet teknolojisi sayesinde küçük ölçekli tanımlayıcı testlerden, büyük ölçekli yeterlik testlerine kadar her türlü amaca yönelik tasarlanabilen bilgisayar tabanlı sınavlar kolaylıkla uygulanabilmektedir (Zheng ve Chang, 2015). Madde tepki kuramı ve bilgisayarların yaratıcı kullanımı ise daha kısa uzunlukta ve daha fazla güvenirliğe sahip testler oluşturulabilmesine olanak sağlamıştır. Günümüzde de madde düzeyinde uyarlanabilir testlerin farklı çeşitleri yaygın biçimde kullanılmaktadır ve bu durum değerlendirmeyi de verimli hale getirmektedir (Mead, 2006). Gelişen teknoloji beraberinde son yirmi yılda bilgisayar tabanlı testlere (Computer Based Test - CBT) olan ilgiyi artırmış (Magis vd., 2017) ve geleneksel kağıt-kalem testlerinin çoğu bilgisayar ortamında uygulanmıştır (Luecht ve Sireci, 2011; Magis vd., 2017). Bilgisayar ortamında yapılan test uygulamaları ise "bilgisayar tabanlı testler (Computer Based Testing-CBT)", "bilgisayar uyarlamalı testler (Computerized Adaptive Test - CAT)" ve "çok aşamalı testler (Multi Stage Test - MST)" olmak üzere üç şekilde yapılabilmektedir (Zheng ve Chang, 2015). Bilgisayar tabanlı testler yanıtlayıcıların test maddelerini bilgisayar aracılığıyla okuyup yanıtladıkları ve ayrıca yanıtlayıcılara cevaplarını gözden geçirip test sonunda çıkış yapma imkanı sağlayan testlerdir (Wang vd., 2004). Bilgisayar tabanlı test uygulamalarında yanıtlayıcıların hepsi aynı maddeleri cevaplamaktadır (Mason vd., 2001). Bilgisayar uyarlamalı test uygulamalarında ise yanıtlayıcının önceki maddelere vermiş olduğu yanıtlara göre hesaplanan

[a] *Bu araştırmanın bir kısmı 8. Uluslararası Eğitim ve Psikolojide Ölçme ve Değerlendirme Kongresi'nde (2022) sözlü bildiri olarak sunulmuştur.*

[b] *Sorumlu yazar, İstanbul Üniversitesi-Cerrahpaşa Hasan Ali Yücel Eğitim Fakültesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı serap.buyukkidik@iuc.edu.tr, ORCID: 0000-0003-4335-2949*

[c] *Afyon Kocatepe Üniversitesi Eğitim Fakültesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, fayva@aku.edu.tr, ORCID: 0000-0002-4555-1987*

Alıntılamak için: Büyükkıdık, S., & Ayva Yörü, F. G. (2024). Çok aşamalı testlerin panel deseni, modül uzunluğu, örneklem büyüklüğü ve yetenek parametresi kestirim yöntemleri açısından farklı koşullar altında karşılaştırılması. *Boğaziçi Üniversitesi Eğitim Dergisi*, *41-3*(2), 9-27, https://doi.org/10.52597/buje.1329338

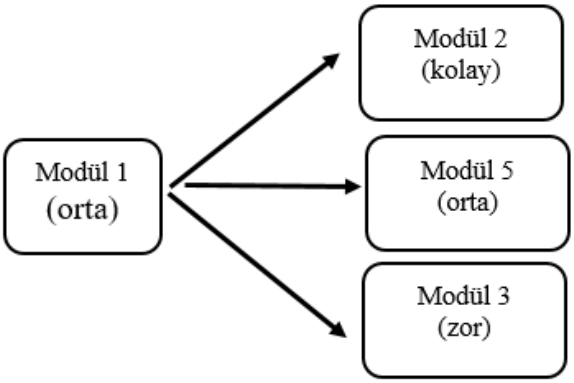
yetenek düzeyine uygun olarak sonraki maddelerin seçildiği test uygulamasıdır (Drasgow ve Mattern, 2006). Dolayısıyla bilgisayar uyarlamalı testlerde farklı yetenek düzeyindeki yanıtlayıcılara farklı maddeler uygulanmaktadır (Weiss, 1983).

Geleneksel kağıt-kalem testlerinde sınava giren bireylerin hepsinin aynı uzunluktaki testi alması, esnek olmayan sınav programı, ölçme için verimsiz olması gibi sebeplerden dolayı bazı dezavantajlara sahiptir. Bilgisayar uyarlamalı testler ise daha kısa test uzunluğu, ölçme için daha verimli olması, sınava girenlere daha esnek sınav programı sağlaması ve kopya çekmenin önüne geçmesini sağlaması noktalarında geleneksel kâğıt-kalem testlerine göre avantajları bulunmaktadır. Bilgisayarlı çok aşamalı testler ise özellikleri ve verimlilikleri nedeniyle son yıllarda popüler hale gelmiştir (Magis vd., 2017; Zheng vd., 2012). Sınavların kalitesini artırmak amacıyla yapılan çok aşamalı test uygulamaları (Stark ve Chernyshenko, 2006), doğrusal test formları (kâğıt-kalem testleri ve bilgisayar tabanlı testler) ile bilgisayar uyarlamalı test uygulamalarının avantajlarını içermektedir (Zheng vd., 2012). Öncelikle çok aşamalı testler geleneksel kâğıt-kalem testlere göre yeterlik aralığının uç noktaları da dâhil olmak üzere yeterlik ölçeğinde daha hassas ve verimli ölçme yapma imkânı sağlar. Test etme ve puanların raporlanma süresi daha kısadır (Hendrickson, 2007). Ayrıca geleneksel sabit uzunluktaki doğrusal testlere (Linear Fixed Length Test - LFT) göre test verimliliğini ve karar doğruluğunu artırma potansiyelini sunar (Stark ve Chernyshenko, 2006). Çok aşamalı testlerin bilgisayar uyarlamalı test uygulamalarına göre farklarından birincisi sınava giren bireylerin serbest hareket edebilmesini sağlamasıdır (Stark ve Chernyshenko, 2006). Daha açık ifadeyle bilgisayar uyarlamalı test uygulamasının aksine bireyler mevcut aşama içinde bulunan maddeler arasında ileri geri gidebilir, bir sonraki maddeye geçmeden önce cevaplarını gözden geçirebilir, cevaplarını değiştirebilir ve bu sayede test içeriği üzerinde daha yüksek kontrol sağlayabilir (Han ve Guo, 2014; Mead, 2006; Sarı vd., 2016; Stark ve Chernyshenko, 2006; Zheng vd., 2012). Böylece çok aşamalı testler bireylerin sınav sırasında daha az stres ve endişe hissetmelerini sağlar (Zheng vd., 2012). Buna rağmen çok aşamalı testlerde bireylerin önceki aşama ya da aşamalara geri dönmelerine veya önceki modüldeki maddeleri yeniden gözden geçirmelerine izin verilmez (Sarı vd., 2016). İkincisi ise konu uzmanları boyutluluk, olumsuz etki ve değişen madde fonksiyonu (Differential Item Functioning - DIF) gibi analizleri paneller yayınlanmadan önce yapabilmektedir. Böylece çok aşamalı testler test yapısı üzerinde daha fazla kontrol sağlamaktadır (Stark ve Chernyshenko, 2006). Bütün bunların yanında çok aşamalı testlerde erken sonlandırma durumu bilgisayar uyarlamalı test uygulamalarına göre daha az esnektir. Bilgisayar uyarlamalı testlerde farklı sonlandırma kurallarına dayalı olarak, ölçme doğruluğu yeterli olduğunda ve içerik gereksinimleri karşılandığı sürece herhangi bir noktada sonlanabilir. Buna rağmen çok aşamalı testler modül tabanlı olduğundan dolayı ancak tüm aşamalar tamamlandıktan sonra test sonlanır (Zheng vd., 2012).

Bilgisayar uyarlamalı test uygulamalarında her bir madde, bireylerin önceki maddelere verdiği yanıtlara dayalı olarak madde havuzundan seçilerek bireylere uygulanır (Zheng ve Chang, 2015). Çok aşamalı testlerde ise maddeleri bireylere uyarlamalı olarak teker teker uygulanmak yerine, her birey için madde grupları (modül) bulunmaktadır (Magis vd., 2017). Modülde yer alan maddeler içerik bakımından farklı olsalar dahi istatistiksel özellikleri (madde ayırt ediciliği, madde güçlüğü vb.) benzerdir. Modülün içinde bulunduğu düzeyler ise aşamaları oluşturur ve aşamaların bir araya gelmesiyle ortaya çıkan desenlere panel denir (Zenisky ve Hambleton, 2014). Birkaç farklı aşamadan oluşan her panel, çeşitli zorluk seviyelerinde ve belirli sayıda modüle yani madde gruplarına sahiptir (Luecht ve Sireci, 2011; Zheng ve Chang, 2015). Bireylerin önceki aşamalarda göstermiş olduğu performanslarına göre her aşamada sadece bir tane modül (daha kolay ya da daha zor) seçilerek test uyarlanır. Bu durum test uzunluğunun azalmasını sağlar. Çok aşamalı testlere ilişkin farklı panel desen örnekleri bulunmaktadır (Bkz. Şekil 1, 2 ve 3). Şekillerde görüldüğü gibi birinci aşamada bir tane yönlendirme modülü bulunmaktadır (bazı çalışmalarda her modülde beş madde olan iki küçük modül kullanılır) ve modül seçim kriterleri yönlendirme olarak adlandırılır (Magis vd., 2017; Yan vd., 2014a). Yönlendirme modülü sınava giren kişinin yeterlik düzeyini belirlemek için kullanılır (Sarı vd., 2016). Sonraki aşamalarda ise olası modül sayısı artırılır. Aşama sayısında olduğu gibi çeşitli zorluklara sahip daha fazla modül eklenmesi, daha fazla adaptasyona ve dolayısıyla da test içinde daha fazla esnekliğe izin verir (Yan vd., 2014a). Ayrıca panellerdeki aşama sayısı ve her modüldeki madde sayısı testin amacına göre değişebilir (Sarı vd., 2016). Şekil 1, 2 ve 3'te çok aşamalı test uygulamalarında sıklıkla kullanılan panel deseni örnekleri sunulmuştur.

Şekil 1

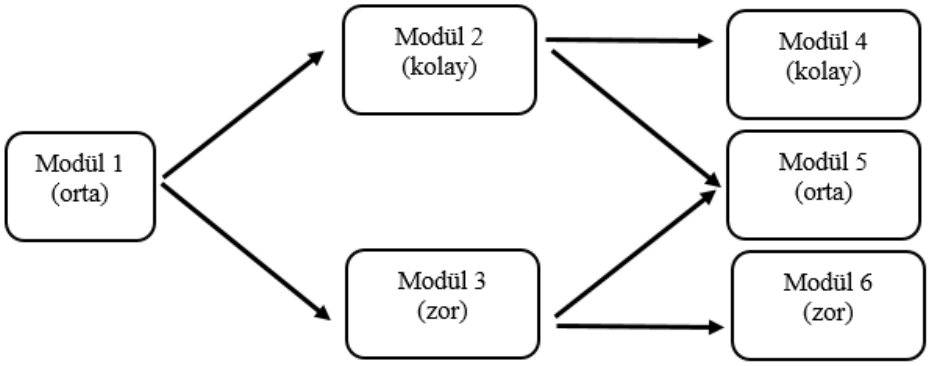
Çok Aşamalı Test Tasarımında 1-3 Panel Desen Örneği (İki Aşamalı Test Tasarımı)



Şekil 1'de iki aşamadan oluşan 1-3 panel desen örneği bulunmaktadır. Öncelikle her birey orta zorlukta ortak bir modülle (modül 1 - yönlendirme testi) maddeleri yanıtlamaya başlar. Birinci aşamadan sonra bireylerin geçici yetenek kestirimi hesaplanır ve test uygulaması ölçüm açısından en bilgilendirici olan ikinci aşamadaki modülü seçer. Her bir modül farklı zorluk seviyesindedir ve bireylere her aşamada sadece bir tane modül uygulanır (Zenisky ve Hambleton, 2014). İkinci aşamada yetenek kestirimi hesaplanarak çok aşamalı test uygulaması sona erer. İki aşamalı 1-3 panel deseninde testi alan bireyin izleyebileceği toplam üç olası yol bulunmaktadır. Bunlar: "1.yol: orta-kolay; 2. yol: orta-orta; 3. yol: orta-zor".

