HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Eğitim Bilimleri Ana Bilim Dalı

Eğitimde Ölçme ve Değerlendirme Programı

SABİT VE ANINDA BİREYSELLEŞTİRİLMİŞ ÇOK AŞAMALI TESTLERİN KARŞILAŞTIRILMASI

Mahmut Sami YİĞİTER

Doktora Tezi

Ankara, 2023

Liderlik, araştırma, inovasyon, kaliteli eğitim ve değişim ile

Daha ileriye... En İyiye...

HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Eğitim Bilimleri Ana Bilim Dalı

Eğitimde Ölçme ve Değerlendirme Programı

SABİT VE ANINDA BİREYSELLEŞTİRİLMİŞ ÇOK AŞAMALI TESTLERİN KARŞILAŞTIRILMASI

COMPARISON OF FIXED AND ON-THE-FLY COMPUTERIZED MULTISTAGE TESTING

Mahmut Sami YİĞİTER

Doktora Tezi

Ankara, 2023

Kabul ve Onay

Eğitim Bilimleri Enstitüsü Müdürlüğüne,

Mahmut Sami YİĞİTER'in hazırladığı "Sabit ve Anında Bireyselleştirilmiş Çok Aşamalı Testlerin Karşılaştırılması" başlıklı bu çalışma jürimiz tarafından **Eğitim Bilimleri Ana Bilim Dalı, Eğitimde Ölçme ve Değerlendirme Bilim Dalında Doktora Tezi** olarak kabul edilmiştir.

Jüri Başkanı	İmza
Jüri Üyesi (Danışman)	İmza
Jüri Üyesi	İmza
Jüri Üyesi	İmza
Jüri Üyesi	İmza

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

Eğitim Bilimleri Enstitüsü Müdürü

Öz

Pek çok geniş ölçekli değerlendirmede Bireyselleştirilmiş Bilgisayarlı Testler (BBT) ve Bireyselleştirilmiş Çok Aşamalı Testler (BÇAT) gibi uyarlanabilir test yaklaşımları benimsenmiştir. BBT'nin yeteneği eksik ya da fazla kestirmesi, maddeler arası gezinmeye izin vermemesi, maddeyi atlamak için mutlaka cevap gerektirmesi ve senaryo tabanlı sorulardan oluşan ortak köklü maddelerin kullanılamaması gibi sorunlarını azaltabileceği düşüncesi ile BBT yerine BÇAT yaklaşımı tercih edilmeye başlandı. BÇAT, uygulama öncesinde birleştirilen panel ve modüller ile BBT'nin sorunlarını azaltsa da; daha az uyarlama noktası olması, tamamen bireyselleştirilmiş olmaması, aynı panel ve modül yapılarının sürekli uygulanması ile oluşan test ve madde güvenliği tehditi dezavantajları bulunmaktadır. Maddelerin katılımcının yetenek düzeyinde biraraya getirilerek modüllerin oluşturulduğu yeni bir yaklaşım olan A-BÇAT ise BBT ve S-BÇAT'ın güçlü yönlerinden faydalandığı ve zayıf yönlerini azalttığı için potansiyel taşımaktadır. Bu çalışmanın amacı farklı benzetim koşulları altında S-BÇAT ve A-BÇAT yaklaşımlarını ölçme kesinliği ve madde güvenliği açılarından karşılaştırmaktır. Araştırmada simülasyon çalışması yürütülmüştür. TIMSS uygulamasında kullanılan 3PL modele dayalı maddelerin parametre dağılımları ile 400 maddenin parametreleri üretilerek madde havuzu oluşturulmuştur. 72 simülasyon koşulu altında 100 replikasyon ile karşılaştırmalar yapılmıştır. Sonuçlar, A-BÇAT'ın hem ölçme kesinliği hem de madde güvenliği açılarından S-BÇAT'tan daha iyi sonuçlar ürettiğini göstermektedir. Kısa test uzunluklarında A-BÇAT'ın S-BÇAT'tan oldukça iyi ölçme kesinliğine sahip olduğu görülmektedir. Yetenek dağılımlarına göre, A-BÇAT'ın özellikle normal olmayan dağılımlarda S-BÇAT'tan daha iyi sonuçlar ürettiği sonucuna ulaşılmıştır. Modül/test uzunluğu oranına göre son modül uzunluğunun arttığı K-K-U oranında A-BÇAT'ın ölçme kesinliği farkının arttığı gözlemlenmiştir. A-BÇAT'ın sunduğu başarılı ölçme kesinliği ve madde güvenliği bulguları hem geniş ölçekli değerlendirmeler hem de literatür doğrultusunda tartışılmaktadır.