Şekil 2

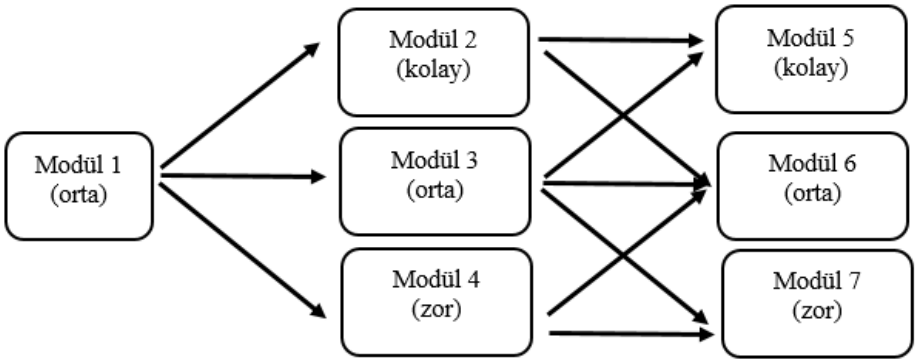
Çok Aşamalı Test Tasarımında 1-2-3 Panel Desen Örneği (Üç Aşamalı Test Tasarımı)



Şekil 2'de üç aşamadan oluşan 1-2-3 panel desen örneği bulunmaktadır. Bu panel deseninde ikinci aşamada iki, üçüncü aşamada ise üç adet modül bulunmaktadır. 1-3 Panel desenindeki gibi tüm bireyler orta zorlukta ortak bir modülle maddeleri yanıtlar ve bireylerin geçici yetenek kestirimi hesaplanarak ikinci aşamadaki modül seçilir. İkinci aşamada da süreç aynı şekilde ilerleyerek yetenek kestirim süreci tekrarlanır ve bu doğrultuda bireyler için en uygun olan üçüncü aşamada yer alan modülden biri uygulanır (Zenisky ve Hambleton, 2014). Üçüncü aşamada da yetenek kestirimi hesaplanarak çok aşamalı test uygulaması sona erer. 1-2-3 panel deseninde testi alan bireyin izleyebileceği toplam dört olası yol bulunmaktadır. Bunlar: "*1.yol*: orta-kolay-kolay; *2. yol*: orta-kolay-orta; *3. yol*: orta-zor-orta; 4. *yol*: orta-zor-zor".

Şekil 3

Çok Aşamalı Test Tasarımında 1-3-3 Panel Desen Örneği (Üç Aşamalı Test Tasarımı)



Şekil 3'te üç aşamadan oluşan 1-3-3 panel desen örneği bulunmaktadır. Bu desende 1-2-3 panel deseninden farklı olarak ikinci aşamada üç adet modül bulunmaktadır. 1-3-3 panel deseninde testi alan bireyin izleyebileceği toplam yedi olası yol bulunmaktadır. Bunlar: "*1.yol*: orta-kolay-kolay; *2. yol*: orta-kolay-orta; *3. yol*: orta-orta-kolay; 4. *yol*: orta-orta-orta; 5. *yol*: orta-orta-zor; *6. yol*: orta-zor-orta; 7. *yol*: orta-zor-zor."

Geniş ölçekli standartlaştırılmış test uygulamalarında, sınava girenlerin farklı alt gruplarına birden çok panel dağıtılmaktadır. Belirli bir amaca yönelik çok aşamalı test tasarlamak için, nihai karar vermeden önce çeşitli tasarım özelliklerinin (her modüldeki madde sayısı, aşama sayısı, test uzunluğu, modüllerin zorluk dereceleri vb.) araştırılması gerekmektedir (Magis vd., 2017). Çünkü çok aşamalı test uygulamasında modüldeki madde sayısı ve panel deseni ölçme doğruluğunu etkileyebilir (Zenisky ve Hambleton, 2014). Ayrıca testin ne kadar süreceği, kaç aşamanın yeterli olacağı, modüller arasında kaç yol olacağı, modüller arasındaki yönlendirme kurallarının ne olacağı, modüllerin nasıl seçileceği, testin tamamı için zorluk derecesinin nasıl elde edileceği, her modülün ne kadar süreceği, her modül için istenen zorluk ve zorlukların dağılımı, modülü birleştirmede maddelerin nasıl seçileceği gibi durumların araştırılması çok aşamalı test tasarımında önemlidir (Yan vd., 2014b). Yapılan araştırmalarda test tasarım sürecinin hala karmaşık bir süreç olmaya devam ettiği belirtilmektedir (Luo ve Kim, 2018). Dolayısıyla madde seçimi ile madde ve kişi parametrelerinin kestiriminde matematiksel modellerin belirlenmesiyle de en uygun tasarım stratejilerini araştırmak için kapsamlı simülasyon çalışmaları yapmak ve testin amaçlarına göre değerlendirme yapmak gerekmektedir (Magis vd., 2017).

Literatürde yer alan çalışmalar incelendiğinde bilgisayar uyarlamalı testler ve çok aşamalı testlerin farklı koşullar altında karşılaştırıldığı birçok çalışma (Davis ve Dodd, 2003; Jodoin, 2003; Kim vd., 2012; Kim ve Plake, 1993; Patsula, 1999; Sari, 2016; Wang, 2017; Zheng, vd., 2012) bulunmaktadır. Sadece yurt içindeki alanyazında çok aşamalı testleri kendi içinde farklı koşullar altında inceleyen sınırlı sayıda çalışmalara (Boztunç Öztürk, 2019; Büyükkıdık ve Ayva Yörü, 2022; Doğruöz, 2018; Erdem Kara, 2019; Ertaş Polat, 2022) rastlanmıştır. Ayrıca çok aşamalı test uygulaması ile ilgili yapılan çalışmalarda yetenek parametre kestirim yöntemlerini ele alan yine sınırlı sayıda çalışmalar bulunmaktadır (Büyükkıdık ve Ayva Yörü, 2022; Ertaş Polat, 2022; Şahin ve Boztunç Öztürk, 2019). Çok aşamalı testlerin avantajları ve önemi dikkate alındığında; çok aşamalı testlerde yetenek parametre kestirim yöntemlerinin ele alınarak incelenmesinin literatüre ve konuyla ilgili çalışma yapan araştırmacılara katkıda bulunabileceği düşünülmektedir. Ayrıca Magis vd.nin (2017) belirttiği gibi bu alanda yapılacak simülasyon çalışmalarına ihtiyaç duyulmaktadır.

Ek olarak yurtdışında uygulanan "Ulusal Eğitimdeki Gelişmelerin Değerlendirilmesi (The National Assessment of Educational Progress - NAEP), Hukuk Fakülteleri Kabul Sınavı (Law School Admission Test – LSAT), Lisansüstü Eğitim Sınavı (Graduate Record Examination - GRE) ve Tıp Alanında Uzmanlık Sınavı (The U.S. Medical Licensure Examination - USMLE)" gibi geniş ölçekli test uygulamalarının çoğunda bilgisayar uyarlamalı testlerin yerine çok aşamalı test uygulamalarına geçildiği görülmektedir. Buna rağmen ülkemizde geniş ölçekli test uygulamalarında çok aşamalı testler henüz kullanılmamaktadır. Çok aşamalı test uygulamaları üzerine yapılacak olan çalışmalarla bu uygulamanın sonuçlarının değerlendirilmesi ve ülkemizde de uygulanan geniş ölçekli testlerde kullanılabilirliğinin araştırılması açısından da alana katkı sağlayacağı düşünülmektedir. Ayrıca bu araştırmada ele alınan simülasyon koşulları, gerçek veriye ait parametre değerleri kullanılarak simülasyon çalışmasının yürütülmesi, üç farklı yetenek kestirim yönteminin kullanılması ve dört

farklı değerlendirme kriterinin bir arada kullanılması boyutlarıyla, mevcut araştırma diğer çalışmalardan farklılık göstermektedir.

Araştırmanın Amacı

Bu araştırmada çok aşamalı test performansının farklı koşullar (panel deseni, modül uzunluğu, örneklem büyüklüğü, yetenek parametresi kestirimi) altında incelenmesi amaçlanmıştır. Elde edilen sonuçlara dayanılarak yapılan kestirimlerin hangi koşullar altında daha iyi sonuç verdiği konusunda daha kesin bir anlayışa ulaşılabileceği düşünülmektedir. Araştırmanın amacı doğrultusunda aşağıdaki soruya yanıt aranmaya çalışılmıştır:

"Çok aşamalı test simülasyonunda panel deseni (1-3, 1-2-3 ve 1-3-3), modül uzunluğu (6, 12 ve 18), örneklem büyüklüğü (300, 1000 ve 3000) ve yetenek parametresi kestirim yöntemlerine (EAP, MAP ve MLEF) göre hata kareler ortalamasının karekökü (RMSE), tahminin standart hatası (SEE), yanlılık (BIAS) ve ortalama mutlak hata (MAE) değerleri nasıl değişim göstermektedir?"

Yöntem

Araştırma Deseni

Bu araştırmada çok aşamalı bireye uyarlanmış test performansları farklı panel deseni, modül uzunluğu, örneklem büyüklüğü ve yetenek parametresi kestirim yöntemleri açısından karşılaştırılmıştır. Araştırma kapsamında ele alınan veriler bilgisayar programı aracılığıyla türetildiğinden dolayı bu çalışma bir simülasyon çalışmasıdır.

Çok Aşamalı Bireye Uyarlanmış Test Simülasyonu

Bu araştırmada simülasyon verisinin üretilmesinde Uluslararası Matematik ve Fen Eğilimleri Araştırması 2015 (Trends in International Mathematics and Science Study - TIMSS) sekizinci sınıf öğrencilerine uygulanan matematik başarı testine ait madde parametre değerleri kullanılmıştır. Bu sayede simülasyon verisi gerçek veriye ait parametre değerleri ile desteklenmiştir. TIMSS, uluslararası eğitim başarılarını değerlendirme kuruluşu (International Association for the Evaluation of Educational - IEA) tarafından yürütülmekte olup uygulamaya katılan ülkelerin dördüncü ve sekizinci sınıf öğrencilerinin matematik ve fen bilimleri alanlarında kazanmış oldukları bilgi ve becerilerin çok yönlü değerlendirilmesini sağlayan ve dört yılda bir yapılan tarama çalışmasıdır. TIMSS uygulaması öğrencilerin bu alanlardaki başarılarını ölçmenin yanı sıra eğitim sisteminin etkinliği ile ülkelerin eğitim sistemleri arasındaki farklılıkların değerlendirilmesini amaçlanmaktadır. Dolayısıyla TIMSS uygulamasında öğrencilere uygulanmak üzere başarı testlerine ek olarak çeşitli anketler de bulunmaktadır (MEB, 2016). TIMSS 2015 sekizinci sınıf matematik testinde toplam 297 madde bulunmakta ve bu maddelerin 159 tanesini 1-0 şeklinde puanlanan çoktan seçmeli maddeler oluşturmaktadır. Bu araştırmada TIMSS 2015 yılına ait uygulama sonuçlarına göre hazırlanan raporda yer alan madde parametre değerleri kullanılmış ve bu değerlere ise TIMSS'in resmi internet sitesinden ulaşılmıştır (International Association for the Evaluation of Educational Achievement, 2021). Madde parametre değerlerine ilişkin bilgiler Tablo 1'de sunulmuştur.

Tablo 1

TIMSS 2015 Sekizinci Sınıf Matematik Testine Ait Madde Parametre Değerleri

Madde parametreleri	Minimum	Maksimum	$\bar{X}$	*SS*
a	0,504	2,351	1,282	0,365
b	-0,833	1,727	0,564	0,541
c	0,077	0,404	0,206	0,077

Tablo 1'de a, b ve c parametrelerine ait değerler yer almaktadır. Madde ayırt edicilik parametresi (a) teorik olarak $-\infty$ ile $+\infty$ değerler almasına rağmen uygulamalarda genellikle $a = 0$ ile $a = +2$ arasındadır. Sıfıra yaklaştıkça maddelerin ayırt etme gücü düşmeye başlar. Madde güçlük parametresi (b) ise genellikle -2 ile +2

arasında değerler alır ve +2'ye yaklaştıkça maddeler zorlaşır (Hambleton vd., 1991). c parametresi ise şans ile maddeyi doğru yanıtlama olasılığıdır ve teorik olarak 0 ve 1 arasında değerler almasına rağmen uygulamada 0 ile 0,35 arasındadır (Baker, 2001). Tabloda yer alan parametre değerleri incelendiğinde ise ayırt edicilik parametresinin ortalaması 1,282 (SS = 0,365; min = 0,504; max = 2,351), güçlük parametresinin ortalaması 0,564 (SS = 0,541; min = -0,833; max = 1,727) ve şans parametresinin ortalaması 0,206'dır (SS = 0,077; min = 0,077; max = 0,404).

Araştırma kapsamında 81 simülasyon koşulu (3 panel deseni x 3 modül uzunluğu x 3 örneklem büyüklüğü x 3 yetenek parametresi kestirim yöntemi) ele alınmıştır. Simülasyon koşulları belirlenirken alanyazında yer alan çalışmalar incelenmiş ve bu doğrultuda koşullar oluşturulmuştur. Koşulların belirlenmesine ilişkin gerekli açıklamalara aşağıda yer verilmiştir. Araştırmada ele alınan simülasyon koşulları Tablo 2'de sunulmuştur.

Tablo 2

Simülasyon Koşulları

Koşullar		Koşul sayısı
Panel deseni	1-3 1-2-3 1-3-3	3
Modül uzunluğu	6 12 18	3
Örneklem büyüklüğü	300 1000 3000	3
Yetenek parametresi kestirim yöntemi	MLEF EAP MAP	3

Panel Deseni

Araştırma kapsamında iki aşamalı (1-3) ve üç aşamalı (1-2-3 ve 1-3-3) olmak üzere üç farklı panel deseni kullanılmıştır. Birden fazla panelin olması panel, modül ve madde kullanım oranının azalmasına yardımcı olur. Bu durum test güvenliği için önemlidir aksi halde kopya çekme ve madde paylaşma gibi sorunlar oluşacaktır (Yan vd., 2014a). Çok aşamalı testlerde en az iki aşama olması gerekmektedir ve literatürde sıklıkla kullanılan iki aşamalı panel desenleri; 1-2 (Wang vd., 2012), 1-3'tür (Boztunç Öztürk, 2019; Kim vd., 2015; Patsula, 1999; Reese vd., 1999; Schnipke ve Reese 1999; Wang vd., 2012). Üç aşamalı panel desenleri; 1-2-2 (Breithaupt ve Hare 2007; Chen, 2010; Patsula, 1999; Şahin, 2020; Wang vd., 2012; Zenisky, 2004), 1-3-3 (Boztunç Öztürk, 2019; Dallas vd., 2012; Edwards vd., 2012; Hambleton ve Xing 2006; Jodoin vd., 2006; Keng ve Dodd, 2009; Luecht vd., 2006; Park, 2015; Patsula, 1999; Şahin, 2020; Wang vd., 2012; Zenisky, 2004; Zheng ve Chang, 2015) ve 1-2-3 (Armstrong ve Roussos 2005; Wang vd., 2012; Yan vd., 2014a; Zenisky, 2004) ve 1-3-2 (Zenisky, 2004) şeklindedir. Dört aşamalı panel desenleri; 1-1-2-3 (Belov ve Armstrong 2008; Weissman vd., 2007); 1-2-2-2, 1-2-3-4, 1-3-3-3 ve 1-3-4-5 (Wang vd., 2012) ve beş aşamalı panel deseni 1-5-5-5-5 (Davey ve Lee 2011) şeklindedir. Araştırma kapsamında ele alınan panel desenleri görüldüğü üzere literatürde en çok araştırılan panel desenleridir.

Çoğu çok aşamalı test uygulamasında iki, üç veya dört aşama kullanılmıştır. İki aşamalı testlerde bireylere bir yönlendirme, bir tane de ölçüm testi uygulanır. İki aşamalı testlerde bir tane adaptasyon noktası olması ise yönlendirme hatası olasılığının yüksek olmasına neden olmaktadır (Yan vd., 2014a). Ayrıca Armstrong vd., (2004) yaptıkları çalışmalarında aşama sayısının dörtten fazla olmasının test sonuçlarında anlamlı kazanımlar sağlamadığını, çok aşamalı test tasarımlarının her aşamasının iki veya üç modülden oluşmasının ve iki veya üç aşamanın da yeterli olduğunu belirtmişlerdir. Patsula (1999) yaptığı çalışmasında ise aşama sayısını ikiden üçe çıkarmanın genel olarak yetenek parametresi kestirimindeki hata miktarını azalttığını belirtmiştir. Ayrıca aşama sayısının artmasıyla yetenek parametresi kestirimin doğruluğunu ve panel deseninin etkililiğini de artırmaktadır (Patsula, 1999).

Modül Uzunluğu

Bu araştırmada ele alınan diğer koşul ise modül uzunluğudur. Modül uzunlukları testin yapısına bağlı olarak küçük (5 ile 10 madde) ve büyük (50 ile 100 madde) arasında değer alabilmektedir. Ayrıca modül uzunlukları aşamalar arasında (Jodoin vd., 2006; Luecht, 2000) ve ortalama güçlük derecesine göre de değişkenlik gösterebilir (Luecht, 2000). Ayrıca madde sayısının artması ölçme doğruluğunu da artırmaktadır (Patsula, 1999). Buna rağmen modül uzunluğunun çok olması test yapısının karmaşık olmasına neden olur. Stark ve Chernyshenko (2006) her modüldeki madde sayısının 15 veya 20 madde olmasını önermektedirler. Literatürde yer alan çalışmalar incelendiğinde ise Kim vd., (2015) "15-20-25-30", Şahin (2020) "10-15-20", Sari (2016) "24 ve 48 madde" ve genel olarak ta 30 ile 60 madde arasında değişen test uzunlukları üzerinde çalışmalar yapıldığı görülmüştür (Hambleton ve Xing, 2006; Jodoin vd., 2006; Patsula, 1999; Zenisky, 2004). Benzer bir şekilde 20 maddelik modülün uygun olduğunu gösteren çalışmalar da mevcuttur (Kim ve Plake, 1993; Zheng ve Chang, 2014). Bu araştırma kapsamında ise 6, 12 ve 18 olmak üzere üç farklı modül uzunluğu incelenmiştir. Örneğin modül uzunluğunun 12 olduğu 1-3 panel deseninde bireyin test sonunda toplam 24 maddeyi, 1-2-3 ve 1-3-3 panel deseninde ise toplam 36 maddeyi cevaplayacağı şekilde tasarlama yapılmıştır. Zenisky (2004) yapmış olduğu çalışmasında modül uzunluğunun her aşamada sabit bir sayı olması gerektiğini belirtmiştir. Bu çalışmada da her bir modüldeki madde sayıları birbirine eşit olarak alınmıştır. Araştırma kapsamında ele alınan panel desenlerindeki modül sayıları ve test uzunlukları Tablo 3'te sunulmuştur.

Tablo 3

Panel Desenlerine Göre Modül Sayıları ve Test Uzunluğu

Panel deseni	Modül sayısı	Modül uzunluğu	Her panel için testin uzunluğu
1-3	4	6	24
1-3	4	12	48
1-3	4	18	72
1-2-3	6	6	36
1-2-3	6	12	72
1-2-3	6	18	108
1-3-3	7	6	42
1-3-3	7	12	84
1-3-3	7	18	126

Araştırma kapsamında her bir modül oluşturulurken MSTgen programında belirtilen koşullar çerçevesinde birden fazla simülasyon gerçekleştirilmiş ve ardından belirlenen değerlere göre en uygun modül kapsama dahil edilmeden önce o modül için test bilgi fonksiyon (TIF) grafikleri incelenmiştir. Örneğin 1-3-3 panel tasarımı için küçük b fark koşulu; yönlendirme modülü bir TIF merkezini (0,00 θ noktası) yansıtacak şekilde yapılandırılmıştır. İkinci aşama üç TIF merkezini (θ noktaları -0,05, 0,00, +0,5) yansıtacak şekilde yapılandırılmıştır. Üçüncü aşama ise üç TIF merkezini yansıtacak şekilde inşa edilmiştir (θ noktaları -1,00, 0,00, +1,00).

Yetenek Parametresi Kestirim Yöntemleri

Madde tepki kuramı ile ilgili yapılan çalışmalarda yetenek parametre kestirimlerinde birçok yöntem kullanılmaktadır (van der Linden ve Pashley, 2010). Bunlardan en sık kullanılanlar; En Çok Olabilirlik Kestirimi [Maximum Likelihood Estimation - MLE] (Lord, 1980); Maksimum Sonsal Dağılım [Maximum a Posteriori - MAP] (Samejima, 1968); Beklenen Sonsal Dağılım [Expected a Posteriori - EAP] (Bock ve Mislevy, 1982); Marjinal En Çok Olabilirlik [Marginal Maximum Likelihood - MML] (Bock ve Aitkin, 1981); Ağırlıklandırılmış En Çok Olabilirlik [Weight Maximum Likelihood - WML] (Warm, 1989); Owen'in Ardışık Bayesçi Yaklaşımı [Owen's Sequential Bayesian - OSB] (Owen, 1975) şekline sıralanabilir. Bunlardan MLE yönteminin bilgisayar uyarlamalı çok aşamalı testlerde kullanımının en önemli avantajı madde parametreleri önceden bilindiğinden dolayı, madde parametrelerinin bilinmediği doğrusal testlere kıyasla daha yansız kestirim yapmasıdır. MAP yönteminin en büyük avantajı ise MLE'den daha iyi performans göstermesidir (Wang ve

Vispoel, 1998). MAP ve EAP yöntemlerinin avantajı ise tam puan ya da sıfır puan alanlar için kestirimler yapmayı sağlamasıdır (De Ayala, 2009). Bununla birlikte tümü 0 ya da tümü 1'lerden oluşanlar olmak üzere belirli yanıt örüntülerinde ve özellikle test uzunluğunun kısa olduğu durumlarda MLE yönteminin kullanılamaması bu yöntemin önemli bir eksikliğidir. Özellikle bilgisayar uyarlamalı test uygulamalarının ilk aşamalarında sadece birkaç yanıtın kaydedilmesi durumlarında MLE yönteminin üstünden gelemeyeceği durumların ortaya çıkmasına neden olabilir (Han, 2016). MLE yönteminin çok aşamalı test uygulamasında kullanıldığında ortaya çıkabilecek dezavantajlı durumlar göz önüne alınarak Han (2016) tarafından alternatif bir yaklaşım olan MLEF önerilmiştir. Bu yöntem temelde MLE ile aynı şekilde çalışmaktadır. Fakat bu yöntemde, sabit alt ve üst sınırları belirleyerek $\hat{\theta}$ ölçeğini sınırlamak yerine, MLEF puan kestiriminde anlamlı bir log olabilirlik fonksiyonu aralığında "sınırlar (fence)" oluşturmak için sabit yanıtlara sahip iki sanal madde yerleştirir. MLEF'de ilk sanal madde alt sınır olarak işlev görür ve b parametresi θ dağılımının beklenen alt sınırının olduğu yerde ayarlanır. b parametre değeri için alt sınır, test formundaki herhangi bir maddeden daha düşük olmalıdır. İkinci sanal madde de benzer şekilde üst sınır olarak işlev görür ve b parametresi θ dağılımının beklenen üst sınırının olduğu yerde ayarlanır. Alt sınır maddesi için yanıt her zaman "1" olarak sabitlenirken üst sınır maddesi için ise yanıt "0" olarak sabitlenir (Han, 2016). Çok aşamalı test uygulaması üzerine yapılan çalışmalarda yetenek parametre kestirim yöntemlerinden MLEF'i de kullanan sınırlı sayıda çalışmaya rastlanmıştır (Şahin ve Boztunç Öztürk 2019; Büyükkıdık ve Ayva Yörü, 2022). Yukarıda yer alan bilgiler ışığında yöntemlerin avantajları dikkate alınarak bu çalışmada yetenek parametre kestirim yöntemlerinden EAP, MAP ve MLEF yöntemleri tercih edilmiştir.