Anahtar sözcükler: bireyselleştirilmiş bilgisayarlı test, bireyselleştirilmiş çok aşamalı test, anında bireyselleştirilmiş çok aşamalı test, uyarlanabilir test, ölçme kesinliği, madde güvenliği.

Abstract

Many large-scale assessments have used adaptive testing approaches such as Computerized Adaptive Testing (CAT) and Computerized Multistage Testing (MST). The MST approach has been preferred instead of the CAT with the idea that it can reduce the problems of the CAT such as under or over estimating ability, not allowing movement between items, requiring an answer to skip an item, the impossibility of using common-root items consisting of scenario-based questions, and the perception of inequality due to participants taking different tests. Although the MST reduces the problems of the CAT with panels and modules that are combined before implementation, it has the disadvantages of having fewer adaptation points, and the test and item security threats caused by the continuous application of the same panel and module structures. The On-the-fly MST, a new approach in which items are combined at the participant's ability level to form modules, has potential as it utilizes the advantages of the CAT and MST and reduces their weaknesses. The aim of this study is to compare F-MST and O-MST approaches in terms of measurement accuracy and item security under different simulation conditions. A simulation study was conducted in the research. An item pool was created by generating the parameters of 400 items with the parameter distributions of the items based on the 3PL model used in the TIMSS application. In the study, comparisons were made with 100 replications under 72 simulation conditions. The results show that the O-MST produces better results than the F-MST in terms of both measurement precision and item security. For short test lengths, the A-BCAT has considerably better measurement accuracy than the F-MST. According to ability distributions, it was concluded that the O-MST produced better results than the F-MST, especially in non-normal distributions. According to the module/test length ratio, it was observed that the difference in the measurement accuracy of O-MST increased in the K-K-U ratio where the last module length increased. The successful measurement accuracy and item security findings of the O-MST are discussed in the light of both large-scale evaluations and the literature.

Keywords: computerized adaptive testing, computerized multistage testing, on-the-fly computerized multistage testing, adaptive testing, measurement precision, item security.

Teşekkür

Doktora eğitim sürecimin başlangıcından sonuna kadar yetişmemiz için büyük gayret gösteren hem akademik mentörlüğü hem bilgisi hem de davranışları ile bizlere rehberlik eden kıymetli danışmanım Prof. Dr. Nuri DOĞAN'a,

Doktora eğitimimde gelişmem ve ilerlememde büyük katkılar sunan değerli hocalarım Prof. Dr. Selahattin GELBAL, Prof. Dr. Hülya KELECİOĞLU ve Prof. Dr. Burcu ATAR'a,

Tez izleme komitemde yer alarak tezimin tüm aşamalarını sabır ve titizlikle inceleyen ve kıymetli öneriler sunan Doç. Dr. Celal Deha DOĞAN'a,

Tez savunma sınavımda yer alarak tezimi inceleyen ve kıymetli zamanını benim için ayıran Doç. Dr. Seher YALÇIN ve Doç. Dr. Sevda ÇETİN'e,

Eğitimde Ölçme ve Değerlendirme alanına yönelmem için beni destekleyen değerli hocam Prof. Dr. Bayram ÇETİN'e,

Tezimin gelişim aşamasında sunduğu katkılardan dolayı Assoc. Prof. Dr. April ZENISKY'e,

Doktora tez sürecimde yanıt aradığım sorularımı içtenlikle cevaplayan Doç. Dr. Tuğba KARADAVUT'a,

Hem dostlukları hem de eşsiz fikirleriyle doktora sürecimizin her aşamasında yanımda olan değerli arkadaşlarım Erdem BODUROĞLU ve Hüseyin YILDIZ'a,

Doktora sürecimizin her aşamasında birbirimizi desteklediğimiz ASBUZEM'deki değerli iş arkadaşlarıma,

Eğitim hayatımda bugünlere gelmeme katkı sağlayan ve adını sayamadığım tüm hocalarıma,

Doktora sürecimin her aşamasını anlaşıyla karşılayan ve hep kolaylık sağlayan sevgili eşime, canım kızıma, anneme, babama ve kardeşlerime teşekkür ederim.