Örneklem Büyüklüğü

Araştırmada üç farklı (300, 1000 ve 3000) örneklem büyüklüğü ele alınmıştır. Çok aşamalı test uygulamalarına ilişkin literatürdeki çalışmalar incelendiğinde Yan vd. (2014a) 250 örneklem üzerinde çalışırken 5000 civarında büyük örneklemler üzerinde çalışılan araştırmalarda mevcuttur (Dallas, 2014; Doğruöz, 2018; Hambleton ve Xing, 2006; Sari, 2016; Wang, 2017; Yang, 2016).

Ek olarak bu araştırmada ortalaması 0, standart sapması 1 olan normal dağılımdan (N[0,1]) elde edilen 8100 simülasyon simüle edilmiştir. Ayrıca Magis vd. (2017) yaptıkları çalışmalarında çok aşamalı test uygulamalarına yönelik madde tepki kuramı yaklaşımında Fisher bilgisine dayalı olarak bir sonraki modülü seçmeyi önermektedirler. Bu araştırmada da benzer şekilde modül seçiminde MFI yöntemi kullanılmış ve her koşul için 100 tekrar yapılmıştır.

Veri Analizi

Bu araştırmada çok aşamalı test uygulaması için MSTgen programı (Han, 2013) kullanılmıştır. Tüm koşullar için 100 replikasyon sonucunda elde edilmiş olan "tahmin edilen yetenek parametre" değerleri ile "gerçek yetenek parametre" değerleri arasındaki farklılığın (hatanın) değerlendirilmesinde yani modellerin öngörü doğruluğunun (tahmin performansının) ölçümünde kullanılan: RMSE, SEE, BIAS ve MAE değerleri hesaplanmıştır.

BIAS: Bireye ait tahmin edilen yetenek düzeyi ile gerçek yetenek düzeyi arasındaki ortalama farklılığın istatistiğidir. Yanlılık değeri pozitif ya da negatif olabilir ve bu değerin sıfıra yakın olması daha yüksek doğruluk için gereklidir. *RMSE:* Bireye ait tahmin edilen yetenek düzeyi ile gerçek yetenek düzeyi arasındaki mutlak farklılığa ilişkin istatistiktir. *MAE:* Bireye ait tahmin edilen yetenek düzeyi ile gerçek yetenek düzeyi arasındaki mutlak farkın ortalamasıdır ve hatanın büyüklüğüne ilişkin tahmin sağlar. *SEE:* Tahmin edilen yetenek düzeyi ile gerçek yetenek düzeyi arasındaki farklılığa ilişkin standart hata değerleridir. MAE ve SEE'nin incelenmesinde de değerlerin sıfıra yakın olması hatanın daha az olmasının bir göstergesidir. Verilerin değerlendirilmesinde kullanılan istatistiklere ilişkin formüller Eşitlik 1, 2, 3 ve 4'te sunulmuştur.

$$\text{BIAS} = \frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)}{n} \qquad \text{Eşitlik (1)}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2}{n}} \qquad \text{Eşitlik (2)}$$

$$\text{MAE} = \frac{\sum_{i=1}^{n}|\hat{\theta}_i - \theta_i|}{n} \qquad \text{Eşitlik (3)}$$

$$SEE = \sqrt{\frac{\sum_{i=1}^{n}(\theta_i - \hat{\theta}_i)^2}{n-(k+1)}} \quad \text{Eşitlik (4)}$$

Eşitlik 1, 2, 3 ve 4'te bulunan gösterimlerde; "n": veri kümesindeki gözlem sayısı; "$\hat{\theta}_i$" bireye ait tahmin edilen yetenek düzeyi ve bireye ait gerçek yetenek düzeyi "θ_i"'; "$n-(k+1)$" serbestlik derecesidir.

Ayrıca modül uzunluğu, örneklem büyüklüğü ve kestirim yöntemlerinin RMSE, BIAS, MAE ve SEE değerleri üzerinde istatistiksel olarak anlamlı etkisinin olup olmadığı ANOVA testi ile incelenmiştir. Ardından farklılıkların hangi koşullar arasında olduğunu tespit edebilmek için ise "Tukey" çoklu karşılaştırma testi uygulanmıştır. Gruplar arasındaki farklılıklara ilişkin etki büyüklüğü değerleri de hesaplanarak yorumlanmıştır. Etki büyüklüğünün yorumlanmasında ise Cohen'in (1988) önerdiği ölçütler dikkate alınmıştır (küçük etki büyüklüğü: 0,01; orta etki büyüklüğü 0,06; büyük etki büyüklüğü 0,14).

Bulgular

Araştırmada kullanılan farklı simülasyon koşulları; panel deseni (1-3, 1-2-3 ve 1-3-3), modül uzunluğu (6, 12 ve 18), örneklem büyüklüğü (300, 1000 ve 3000) ve yetenek parametresi kestirim yöntemi (EAP, MAP ve MLEF) üretilen çok aşamalı bireye uyarlanmış test verisine ait RMSE, SEE, BIAS ve MAE değerleri Tablo 4'te sunulmuştur.

Tablo 4 incelendiğinde, en yüksek RMSE değerinin (1,30) 3000 örneklem büyüklüğünde, 6 modül uzunluğunda, 1-3 panel deseninde ve MLEF kestiriminde olduğu; ikinci yüksek RMSE değerinin (1,25) 1000 örneklem büyüklüğünde, 6 modül uzunluğunda, 1-3 panel deseninde ve MLEF kestiriminde olduğu görülmüştür. En düşük RMSE değerlerinin ise 300 örneklem büyüklüğünde, 12 modül uzunluğunda, 1-3-3 panel deseninde EAP (0,47), MLEF (0,47) ve MAP (0,48) kestirimlerinde olduğu bulunmuştur. Tüm koşullar açısından incelendiğinde ise modül uzunluğu arttıkça genellikle RMSE değerlerinin azaldığı görülmektedir. Farklı panel deseni koşullarına göre RMSE değerlerinin sırasıyla 1-3 panel deseninde en yüksek değerleri aldığı, ardından 1-2-3 panel deseninde ikinci en yüksek değerleri aldığı, en düşük RMSE değerlerinin ise 1-3-3 panel desenindeki koşullar için elde edildiği görülmüştür.

Tablo 4'da yer alan SEE ile ilgili bulgular incelendiğinde ise en yüksek SEE değerinin (5,52) 1000 örneklem büyüklüğünde, 6 modül uzunluğunda, 1-3 panel deseninde ve MAP kestiriminde olduğu; ikinci yüksek SEE değerinin (5,49) 1000 örneklem büyüklüğünde, 6 modül uzunluğunda, 1-3 panel deseninde ve EAP kestiriminde olduğu görülmüştür. En düşük SEE değerlerinin (0,41) ise 3000 örneklem büyüklüğünde, 12 modül uzunluğunda, 1-3-3 panel deseninde hem EAP hem de MAP kestiriminde olduğu bulunmuştur. Tüm koşullar açısından incelendiğinde ise modül uzunluğu arttıkça genellikle SEE değerlerinin azaldığı görülmüştür. Farklı panel desenlerinde ise SEE değerlerinin sırasıyla genellikle 1-3 panel deseninde en yüksek değerleri aldığı, ardından 1-2-3 panel deseninde ikinci en yüksek değerleri aldığı ve en düşük SEE değerlerinin 1-3-3 panel desenindeki koşullarda elde edildiği görülmüştür.

Tablo 4'te sıfıra en uzak BIAS değerlerinin (-0,32) 300 örneklem büyüklüğünde, 6 modül uzunluğunda, 1-2-3 panel deseninde hem EAP hem de MAP kestiriminde olduğu tespit edilmiştir. Sıfıra en yakın BIAS değerlerinin ise (0,01) 3000 örneklem büyüklüğünde, 12 modül uzunluğunda, 1-3-3 panel deseninde hem EAP hem de MAP kestiriminde olduğu görülmüştür.

Tablo 4'teki MAE ile ilgili bulgular incelendiğinde ise en yüksek MAE değerinin (0,95) 3000 örneklem büyüklüğünde, 6 modül uzunluğunda, 1-3 panel deseninde ve MLEF kestiriminde olduğu; ikinci en yüksek MAE değerinin ise (0,90) 1000 örneklem büyüklüğünde ve yine 6 modül uzunluğunda, 1-3 panel deseninde MLEF kestiriminde olduğu görülmüştür. En düşük MAE değerinin (0,29) ise 300 örneklem büyüklüğünde, 18 modül uzunluğunda, 1-3-3 panel deseninde EAP ve MAP kestiriminde olduğu bulunmuştur. Tabloda yer alan MAE değerleri genel olarak incelendiğinde modül uzunluğu arttıkça azaldığı görülmüştür. Farklı panel deseni koşullarında MAE değerlerinin sırasıyla 1-3 panel deseninde en yüksek değerleri aldığı ardından 1-2-3 panel deseninde ikinci en yüksek değerleri aldığı; en düşük MAE değerlerinin ise 1-3-3 panel desenindeki koşullarda elde edildiği görülmüştür. 300, 1000 ve 3000 örneklem büyüklükleri için EAP, MAP ve MLEF yetenek parametresi kestirim yöntemlerinden elde edilen RMSE, SEE, BIAS ve MAE değerleri Şekil 4, 5, 6 ve 7'de sunulmuştur.