İçindekiler

Tablolar Dizini

Şekiller Dizini

Simgeler ve Kısaltmalar Dizini

A-BÇAT: Anında Bireyselleştirilmiş Çok Aşamalı Test

DT: Doğrusal Test

BÇAT: Bireyselleştirilmiş Çok Aşamalı Test

BBT: Bireyselleştirilmiş Bilgisayarlı Test

S-BÇAT: Sabit Bireyselleştirilmiş Çok Aşamalı Test

OTB: Otomatik Test Birleştirme

PISA: Sosyoekonomik Düzey

PIAAC: Sosyoekonomik Düzey

LGS: Liselere Geçiş Sistemi

RMSE: Root Mean Square Deviation

MTK: Madde Tepki Kuramı

MIP: Mixed Integer Programming

TIF: Test Information Function

KPSS: Kamu Personeli Seçme Sınavı

ALES: Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı

CPA: Certified Public Accountant

GRE: Graduate Record Examinations

MEB: Milli Eğitim Bakanlığı

ÖSYM: Öğrenci Seçme ve Yerleştirme Merkezi

YDS: Ölçme, Seçme ve Yerleştirme Merkezi

Bölüm 1

Giriş

Bu bölümde sırasıyla problem durumu, araştırmanın amacı ve önemi, araştırma problemi, alt problemler, sayıltılar, sınırlılıklar ve tanımlar sunulmuştur.

Problem Durumu

Doğrusal (Lineer) Testler (DT), yüzyıllardır eğitim alanında yapılan ölçmelerde sınava girenlerin bilgi, beceri ve yeteneklerini ölçmenin en popüler yolu olmuştur. Son yarım yüzyılda bilgisayar donanımı ve yazılımındaki gelişmelerle birlikte, Bireyselleştirilmiş Bilgisayarlı Testler (BBT) önemli ölçüde gelişim göstermiş ve popülerlik kazanmıştır. Özellikle etkili ve verimli yetenek kestirimleri sunması, test uzunluğunu kısaltması nedeniyle BBT, Dünya'da pek çok ulusal ve uluslararası uygulamada benimsenmiştir ve kullanılmaktadır (Khorramdel, Pokropek, Joo, Kirsch & Halderman, 2020; Kirsch & Lennon, 2017). BÇAT ise daha çok yakın zamanlarda sunduğu yenilikler ile dikkat çekmektedir.

Doğrusal testlerde, tüm test katılımcıları her maddeyi alır. Zor maddeler, düşük puanlı bireylerin arasındaki farkları ayırt etmek için çok fazla bilgi vermezler. Benzer şekilde kolay maddeler ise yüksek puanlı bireylerin arasındaki farkları ayırt etmek için çok fazla bilgi vermezler. Dolayısıyla kolay ve zor maddeler, teste girenlerin puan ölçeğinin sırasıyla üst ve alt uçlarındaki yeteneklerinin kestirilmesinde çok az bilgi sağlar. Yani doğrusal testlerden elde edilen puanların etkili kestirimlerini elde etmek için çok sayıda maddeye ihtiyaç vardır (Huang, Lin & Chen, 2009). BBT'de ise maddelerin özellikleri algoritmalar tarafından analiz edilerek bireye özgü testler elde edilmektedir (Raborn & Sari, 2021). BBT, bilgisayar tabanlı bir testtir, sabit veya değişken uzunluğa sahip olabilir. Test yönetim algoritması, katılımcıya art arda maddeler sunmakta ve test ilerledikçe katılımcının yetenek düzeyinin kestirilmesi için maddelerin zorluklarını ayarlamaktadır. BBT ile bir katılımcıya madde uygulanır ve madde yanıtlandıktan sonra bireyin geçici yeteneği kestirilir. Test algoritması bireyin geçici yeteneğini madde havuzundan bir sonraki uygulanacak maddenin seçiminde kullanır.

Sonlandırma kuralına ulaşana kadar tüm maddeler için bu adım tekrarlanır (Wainer, 1990; Hendrickson, 2007). Doğrusal testlerde ölçüm kesinliği teste giren bireylerin yetenek düzeyine göre değişebilir. Bir doğrusal testte genellikle teste katılan grup içinde ortalama düzeydeki bireylerin yetenekleri daha hassas bir şekilde kestirilebilir, fakat puan ölçeğinin uç noktalarına yakın olanlar için ölçme kesinliği azalmaktadır (Hambleton & Swaminathan 1985). Diğer taraftan, BBT'de ise puan ölçeğinin uçlarında olan bireyler de dâhil olmak üzere tüm sınava girenler için eşit derecede ölçme kesinliği sağlayabilmek için katılımcının yetenek düzeyinde kestirim yapılmasına odaklanılır (Mills, Potenza, Fremer & Ward, 2002). Yapılan araştırmalar, iyi yapılandırılmış bir madde havuzu ile BBT'den doğrusal testlerden çok daha etkili kestirimler gerçekleştirebileceğini belirtmektedir (Choi & van der Linden, 2018; Çoban, 2020; Hendrickson, 2007; Lord, 1980; Wainer, 1990).

BBT, doğrusal testlere göre avantajlı olmasının yanında bir takım potansiyel sorunları da içerir. Geniş ve yeterli madde havuzu gerektirmesi, test oluşturma algoritmalarının karmaşık olması, bilgisayar donanımlarının maliyetli olması, bireylerin farklı testleri almasından dolayı oluşan eşitsizlik algısı, bireyin zorlandığı maddeyi atlama (skipping) imkanının olmaması, ekranda sadece bir sorunun görünmesi, maddeyi atlamak için mutlaka cevap gerektirmesi sorunları bulunmaktadır. Bu sorunların bir kısmının azaltılması açısından BÇAT tercih edilebilmektedir (Zenisky & Hambleton, 2014).

BÇAT, BBT'ye benzer şekilde testin uygulanması sırasında testin güçlüğünün teste giren kişinin yetenek düzeyine göre uyarlandığı bir test tasarımıdır. BÇAT, BBT ve DT'nin bir melezi olarak görülebilir, her iki tasarımın da özelliklerini içermektedir. BBT ve BÇAT, testteki maddelerin teste girenlere uygulanma sırasının teste girenlerin önceki maddelerdeki performansına bağlı olması bakımından benzerdir. Bununla birlikte, bir BÇAT'da algoritma, BBT'de olduğu gibi her maddeden sonra değil, modül adı verilen bir dizi madde uygulandıktan sonra harekete geçer ve güncel yetenek kestirilir. Daha özel olarak söylemek gerekirse, ilk başta bireye yönlendirme modülü adı verilen bir grup madde uygulanır. Ardından her katılımcının yönlendirme modülündeki performansına göre

yeteneği hesaplanır ve bir kriter puan (kesme puanı) ile karşılaştırılır. Daha sonra bireyin yeteneği kriter puandan büyükse bireye zor modül (zor madde seti), düşükse kolay modül (kolay madde seti) uygulanır. BÇAT, esneklik ve karmaşıklık açısından DT ile BBT'nin bir uzlaşısı olarak görülebilir. BÇAT, DT'ye kıyasla yetenek ölçeğinde daha verimli ve hassas ölçmeler sağlar. Dolayısıyla ölçme kesinliği açısından BÇAT'ın DT'den daha etkili olduğu belirtilmektedir (Choi & van der Linden, 2018; Kim & Plake, 1993; Lord, 1971; Patsula & Hambleton 1999). Özellikle BÇAT, DT'ye göre test uzunluğunun azalmasını da sağlar (Drasgow & Luecht, 2006; Mead, 2006; Stark & Chernyshenko, 2006; van der Linden, 2010).

BÇAT'da kullanılan modüller, test uygulamasından önce tasarlanıp birleştirildiğinden ve teste giren kişiye bir birim olarak sunulduğundan, test geliştiricilerinin kapsam dengelemesi, test yapısının kalitesi ve testin kendi içinde yönetimi üzerinde daha fazla kontrol sağlanmasına olanak tanır (van der Linden & Glas, 2010). Ayrıca, BBT'nin aksine BÇAT'ta sınav katılımcılarının her modül içerisinde madde yanıtlarını gözden geçirmeye, maddeyi atlamaya, dönüp tekrar cevap vermeye izin verdiğini belirtmekte fayda vardır (Chang, 2015). Hambleton ve Xing (2006), kaldı-geçti kararı vermek için BÇAT'ı BBT ve doğrusal testler ile karşılaştırmıştır. Çalışma sonuçları, BBT'nin ölçme kesinliği açısından BÇAT'tan sadece biraz daha iyi olduğunu göstermiştir. Araştırmacılar, test geliştiricilerinin daha geniş bir güçlük aralığına sahip modüller ve içeriğe göre daha geniş madde bankaları kullanırlarsa BÇAT'tan daha iyi yararlanabilecekleri sonucuna varmışlardır.