Tablo 4

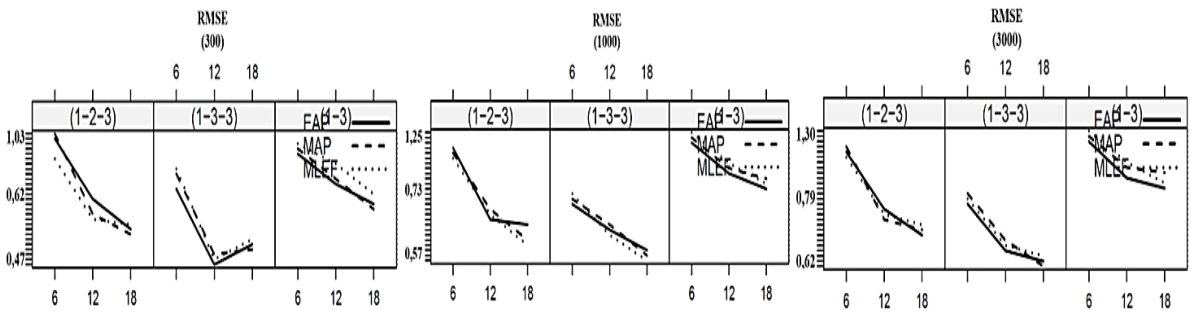
Yetenek Parametresine Ait RMSE, SEE, BIAS ve MAE Değerleri

			RMSE			SEE			BIAS			MAE		
			PD			PD			PD			PD		
AEM	SS	ML	1-3	1-2-3	1-3-3	1-3	1-2-3	1-3-3	1-3	1-2-3	1-3-3	1-3	1-2-3	1-3-3
EAP	300	6	0,92	0,99	0,71	4,97	3,13	1,37	-0,07	-0,32	-0,09	0,60	0,62	0,47
EAP	300	12	0,73	0,63	0,47	3,77	0,78	0,60	-0,19	-0,09	-0,06	0,44	0,36	0,31
EAP	300	18	0,62	0,56	0,51	1,09	0,64	0,58	-0,10	-0,11	-0,08	0,38	0,32	0,29
MAP	300	6	0,96	1,03	0,76	5,16	3,04	1,65	-0,10	-0,32	-0,10	0,62	0,65	0,49
MAP	300	12	0,73	0,61	0,48	3,55	0,74	0,62	-0,17	-0,07	-0,06	0,44	0,35	0,31
MAP	300	18	0,61	0,56	0,50	1,09	0,65	0,59	-0,09	-0,12	-0,08	0,37	0,32	0,29
MLEF	300	6	0,98	0,89	0,77	4,52	1,48	1,60	-0,24	-0,23	-0,12	0,66	0,58	0,49
MLEF	300	12	0,81	0,60	0,47	2,63	0,72	0,62	-0,21	-0,08	-0,07	0,49	0,35	0,30
MLEF	300	18	0,63	0,58	0,52	1,10	0,65	0,56	-0,10	-0,12	-0,09	0,38	0,34	0,30
EAP	1000	6	1,11	1,10	0,73	5,49	2,63	1,33	-0,19	-0,25	-0,05	0,79	0,77	0,54
EAP	1000	12	0,90	0,71	0,64	3,85	0,63	0,46	-0,14	-0,02	0,02	0,64	0,47	0,46
EAP	1000	18	0,82	0,64	0,58	0,94	0,49	0,47	-0,05	-0,05	-0,05	0,59	0,43	0,39
MAP	1000	6	1,12	1,08	0,75	5,52	2,38	1,52	-0,18	-0,22	-0,04	0,80	0,76	0,54
MAP	1000	12	0,91	0,72	0,64	3,83	0,66	0,47	-0,13	-0,03	0,02	0,64	0,48	0,46
MAP	1000	18	0,83	0,63	0,57	0,98	0,46	0,46	-0,05	-0,03	-0,04	0,59	0,42	0,39
MLEF	1000	6	1,25	1,00	0,79	4,28	1,23	1,74	-0,25	-0,15	-0,07	0,90	0,71	0,56
MLEF	1000	12	0,99	0,72	0,64	2,95	0,63	0,45	-0,20	-0,02	0,03	0,70	0,48	0,46
MLEF	1000	18	0,82	0,61	0,57	0,93	0,45	0,45	-0,04	-0,03	-0,05	0,59	0,41	0,39
EAP	3000	6	1,15	1,06	0,79	4,80	2,33	1,28	-0,21	-0,22	-0,09	0,82	0,75	0,56
EAP	3000	12	0,88	0,75	0,65	2,86	0,61	0,41	-0,11	-0,04	0,01	0,64	0,51	0,48
EAP	3000	18	0,87	0,68	0,64	0,88	0,44	0,44	-0,07	-0,02	-0,03	0,65	0,46	0,45
MAP	3000	6	1,17	1,08	0,80	4,86	2,04	1,51	-0,21	-0,22	-0,08	0,83	0,75	0,57
MAP	3000	12	0,89	0,74	0,65	2,81	0,59	0,41	-0,10	-0,03	0,01	0,65	0,51	0,48
MAP	3000	18	0,89	0,69	0,62	0,95	0,44	0,43	-0,08	-0,02	-0,02	0,66	0,47	0,44
MLEF	3000	6	1,30	1,01	0,79	4,06	1,24	1,50	-0,28	-0,18	-0,08	0,95	0,71	0,57
MLEF	3000	12	0,99	0,75	0,65	2,39	0,60	0,42	-0,16	-0,04	0,02	0,72	0,51	0,48
MLEF	3000	18	0,87	0,70	0,64	0,90	0,44	0,43	-0,07	-0,02	-0,04	0,65	0,47	0,45

Not: AEM: Yetenek parametresi kestirim yöntemi; ML: Modül uzunluğu; PD: Panel deseni; SS: Örneklem büyüklüğü

Şekil 4

300, 1000 ve 3000 Örneklem İçin Yetenek Parametresi Kestirim Yöntemlerinden Elde Edilen RMSE Değerleri

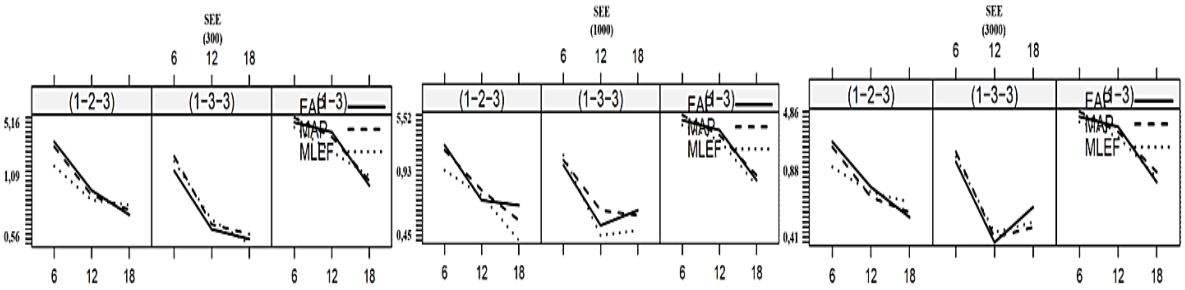


Şekil 4 incelendiğinde 300, 1000 ve 3000 örneklem büyüklüklerinin hepsinde modül uzunluğu arttıkça RMSE değerlerinin genel olarak düşme eğiliminde olduğu görülmektedir. Fakat 300 örneklem büyüklüğü ve 1-3-3 panel deseni için EAP, MAP ve MLEF yöntemlerinin üçünden de elde edilen RMSE değerlerinin 12 modül

uzunluğunda en düşük değerleri aldığı ve modül uzunluğu 18'e çıktığında RMSE değerinde bir miktar artış olduğu görülmektedir. Panel desenlerine göre incelendiğinde ise en yüksek RMSE değerlerinin 1-3 panel deseninde ([0,61- 1,30] aralığında) ardından 1-2-3 ([0,56- 1,10] aralığında) olduğu, en düşük ise 1-3-3 panel deseninde ([0,47-0,80] aralığında) elde edildiği açıkça görülmektedir. Yetenek parametresi kestirim yöntemleri açısından bulgular ele alındığında, koşuldan koşula değişmesine rağmen benzer koşullarda EAP ve MAP'ın genellikle benzer sonuçlar verdiği görülmüştür. Ayrıca RMSE değerlerinin 300 örneklem büyüklüğü için [0,47-1,03] aralığında, 1000 örneklem büyüklüğü için [0,57-1,25] aralığında ve 3000 örneklem büyüklüğü için [0,62-1,30] aralığında değerler aldığı saptanmıştır. 1000 ve 3000 örneklem büyüklüklerinden elde edilen RMSE değerlerinin ise genellikle benzer olduğu, 300 örneklem büyüklüğünde elde edilen değerlerin genellikle daha düşük olduğu sonucuna ulaşılmıştır.

Şekil 5

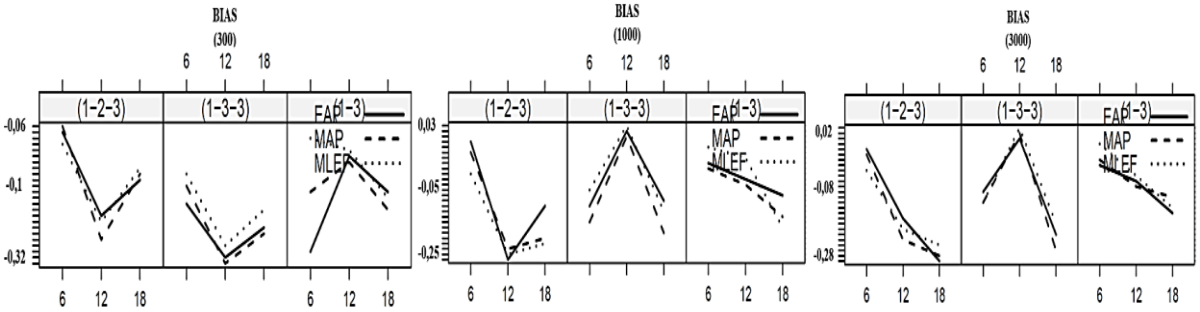
300, 1000 ve 3000 Örneklem İçin Yetenek Parametresi Kestirim Yöntemlerinden Elde Edilen SEE Değerleri



Şekil 5 incelendiğinde 300, 1000 ve 3000 örneklem büyüklüklerinin hepsinde modül uzunluğu arttıkça SEE değerlerinin genel olarak düşme eğiliminde olduğu saptanmıştır. Fakat 1000 ve 3000 örneklem büyüklüğü ve 1-3-3 panel deseni için EAP, MAP ve MLEF yöntemlerinin üçünden de elde edilen SEE değerlerinin 12 modül uzunluğunda daha düşük değerleri aldığı ve modül uzunluğu 18'e çıktığında SEE değerinde bir miktar artış olduğu tespit edilmiştir. Bu durumda özellikle EAP yöntemindeki artışın diğer iki yöntemdeki artışa göre nispeten fazla olduğu görülmektedir. Panel desenlerine göre incelendiğinde ise en yüksek SEE değerlerinin 1-3 panel deseninde ([0,88-5,52] aralığında) ardından 1-2-3 ([0,44-3,13] aralığında) olduğu, en düşük ise 1-3-3 panel deseninde ([0,41-1,74] aralığında) elde edildiği açıkça görülmektedir. Yetenek parametresi kestirim yöntemleri açısından bulgular ele alındığında, koşuldan koşula değişmesine rağmen benzer koşullarda EAP ve MAP'ın genellikle benzer sonuçlar verdiği görülmüştür. SEE değerlerinin 300 örneklem büyüklüğü için [0,56-5,16] aralığında, 1000 örneklem büyüklüğü için [0,45-5,52] aralığında ve 3000 örneklem büyüklüğü için [0,41-4,86] aralığında değerler aldığı saptanmıştır. Dolayısıyla 3000 örneklem büyüklüğünden elde edilen SEE değerlerinin genellikle daha düşük olduğu sonucuna ulaşılmıştır. Ayrıca üç örneklem büyüklüğünden elde edilen SEE değerleriyle elde edilen performansların ise koşullara göre farklılaşabildiği söylenebilir.

Şekil 6

300, 1000 ve 3000 Örneklem İçin Yetenek Parametresi Kestirim Yöntemlerinden Elde Edilen BIAS Değerleri

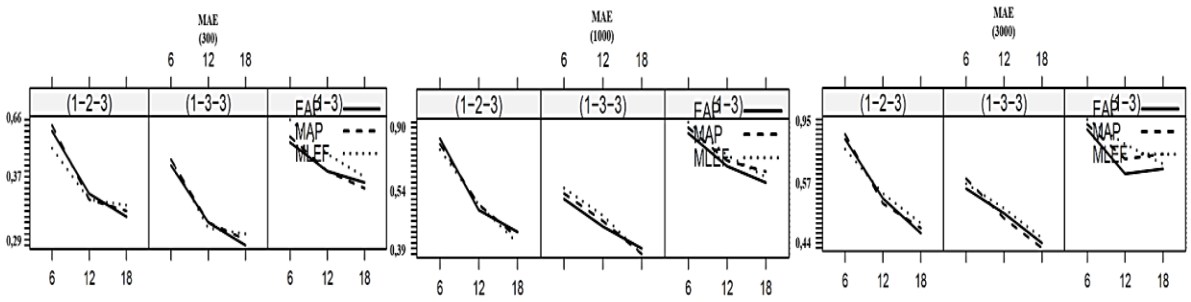


Şekil 6'da BIAS değerleri panel desenlerine göre incelendiğinde 1-3 panel deseninde [-0,28-(-0,04)] aralığında, 1-2-3 panel deseninde [-0,32-(-0,02)] aralığında ve 1-3-3 panel deseninde [-0,12-(0,03)] aralığındadır. BIAS

değerlerinin 300 örneklem büyüklüğü için [-0,32-(-0,06)] aralığında, 1000 örneklem büyüklüğü için [-0,25-(0,03)] aralığında ve 3000 örneklem büyüklüğü için [-0,28-(0,02)] aralığındadır.

Şekil 7

300, 1000 ve 3000 Örneklem İçin Yetenek Parametresi Kestirim Yöntemlerinden Elde Edilen MAE Değerleri



Şekil 7 incelendiğinde 300, 1000 ve 3000 örneklem büyüklüklerinin hepsinde modül uzunluğu arttıkça MAE değerlerinin genel olarak azalma eğiliminde olduğu saptanmıştır. Fakat sadece 3000 örneklem büyüklüğü ve 1-3 panel deseninde EAP ve MAP yöntemlerinden elde edilen MAE değerlerinin 12 modül uzunluğunda en düşük değeri aldığı ve modül uzunluğu 18'e çıktığında MAE değerlerinde bir miktar artış olduğu tespit edilmiştir. Panel desenlerine göre incelendiğinde ise MAE değerlerinin en yüksek 1-3 panel deseninde ([0,37-0,95] aralığında) ardından 1-2-3 panel deseninde ([0,32-0,77] aralığında) olduğu, en düşük ise 1-3-3 panel deseninde ([0,29-0,57] aralığında) elde edildiği açıkça görülmektedir. Yetenek parametresi kestirim yöntemleri açısından bulgular ele alındığında, koşuldan koşula değişmesine rağmen benzer koşullarda EAP ve MAP'ın genellikle benzer sonuçlar verdiği görülmüştür. MAE değerlerinin 300 örneklem büyüklüğü için [0,29-0,66] aralığında, 1000 örneklem büyüklüğü için [0,39-0,90] aralığında ve 3000 örneklem büyüklüğü için [0,44-0,95] aralığında değerler aldığı saptanmıştır. Dolayısıyla 300 örneklem büyüklüğünde elde edilen MAE değerlerinin diğer örneklem büyüklüklerine göre daha düşük olduğu sonucuna ulaşılmıştır. Ayrıca üç örneklem büyüklüğünden elde edilen SEE değerleriyle elde edilen performansların koşullara göre farklılaşabildiği söylenebilir.

Yetenek parametresi kestirimlerinin değerlendirilmesinde kullanılan RMSE, SEE ve MAE bulguları üzerindeki modül uzunluğu, örneklem büyüklüğü ve kestirim yöntemlerinin istatistiksel olarak anlamlı etkisinin olup olmadığı ANOVA testi ile incelenmiştir. RMSE, SEE ve MAE değerlerine ilişkin ANOVA sonuçları, etki büyüklükleri ve Tukey testi sonuçları Tablo 5 ve 6'da sunulmuştur. ANOVA, grupların ortalamaları arasındaki varyansın istatistiksel olarak anlamlı olup olmadığını belirlemek için kullanılan bir yöntemdir. BIAS matematiksel hesaplamasından (bk. Eşitlik 1) kaynaklı olarak negatif değerler de alabilir ve bu yüzden ANOVA testi sonuçlarının hatalı yorumlanmasına sebep olabileceğinden ötürü BIAS Tablo 5 ve 6'ya dâhil edilmemiştir. Ayrıca BIAS -1 ve +1 arasında değerler almakta ve bu değerlerin sıfıra yakın olması ölçme doğruluğuna işaret etmektedir. Araştırma kapsamında ele alınan diğer RMSE, SEE ve MAE değerleri ise sadece pozitif değerler aldığından dolayı ANOVA analizi yapılmasına uygundur.

Tablo 5

RMSE, SEE ve MAE Değerlerine İlişkin ANOVA Sonuçları

	RMSE			SEE			MAE		
	sd	*F*	η^2	*sd*	*F*	η^2	*sd*	*F*	η^2
AEM	2	0,12	-	2	0,45	-	2	0,16	-
Ö	2	7,13	-	2	0,31	-	2	19,04*	0,24
ML	2	30,79*	0,47	2	17,08*	0,38	2	32,03*	0,41
AEM*Ö	4	0,01	-	4	0,01	-	4	0,01	-
AEM*ML	4	0,04	-	4	0,16	-	4	0,03	-
Ö*ML	4	0,12	-	4	0,03	-	4	0,14	-
AEM*Ö*ML	8	0,02	-	8	0,00	-	8	0,01	-

NOT: AEM: Yetenek parametresi kestirim yöntemi; ML: Modül uzunluğu; Ö: Örneklem büyüklüğü,
**p < 0,05*

Tablo 5'e göre yetenek parametresi kestiriminin değerlendirilmesinde kullanılan RMSE ve SEE değerlerinin sadece modül uzunluğuna göre anlamlı şekilde farklılaştığı görülmektedir (RMSE: $F_{2;\,54} = 30{,}79$, $p < 0{,}05$; SEE: $F_{2;\,54} = 17{,}08$, $p < 0{,}05$). MAE değerleri ise hem örneklem büyüklüğüne hem de modül uzunluğuna göre anlamlı şekilde farklılaşmaktadır (Ö: $F_{2;\,54} = 19{,}04$, $p < 0{,}05$; ML: $F_{2;\,54} = 32{,}03$, $p < 0{,}05$). Buna rağmen RMSE, SEE ve MAE değerlerinin kestirim yöntemlerine (EAP, MAP ve MLEF) göre anlamlı şekilde farklılaşmadığı tespit edilmiştir ($p > 0{,}05$). Farklılığın hangi modül uzunlukları ve hangi örneklem büyüklükleri arasında olduğunu tespit etmek için yapılan Tukey testi sonuçları incelenmiş ve Tablo 6 ve Tablo 7'de sunulmuştur.

Tablo 6

Modül Uzunluklarına İlişkin Tukey Testi Sonuçları

	RMSE			SEE			MAE		
ML	Ortalama farkı	*sd*	*t*	Ortalama farkı	*sd*	*t*	Ortalama farkı	*sd*	*t*
6-12	0,25	54	5,97*	1,39	54	3,70	0,18	54	6,02*
6-18	0,31	54	7,40*	2,17	54	5,77*	0,23	54	7,58*
12-18	0,06	54	1,43	0,78	54	2,07	0,05	54	1,56

NOT: ML: Modül uzunluğu
**p < 0,05*

Tablo 6'daki bulgulara göre RMSE ve MAE değerlerindeki tüm farklılıkların 6-12 ve 6-18 modül uzunlukları arasında olduğu görülmektedir. Ortalama fark değerleri incelendiğinde ise bu farklılıkların 6 modül uzunluğu yerine 12 ve 18 modül uzunluklarının lehine olduğu tespit edilmiştir. Yine tabloda yer alan SEE değerlerindeki tüm farklılıkların ise sadece 6-18 modül uzunlukları arasında olduğu ve ortalama fark değerlerine bakıldığında da bu farklılığın 18 modül uzunluğu lehine olduğu saptanmıştır. Ayrıca Tablo 5'te görüldüğü gibi elde edilen etki büyüklüğü değerleri $\eta^2 = 0{,}47$ (RMSE), $\eta^2 = 0{,}38$ (SEE) ve $\eta^2 = 0{,}41$ (MAE)'dir. Cohen'in (1988) önerdiği ölçütler dikkate alındığında, modül uzunluğunun RMSE, SEE ve MAE değerleri üzerinde büyük etkiye sahip olduğu görülmektedir.

Tablo 7

Örneklem Büyüklüklerine İlişkin Tukey Testi Sonuçları

Örneklem büyüklüğü	MAE		
	Ortalama farkı	*SD*	*t*
1000-3000	-0,03	54	-1,06
1000-300	0,14	54	4,73*
3000-300	0,17	54	5,80*

**p < 0,05*

Tablo 7'ye göre ise MAE değerlerindeki farklılıkların 1000-300 ve 3000-300 örneklem büyüklükleri arasında olduğu tespit edilmiştir. Ortalama fark değerleri incelendiğinde ise bu farklılığın 1000 ve 3000 örneklem büyüklüğü yerine 300 örneklem büyüklüğü lehine olduğu saptanmıştır. Tablo 5'te görüldüğü gibi örneklem büyüklüğünün MAE değeri üzerinde büyük etkiye sahip olduğu görülmektedir ($\eta^2 = 0{,}24$).

Tartışma, Sonuç ve Öneriler

Bu araştırmada farklı koşullar (panel deseni, modül uzunluğu, örneklem büyüklüğü, yetenek parametresi kestirimi) altında elde edilen sonuçlara dayalı olarak çok aşamalı test performanslarının RMSE, SEE, BIAS ve MAE açısından karşılaştırılmasına odaklanılmıştır. Bu sayede yapılan kestirimlerin hangi koşullar altında daha iyi sonuç verdiği araştırılmıştır. 81 farklı koşul altında yapılan kestirimlerle elde edilen araştırma sonuçları panel deseni, modül uzunluğu, örneklem büyüklüğü ve yetenek parametresi kestirim yöntemleri açısından aşağıda sunulmuştur.

Araştırma bulguları modül uzunluklarının, kestirimlerden elde edilen RMSE, MAE ve SEE değerlerini etkilediğini göstermiştir. Daha açık ifadeyle araştırma kapsamında ele alınan üç farklı modül uzunluklarında;

kısa modül uzunluğu (6) orta (12) ve uzun modül uzunluğuna (18) göre daha yüksek RMSE, MAE ve SEE değerleri üretmiştir. Dolayısıyla modül uzunluğu arttıkça RMSE, MAE ve SEE değerlerinin genellikle azaldığı sonucuna ulaşılmıştır. Bu durum test uzunluğu arttıkça elde edilen hata değerlerinin azalması şekilde de yorumlanabilir. Elde edilen bu bulgu literatürde yer alan ve test uzunluğu arttıkça RMSE ve standart hata değerlerinin azaldığı bulgusuna ulaşan çalışma sonuçlarıyla paraleldir (Boztunç Öztürk, 2019; Doğruöz, 2018; Hembry, 2014; Sari, 2016; Yang, 2016). Ayrıca Park (2015) çok aşamalı test uygulaması üzerine yaptığı çalışmasında testin uzun olmasıyla ölçme doğruluğunun daha yüksek olduğu sonucuna ulaşmıştır. Sari (2016) içerik sayısı, test yönetimi ve test uzunluğunun test birleştirme yöntemi üzerindeki etkisini incelediği çalışmasında ise yalnızca test uzunluğu faktörünün RMSE değerleri üzerinde anlamlı etkiye sahip olduğu sonucuna ulaşmıştır. Araştırmadan elde edilen ve literatürün de desteklediği gibi modül uzunluğunun artmasıyla RMSE, MAE ve standart hata değerlerinin azalmasının sebebi, kısa testlerden oluşan çok aşamalı testlerin ölçme hassasiyetinin daha düşük olmasından kaynaklı olabilir. Dolayısıyla testlerdeki madde sayısının artması, test puanlarının güvenirliğini ve dolayısıyla ölçme doğruluğunu artırmaktadır (Crocker ve Algina, 1986). Modül uzunluğunun arttırılması teoride önerilse de elde edilen sonuçların uygulanabilirliği ve kullanışlık dikkate alınarak çok aşamalı testlerde en uygun modül uzunluğu ile çalışılmalı ve mümkünse 12 ve üzeri modül uzunluğunun kullanılması önerilmektedir.