Son yıllarda, bireyselleştirilmiş çok aşamalı testler (BÇAT) sunduğu faydalı yönleri ile pek çok geniş ölçekli sınavda kullanılmaktadır. Yeminli Mali Müşavirler (Certified Public Accountants - CPA) Sınavı, 2004'ten beri BÇAT'ı benimsemiştir (Breithaupt, Mills & Melican, 2006). Lisansüstü Kayıt Sınavları'nın (Graduate Record Examinations - GRE) 2011'de gözden geçirilmiş bir versiyonunu BÇAT ile uygulamıştır. Yine 2011 yılında OECD, Uluslararası Yetişkin Yeterliliklerini Değerlendirme Programında (PIAAC) BÇAT tasarımı altında uyarlanabilir bir uluslararası geniş ölçekli testi uygulamıştır (Kirsch & Lennon, 2017).

PIAAC'ı takiben, 2018 yılında gerçekleştirilen PISA (Programme for International Student Assessment) sınavında ise ele alınan 3 temel alandan sadece okuma alanında bir BÇAT tasarımı kullanılmıştır (Khorramdel, Pokropek, Joo, Kirsch & Halderman, 2020). 2021 yılında ise okuma alanına ek olarak matematik okuryazarlığı alanında da BÇAT tasarımı kullanılacağı belirtilmektedir (NCES, 2019).

BÇAT, sınav esnasında bireyin yetenek düzeyine göre uyarlanmış maddeler sunarak, sınava katılanların katılımını ve motivasyonunu artırmaya katkıda bulunabilme potansiyeline sahiptir. PISA, Bilgisayar Tabanlı Testlerde (CBA), Kağıt Tabanlı Testlere (paper-based assessment (PBA) göre daha düşük yanıtsız bırakma oranları bildirmektedir (OECD, 2017). Dolayısıyla BÇAT'ın değerlendirmelerde yanıtsız bırakma ve rastgele yanıtlama davranışlarını azaltabileceği beklenmektedir (Yamamoto, Shin & Khorramdel, 2018). Uyarlanabilir testlerin motivasyon üzerine etkililiğini ele alan pek çok çalışma bulunmaktadır (Arvey, Strickland, Drauden & Martin, 1990; Bergstrom, Lunz & Gershon, 1992; Ortner, Weisskopf & Koch, 2013; Pine, Church, Gialluca ve Weiss, 1979). Ling, Attali, Finn ve Stone (2017), uyarlanabilir testin, sabit maddeli testlerden daha yüksek katılım ve daha düşük kaygı ile sonuçlandığını bildirmiştir. Martin ve Lazendic (2018), BÇAT'ın bilgisayar tabanlı teste göre daha iyi ölçüm hassasiyeti sunduğunu ve motivasyon, katılım ve test deneyimi açısından olumlu sonuçlar verdiğini bildirmektedir.

Son zamanlarda, gerçekçi koşullar altında veya mümkün olduğunca gerçek dünyaya yakın durumlarda bilişsel becerileri ölçmeyi amaçlayan etkileşimli değerlendirmelere olan ilgi artmıştır (Bulut, 2021). Senaryo tabanlı madde grupları, öğrencilerin madde üzerinde değişiklikler yaparak etkileşime girdikleri zengin ortamlar sağlamaktadır. Dolayısıyla hem gerçek yaşam durumlarının incelenmesi hem de senaryo tabanlı değerlendirmenin ele alınması için bir dizi maddeden oluşan daha bütünleştirici sorulara ihtiyaç vardır. Örneğin bir metne dayalı olarak öneriler yazmak veya bir gerçek yaşam durumunda yaşanan bir dizi sorunla baş etmek gibi üst düzey görevleri tamamlamak amacıyla birden fazla maddenin sunulması muhtemeldir. Bu durumda sınav katılımcısının