Benzer koşullarda örneklem büyüklüğü arttıkça SEE değerlerinde genel olarak bir miktar azalma tespit edilmiştir. Buna rağmen RMSE ve MAE değerlerinin örneklem büyüklüğü arttıkça bir miktar artma eğiliminde olduğu, yanlılık değerlerinin ise örneklem büyüklüğü açısından farklı koşullarda farklı sonuçlar verdiği gözlemlenmiştir. Yan vd. (2014a) ise çok aşamalı test uygulamasının küçük örneklemlerde iyi performans gösterdiğini yaptıkları çalışmada belirtmişlerdir. Doğruöz (2018) test birleştirme yöntemlerinin çok aşamalı test uygulaması üzerindeki etkisini incelediği çalışmasında örneklem büyüklüklerinin (250-2000) hangi yöntemde daha etkili olduğu konusunda net bir sonuca ulaşamamıştır. Yine Doğruöz'ün (2018) yaptığı aynı çalışmada örneklem büyüklüğünün 250'den 2000'e artmasının 1-2-3 panel deseninde RMSE değerlerini değiştirmezken 1-2 ve 1-2-2 panel desenlerinde ortalama RMSE değerlerini küçülttüğünü bulmuştur. Buna rağmen Doğruöz (2018) örneklem büyüklüğündeki artışın tüm panel desenlerine ait yanlılık değerlerinde çok küçük miktarda düşüşe neden olduğunu gözlemlemiştir. Bu çalışmada 300 örneklem büyüklüğünde elde edilen RMSE ve MAE değerlerinin 1000 ve 3000 örneklem büyüklüğünden genel olarak iyi performans verdiği sonucuna ulaşılmıştır. Tam aksine 300 örneklemde SEE değerlerinin, 1000 ve 3000 örnekleme göre genel olarak daha düşük performans sağladığı söylenebilir. Bu sonuçtan yola çıkarak küçük örneklemlerde bile çok aşamalı testlerin uygulanabileceği sonucuna varılabilir. Alanyazın incelendiğinde 250 örneklem ile yapılan çok aşamalı test araştırmalarının da olduğu görülmektedir (örn., Yan vd., 2014a).

Araştırmadan elde edilen sonuçlar doğrultusunda RMSE, MAE ve SEE 1-3 panel deseninde en yüksek değerlere sahip olduğu ardından 1-2-3 panel deseninde ikinci en yüksek değerleri aldığı, 1-3-3 panel deseninde ise en düşük değerleri aldığı görülmüştür. Araştırma sonucunda genellikle sıfıra en yakın yanlılık değerlerinin 1-3-3 panel deseninde olduğu tespit edilmiştir. Dolayısıyla 1-3-3 panel deseninin 1-3 ve 1-2-3 panel desenlerinden daha iyi performans sergilediği söylenebilir. Aşama sayısının artmasının hata değerlerinin azalmasına neden olduğu söylenebilir. Patsula (1999) yaptığı çalışmada aşama sayısının artmasının yetenek parametresi kestirimin doğruluğunu ve panel deseninin etkililiğini artırdığı sonucuna ulaşmıştır. Ayrıca aşama sayısının ikiden üçe çıkarılmasının yetenek parametresi kestirimindeki hata miktarını azalttığını da çalışmasında belirtmiştir. Benzer şekilde 1-3 ve 1-3-3 panel desenlerinin performanslarını karşılaştıran Boztunç Öztürk (2019), iki aşamalı yapıdan üç aşamalı yapıya geçildiğinde RMSE değerlerinin düştüğü, dolayısıyla üç aşamalı panel tasarımının ölçme hassasiyeti açısından genel olarak daha iyi sonuçlar verdiği sonucuna ulaşmıştır. Park (2015) panel deseninin ölçme doğruluğunu üzerindeki etkisini araştırdığı çalışmasında 1-3-3 panel deseninin 1-2-2'den daha iyi performans sergilediğini tespit etmiştir. Doğruöz (2018) ise farklı test birleştirme yöntemlerine göre karşılaştırma yaptığı simülasyon çalışmasında 1-2 panel deseninden 1-2-2 ve 1-2-3 panel desenine geçişte ortalama RMSE değerlerinde düşüş olduğu sonucuna ulaşmıştır. Sari (2016) ise farklı panel desenlerini (1-3 ve 1-3-3) ele aldığı çalışmasında test deseninin çalışma sonuçlarını etkilediğini tespit etmiştir. Alanyazındaki bu çalışmalar incelendiğinde ise bu araştırmada kullanılan panel desenlerinin diğer çalışmalardan farklılaşmasına rağmen araştırma sonuçları alanyazındaki çalışmalarla paralellik göstermektedir. Elde edilen sonuçlar doğrultusunda araştırmacılara ve çok aşamalı test uygulayıcılarına aşama sayısının daha fazla olduğu 1-3-3 panel desenini kullanmaları önerilebilir.

Araştırmada yetenek parametresi kestirim yöntemlerine göre elde edilen RMSE, SEE, BIAS ve MAE değerlerinin koşuldan koşula değişmesine rağmen benzer koşullarda EAP ve MAP'ın genellikle benzer sonuçlar verdiği görülmüştür. Yetenek kestirim yöntemlerine ilişkin RMSE değerleri incelendiğinde 1-3 panel deseninde

en düşük değerlerin genellikle EAP kestiriminden elde edildiği bulunmuştur. Bu araştırmadan elde edilen sonuçlar doğrultusunda benzer şekilde Han'ın (2016) MLE, MLET, MLEF, MAP ve EAP yöntemlerini karşılaştırdığı araştırmasında benzer koşullarda MAP ve EAP yöntemlerinin neredeyse aynı tahmin yanlılığına sahip olduğunu belirtmiştir. Benzer şekilde Ertaş Polat (2022) farklı koşullardan (modül uzunluğu, yönlendirme yöntemi ve kestirim yöntemi) elde ettiği yetenek kestirimlerini (EAP ve MLE) karşılaştırdığı çalışmasında da EAP yöntemiyle yapılan kestirimlerin tüm koşullar için ortalama hata değerinin MLE'ye göre daha düşük olduğunu belirtmiştir. Yine aynı çalışmada tümünün doğru veya tümünün yanlış yanıtlandığı durumlarda MLE yönteminin çalışmadığı ve uç değerlerde bulunan yetenek değerlerinde de bu koşulların oluşmasından kaynaklı olarak yapılan kestirim hatalarının yükselmesine neden olabileceği belirtilmiştir. MLE yönteminin bu tarz dezavantajları göz önünde bulundurularak mevcut çalışmada kestirim yöntemlerinden Han'ın (2016) önerdiği alternatif yaklaşımlardan MLEF yöntemi kullanılmıştır. Buna rağmen MLEF yönteminden elde edilen hata değerlerine ilişkin sonuçlarda, benzer koşullardaki EAP ve MAP yönteminden elde edilen hata değerlerinin biraz daha yüksek olduğu sonucuna ulaşılmıştır. Üç kestirim yöntemi arasından genellikle en düşük RMSE değerleri ise EAP yöntemi ile yapılan kestirimlerden elde edilmiştir. Bu doğrultuda bu çalışmanın koşulları dikkate alınarak benzer çalışmalarda kestirim yöntemlerinden EAP yönteminin seçilmesi önerilebilir.

Araştırmada değerlendirme kriterleri olarak RMSE, SEE, BIAS ve MAE değerleri ele alınmıştır. Değerlendirme kriterleri 81 simülasyon koşulu altında incelendiğinde RMSE ve MAE değerlerinin genellikle benzer sonuçlar verdiği görülmüştür. SEE değerlerinin ise RMSE ve MAE değerlerine göre daha yüksek değerler verdiği tespit edilmiştir. Literatürde yer alan çok aşamalı testlere yönelik yapılan çalışmalarda ve Ertaş Polat (2022)'ın da yapmış olduğu çalışmada belirtildiği gibi bu araştırma kapsamında da ele alınan koşullara yönelik bulgular doğrultusunda, her koşula uyan ve en iyi olan bir tane çok aşamalı test tasarımının olmadığı sonucuna ulaşılmıştır.

Bu araştırmanın çeşitli sınırlılıkları bulunmaktadır. Mevcut araştırma bir simülasyon çalışmasıdır. Öncelikle araştırmada verilerin üretilmesinde TIMSS 2015 matematik başarı testi parametreleri 300, 1000 ve 3000 örneklem için dikkate alınmıştır. Gerçek veriler üzerinden yetenek parametresi kestirim yöntemleri karşılaştırılabilir. Ayrıca araştırmada 81 simülasyon koşulu (3 örneklem büyüklüğü x 3 panel deseni x 3 modül uzunluğu x 3 yetenek parametresi kestirim yöntemi) alanyazın sonucunda karar verilerek incelenmiştir. Çok aşamalı testler üzerine yapılan çalışmaların sonuçlarına göre en iyi performansı sergileyen tek bir tasarımın olmadığı görülmektedir. Bu nedenle farklı simülasyon koşulları ele alınarak simülasyon çalışmaları gerçekleştirilebilir. Bunun yanı sıra araştırmada 100 replikasyon yapılmıştır. Farklı replikasyonların kestirimlere etkisi başka araştırmalarda incelenebilir. Ek olarak madde kullanım sıklığının kontrolü ve içerik dengelemeye yer verilmemiştir. Bu durumlar dikkate alınarak benzer koşullar üzerinde çalışmalar yapılabilir. Son olarak ise araştırmada her bir modüldeki madde sayıları birbirine eşit olarak alınmıştır. Benzer koşullarda modüldeki madde sayıları değiştirilerek farklı çalışmalar tasarlanabilir.

Kaynakça

Armstrong, R. D., & Roussos, L. (2005). *A method to determine targets for multi-stage adaptive tests*. Research Report 02-07. Newton, PA: School Admission Council.

Armstrong, R. D., Jones, D. H., Koppel, N. B., & Pashley, P. J. (2004). Computerized adaptive testing with multiple-form structures. *Applied Psychological Measurement*, *28*(3), 147–164. https://doi.org/10.1177/0146621604263652

Baker, F. B. (2001). *The basics of item response theory*. United States of America: ERIC Clearinghouse on Assessment and Evaluation.

Belov, D. I., & Armstrong, R. D. (2008). A monte carlo approach to the design, assembly, and evaluation of multistage adaptive tests. *Applied Psychological Measurement, 32*(2), 119–137. https://doi.org/10.1177/0146621606297308

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443–459. https://doi.org/10.1007/BF02293801

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*(4), 431–444. https://doi.org/10.1177/014662168200600405

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., & Hare, D. R. (2007). Automated simultaneous assembly of multistage testlets for a high-stakes licensing examination. *Educational and Psychological Measurement, 67*(1), 5–20. https://doi.org/10.1177/0013164406288162

Büyükkıdık, S. & Ayva Yörü, F. G. (2022, Eylül). *Çok aşamalı testlerin panel deseni, modül uzunluğu, örneklem büyüklüğü ve yetenek kestirim yöntemleri açısından karşılaştırılması* [Sözlü bildiri]. 8. Uluslararası Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi, İzmir.

Chen, L. Y. (2010). *An investigation of the optimal test design for multi-stage test using the generalized partical credit model* [Yayımlanmamış doktora tezi]. The University of Texas at Austin.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. baskı). Routledge.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart & Winston.

Dallas, A. (2014). *The effects of routing and scoring within a computer adaptive multi-stage framework.* [Yayımlanmamış doktora tezi]. The University of North Carolina.

Dallas, A., Wang, X., Furter, R., & Luecht, R. M. (2012, Nisan). *Item pool size, targeted item writing and panel replication strategies for a 1-3-3 multistage test design* [Sözlü bildiri]. National Council on Measurement in Education (NCME), Vancouver, BC.

Davey, T., & Lee, Y. H. (2011). *Potential impact of context effects on the scoring and equating of the multistage GRE® revised general test.* ETS Research Report Series, 2011(2), i–44. https://www.ets.org/research/policy_research_reports/publications/report/2011/itjm.html

Davis, L. L., & Dodd, B. G. (2003). Item exposure constraints for testlets in the verbal reasoning section of the MCAT. *Applied Psychological Measurement, 27*(5), 335–356. https://doi.org/10.1177/0146621603256804

De Ayala, R. J. (2009*). The theory and practice of item response theory*. New York: The Guilford Press.