performansını değerlendirmek için tek bir madde yerine birbiriyle ilişkili madde grubunun yer alması gerekir. BBT uygulamalarında madde bazında yetenek kestirimi yapılmaktadır. Bahsedildiği gibi ortak bir köke ya da senaryoya dayalı madde grubunu ayrı ayrı maddelere bölerek sınav katılımcılarına sunmak bu tür yapıların ölçülmesini geçersiz kılar. Bu nedenle BBT bu senaryoyu desteklemez. Diğer taraftan, örneğin PISA'da maddelerin yaklaşık %30"u insan kodlayıcılar tarafından puanlanan yapılandırılmış yanıtlı (constructed response) maddelerdir. Bu maddelerin puanlaması o esnada hazır olmadığı için BBT ile bu maddeleri kullanmak yine uygun olmayabilir. Madde düzeyinde uyarlanabilir test olan BÇAT tasarımı bu bağlamda daha uygun görünmektedir. Sağlam biçimde oluşturulmuş ve ölçülen yapının çerçevesini yansıtan madde grupları (modüller), birbirinden bağımsız ve olgusal bilgi içeren çok sayıda maddeden daha iyi ölçümler gerçekleştirebilir (Yamamoto ve diğerleri, 2018).

BBT'nin en önemli sınırlılıklarından birisi de sınav katılımcılarının yetenek düzeyini gerçek yetenek düzeyinden az ya da fazla kestirebilmesidir (Chang & Ying, 2008). Bunun nedeni, yaygın olarak kullanılan Fisher Bilgisini en üst düzeye çıkaran MFI yöntemi ile yanıtlanan her maddeden sonra yetenek düzeyinin güncellenmesidir. Örneğin yüksek yetenek düzeyine sahip bir katılımcı ilk birkaç maddeyi sınav heyecanıyla, motivasyonsuzlukla ya da yanlışlıkla ilk birkaç maddeyi yanlış yanıtlarsa ardından gelecek maddelere vereceği yanıtlar ile gerçek yetenek düzeyine ulaşması zor görünmektedir. Benzer şekilde, düşük yetenek düzeyine sahip bir katılımcı şansla ya da geçmiş yaşantıları ile ilk birkaç maddeyi doğru yanıtlarsa yetenek kestirilen yetenek düzeyi gerçek yetenek düzeyinden daha yüksek olacaktır. BÇAT ise yetenek düzeyini her aşamanın sonunda kestirir ve BBT'nin bu dezavantajlı yönünü azaltabilir. Bazı test kuruluşları BBT'nin geniş ölçekli değerlendirmelerde uygulanması ile BBT'nin bu dezavantajlı yönünü ortaya çıkarmışlardır. 2000 yılında Eğitimsel Test Servisi (ETS), bilgisayarlı Lisansüstü Kayıt Sınavları Genel Testi (GRE-CAT) sisteminin sınava giren birkaç bin kişi için yanlış puanlar kestirdiğini fark etmiş ve bu durumdan etkilenen katılımcılara ücretsiz olarak sınava yeniden

girmelerini teklif etmiştir (Carlson, 2000). Bir benzeri durum ise 2002 yılında Lisansüstü Yönetim Kabul Testi'nde (GMAT) yaşanmıştır. Bu sınavda yaklaşık bin adaya yanlış puanlar verildiği bildirilmektedir (Chang, 2004). Yaşanan bu durumlardan sonra BÇAT, BBT'nin bu sorunlarını hafifleteceği düşüncesi ile hızla gelişmeye ve yaygınlaşmaya başlamıştır. Dolayısıyla pek çok test kuruşulu BBT'den vazgeçmiş ve BÇAT'ı benimsemeye başlamıştır (Hendrickson, 2007).