Doğruöz, E. (2018). *Bireyselleştirilmiş çok aşamalı testlerin test birleştirme yöntemlerine göre incelenmesi* [Yayımlanmamış doktora tezi]. Hacettepe Üniversitesi.

Drasgow, F., & Mattern, K. (2006). New tests and new items: Opportunities and issues. D. Bartram & R. Hambleton (Haz.), Computer based testing and internet: Issues and advances içinde (s. 59–77). Educational testing service: London.

Edwards, M. C., Flora, D. B., & Thissen, D. (2012). Multistage computerized adaptive testing with uniform item exposure. *Applied Measurement in Education, 25*(2), 118–141. https://doi.org/10.1080/08957347.2012.660363

Erdem Kara, B. (2019). *Değişen madde fonksiyonu gösteren madde oranının bireyselleştirilmiş bilgisayarlı ve çok aşamalı testler üzerindeki etkisi* [Yayımlanmamış doktora tezi]. Hacettepe Üniversitesi.

Ertaş Polat, F. G. (2022). *Çok aşamalı bireye uyarlanmış testlerde farklı koşullardan elde edilen yetenek kestirimlerinin karşılaştırılması* [Yayımlanmamış doktora tezi]. Hacettepe Üniversitesi.

Hambleton, R. K., & Xing, D. (2006). Optimal and nonoptimal computerbased test designs for making pass-fail decisions. *Applied Measurement in Education, 19*(3), 221–239 https://doi.org/10.1207/s15324818ame1903_4

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

Han, K. C. T., & Guo, F. (2014). Multistage testing by shaping modules on the fly. D. Yan, A. A. von Davier, & C. Lewis (Haz.), *Computerized multistage testing: Theory and applications* içinde (s. 119–133). Chapman and Hall/CRC.

Han, K. T. (2013). MSTGen: Simulated data generator for multistage testing. *Applied Psychological Measurement, 37*(8), 666–668. https://doi.org/10.1177/0146621613499639

Hembry, I. F. (2014). *Operational characteristics of mixed format multistage tests using the 3PL testlet response theory model* [Yayımlanmamış doktora tezi]. The University of Texas at Austin.

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

International Association for the Evaluation of Educational Achievement. (2021). *TIMSS 2019 international database* [Veri seti]. TIMSS & PIRLS International Study Center. https://timss2019.org/international-database/?_gl=1*1gitpgj*_ga*OTg0NzE0MzYuMTY0NTk5NzE4MQ..*_ga_L2FMXN42HR*MTY0Njc3OTQ2OC41LjAuMTY0Njc3OTQ2OC4w

Jodoin, M. G. (2003). *Psychometric properties of several computer-based test designs with ideal and constrained item pool* [Yayımlanmamış doktora tezi]. University of Massachusetts-Amherst.

Jodoin, M. G., Zenisky, A., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test design for credentialing exams with multiple purposes. *Applied Measurement in Education*, *19*(3), 203–220. http://doi.org/10.1207/s15324818ame1903_3

Keng, L., & Dodd, B.G. (2009, Nisan). *A comparison of the performance of testlet based computer adaptive tests and multistage tests* [Sözlü bildiri]. National Council on Measurement in Education (NCME), San Diego, CA.

Kim, H., & Plake, B. S. (1993, Nisan). *Monte carlo simulation comparison of two-stage testing and computerized adaptive testing* [Sözlü bildiri]. National Council on Measurement in Education (NCME), Atlanta, GA.

Kim, J., Chung, H., Dodd, B. G., & Park, R. (2012). Panel design variations in the multistage test using the mixed-format tests. *Educational and Psychological Measurement*, *72*(4), 574–588. https://doi.org/10.1177/0013164411428977

Kim, S., Moses, T., & Yoo, H. H. (2015). A comparison of IRT proficiency estimation methods under adaptive multistage testing. *Journal of Educational Measurement, 52*(1), 70–79. https://onlinelibrary.wiley.com/doi/abs/10.1111/jedm.12063

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

Luecht, R. M. (2000, Nisan). *Implementing the Computer-Adaptive Sequantial Testing (CAST) framework to mass produce high quality computer adaptive and mastery tests* [Sözlü bildiri]. National Council on Measurement in Education (NCME), New Orleans, LA.

Luecht, R., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, *19*(3), 189–202. https://doi.org/10.1207/s15324818ame1903_2

Luecht, R. M., & Sireci, S. G. (2011). *A review of models for computer-based testing*. (Rapor No. RR-2011-12). College Board, New York. https://doi.org/10.1111/j.1745-3984.1998.tb00537.x

Luo, X., & Kim, D. (2018). A top-down approach to designing the computerized adaptive multistage test. *Journal of Educational Measurement*, *55*(2), 243–263. https://doi.org/10.1111/jedm.12174

Magis, D., Yan, D. & von Davier, A. A. (2017). *Computerized adaptive and multistage testing with R: Using packages catR and mstR*. Cham, Switzerland: Springer International Publishing.

Mason, B. J., Patry, M., & Bernstein, D. J. (2001). An examination of equivalence between non adaptive computer-based test and traditional testing. *Journal of Educational Computing Research 24*(1), 29–39. https://doi.org/10.2190/9EPM-B14R-XQWT-WVNL

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

Milli Eğitim Bakanlığı [MEB]. (2016). *TIMSS 2015 ulusal matematik ve fen bilimleri ön raporu 4. ve 8. sınıflar.* https://timss.meb.gov.tr/meb_iys_dosyalar/2022_03/07135609_TIMSS_2015_Ulusal_Rapor.pdf

Owen, R. J. (1975). A bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*(350), 351–356. https://doi.org/10.2307/2285821

Park, R. (2015). *Investigating the impact of a mixed-format item pool on optimal test designs for multistage testing* [Yayımlanmamış doktora tezi]. The University of Texas at Austin.

Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multistage testing* [Yayımlanmamış doktora tezi). The University of Massachusetts Amherst. https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=4283&context=dissertations_1

Reese, L. M., Schnipke, D. L., & Luebke, S. W. (1999). *Incorporating content constraints into a multi-stage adaptive testlet design.* Law School Admission Council Computerized Testing Report. LSAC Research Report Series.

Samejima, F. (1968). Estimation of latent ability using a response patterns of graded scores. *Psychometrika Monograph, 17*, i–169.https://doi.org/10.1002/j.2333-8504.1968.tb00153.x

Sarı, H. İ., Yahşi Sarı, H., & Huggins Manley, A. C. (2016). Computer adaptive multistage testing: Practical issues, challenges and principles. *Journal of Measurement and Evaluation in Education and Psychology*, *7*(2), 388–406. https://doi.org/10.21031/epod.280183

Sari, H. İ. (2016). *Examining content control in adaptive tests: Computerized adaptive testing vs. computerized multistage testing* [Yayımlanmamış doktora tezi]. University of Florida.

Schnipke, D. L., & Reese, L. M. (1999). *A comparison of testlet-based test designs for computerized adaptive testing.* (Rapor No: 97–01). ERIC Database.

Stark, S., & Chernyshenko, O. S. (2006). Multistage testing: Widely or narrowly applicable? *Applied Measurement in Education, 19*(3), 257–260. https://doi.org/10.1207/s15324818ame1903_6

Şahin, M. G. (2020). Analyzing different module characteristics in computer adaptive multistage testing. *International Journal of Assessment Tools in Education, 9*(2), 191–206. https://doi.org/10.21449/ijate.676947

Şahin, M. G., & Boztunç Öztürk, N. (2019). Analyzing the maximum likelihood score estimation method with fences in ca-MST. *International Journal of Assessment Tools in Education 6*(4), 555–567. https://dx.doi.org/10.21449/ijate.634091

van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. W. J. van der Linden, & C. A. W. Glas (Haz.), *Elements of adaptive testing* içinde (s. 3–30). New York: Springer.

Wang, K. (2017). *Fair comparison of the performance of computerized adaptive testing and multistage adaptive testing* [Yayımlanmamış doktora tezi]. Michigan State University.

Wang, T. H., Wang, K. H., Wang, W. L., Huang, S. C., & Chen, S. Y. (2004). Web-based assesment and test analyses (WATA) Q3 system: Development and evaluation. *Journal of Computer Assisted Learning, 20*(1), 59–71. https://doi.org/10.1111/j.1365-2729.2004.00066.x

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*(2), 109–135. https://doi.org/10.1111/j.1745-3984.1998.tb00530.x

Wang, X., Fluegge, L., & Luecht, R. M. (2012, Nisan). *A large-scale comparative study of the accuracy and efficiency of ca-MST panel design configurations* [Sözlü bildiri]. National Council on Measurement in Education (NCME), Vancouver, BC.

Warm, A. W. (1989). Weighted likelihood estimation of ability in item response theory with tests of finite length. *Psychometrika, 54*(3), 427–450. https://doi.org/10.1007/BF02294627

Weiss, D. J. (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing.* Academic Press: New York.

Weissman, A., Belov, D., & Armstrong, R. (2007). *Information-based versus number-correct routing in multistage classification tests*. (LSAC Research Report No:07–05). Newtown, PA.

Yan, D., Lewis, C., & von Davier, A. (2014a). Overview of computerized multistage tests. D. Yan, A. A. von Davier, & C. Lewis (Haz.), *Computerized multistage testing: Theory and applications* içinde (s. 3–20). London, England: Chapman & Hall.

Yan, D., von Davier, A. A., & Lewis, C. (Haz.). (2014b). *Computerized multistage testing: Theory and applications*. CRC Press.

Yang, L. (2016). *Enhancing item pool utilization when designing multistage computerized adaptive tests* [Yayımlanmamış doktora tezi]. Michigan State University.

Zenisky, A. L. (2004). *Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment* [Yayımlanmamış doktora tezi]. University of Massachusetts Amherst. https://scholarworks.umass.edu/dissertations/AAI3136800 adresinden erişilmiştir.

Zenisky, A., & Hambleton, R. (2014). Multistage test designs: Moving research results into practice. Yan, D., Von Davier, A., & Lewis, C. (Haz*.), Computerized multistage testing: Theory and applications*, içinde (s. 21–36). London, England: Chapman & Hall.

Zheng, Y. & Chang, H. H. (2014). Multistage testing, on-the-fly multistage testing, and beyond. Y. Cheng, & H. H. Chang (Haz.), *Advancing methodologies to support both summative and formative assessments* içinde (s. 21–40). Charlotte, NC: Information Age Publishing.

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

Zheng, Y., Nozawa, Y., Gao, X., & Chang, H. H. (2012). *Multistage adaptive testing for a large-scale classification test: Design, heuristic assembly, and comparison with other testing modes*. ACT Research Report Series, ACT.

Comparing Multi-Stage Tests under Different Conditions in Terms of Panel Design, Module Length, Sample Size and Ability Parameter Estimation Methods

Abstract

In this research, the performances of multi-stage tests under various simulation conditions have been compared in terms of evaluation criteria, including root mean square error (RMSE), standard error of estimate (SEE), bias, and mean absolute error (MAE). In the test simulation, 81 conditions (3x3x3x3) have been determined, including panel design (1-3, 1-2-3, 1-3-3), module length (6, 12, 18), sample size (300, 1000, 3000), and ability parameter estimation methods (expected a posteriori [EAP], maximum a posteriori [MAP], and maximum likelihood estimation with fences [MLEF]). The research findings indicate that RMSE and MAE values generally produce similar results, and measurement accuracy tends to increase with the lengthening of the module. Additionally, it was observed that RMSE, SEE, and MAE have the highest values in the 1-3 panel design and the lowest values in the 1-3-3 design. Researchers are recommended to conduct their studies using a 1-3-3 panel design, with a minimum module length of 12, and employing the EAP method.

Keywords: multi-stage test, panel design, module length, sample size, ability parameter estimation method