Sabit (Fixed) Bireyselleştirilmiş Çok Aşamalı Testler (S-BÇAT), test tasarımında yer alan modül ve aşamaların testten önce birleştirilerek oluşturulduğu ve maddelerin test yönetiminde konumlarının değişmediği uyarlanabilir bir test yaklaşımıdır. Anında Bireyselleştirilmiş Çok Aşamalı Testler (A-BÇAT) ise modüllerin anında birleştirilerek oluşturulduğu bir BÇAT tasarımıdır. S-BÇAT'ta tüm modüller sınav öncesinde birleştirilerek hazır hale getirilirken, A-BÇAT'ta sınav anında her sınav katılımcısına özel olarak bireyin yetenek düzeyinde bir modül birleştirilir. Sınav katılımcısı, bu modülü yanıtladıktan sonra yetenek düzeyi kestirilir ve kestirilen geçici yetenek düzeyine göre ikinci modül anında birleştirilir. Bu işlem son modül uygulanana kadar devam eder ve nihai yetenek kestirimi elde edilir ve test sonlandırılır. Zheng ve Chang (2015) tarafından önerilen ve gölge test (shadow-test) yaklaşımı altında birleştirilen modüller ile yürütülen A-BÇAT tasarımı yeni bir yöntemdir. A-BÇAT, her modüldeki madde sayısı ve aşama sayısı da dahil olmak üzere pek çok açıdan esnektir. Test geliştiricisi dilerse aşama sayısını arttırabilir. Ayrıca A-BÇAT'ta aşamalardaki maddelerin seçimi, maksimum test bilgisi, madde kullanım sıklığı kontrolü ve kapsam dengeleme gibi test özellikleri kontrol edilebilmekte ve yönetilebilmektedir. Başta uluslararası geniş ölçekli değerlendirmeler olmak üzere pek çok ulusal ve uluslararası değerlendirme programlarının önceki paragraflarda değinilen avantajlı yönlerinden dolayı BÇAT benzeri yaklaşımları benimsediği görülmektedir. Bu çalışmada A-BÇAT ve S-BÇAT yöntemleri farklı koşullar altında karşılaştırılarak incelenmiştir.

Araştırmanın Amacı ve Önemi

Bu araştırmanın amacı, S-BÇAT ve A-BÇAT tasarımlarının farklı benzetim koşulları altında ölçme kesinliği ve madde güvenliği açılarından karşılaştırılmasıdır.

Ülkemizde her yıl onlarca merkezi sınav gerçekleştirilmektedir. Yapılan sınavlara bakıldığında bu sınavların neredeyse tamamının Klasik Test Kuramı'na dayalı olarak kağıt-kalem sınavı şeklinde uygulandığı görülmektedir. Dünya geneline bakıldığında, CPA, GRE, PIAAC, PISA gibi uluslararası geniş ölçekli sınavlarda Madde Tepki Kuramı'na dayalı bireyselleştirilmiş test yaklaşımlarından faydalanıldığı görülmektedir. Dijital çağın getirdiği bilgisayar teknolojilerinden ve Madde Tepki Kuramı'nın avantajlarından faydalanmak amacıyla ülkemizde de bireyselleştirilmiş test yaklaşımlarının kullanılmasının faydalı olacağı düşünülmektedir.

Ülkemizde Millî Eğitim Bakanlığı, ehliyet sınavlarında bilgisayar tabanlı test uygulamalarını 2017 yılında başlatmıştır. Benzer şekilde ÖSYM, Elektronik Yabancı Dil Sınavı'nı (e-YDS) 2017 yılından bu yana bilgisayarlı doğrusal test uygulaması ile sınav merkezlerinde gerçekleştirmektedir. Bu gelişmelerin devamında teknolojideki ve ölçme alanındaki gelişmelerle birlikte yakın gelecekte uyarlanabilir test uygulamalarına geçiş yapılacağı düşünülmektedir. Dolayısıyla ülkemizde yapılan KPSS, ALES, YDS, TUS vb. gibi ülke çapında yapılan sınavlarda uyarlanabilir test uygulamalarına geçiş süreci için araştırmaların yapılması önemlidir.

Sınavlarda yer alan maddeler sınav sonrasında adaylarla paylaşıldığında her yeni uygulamada yeni maddelerin yazılması gerekmektedir. Her sınavda yeniden madde yazılması bir maddi yük getirmektedir. Oysa ki, bireyselleştirilmiş testlerde madde havuzunda yer alan maddeler tekrar tekrar kullanılabilmekte, yeni maddeler havuza eklenebilmekte ve açığa çıkma ihtimali yüksek olan maddeler havuzdan çıkarılabilmektedir. Dolayısıyla yazılan maddelerin daha verimli kullanıldığı düşünülmektedir.

2021 LGS raporu incelendiğinde, Türkçe, Matematik ve Fen Bilimleri alt testlerinin sağa çarpık, TC İnkılap Tarihi ve Atatürkçülük alt testinin uniform, Din Kültürü ve Ahlak Bilgisi alt testinin sola çarpık dağılıma benzer dağılımlar gösterdiği görülmektedir (MEB, 2021). Dolayısıyla gelecekte yeni test yaklaşımlarının Türkiye'de uygulanabilmesi için farklı yetenek dağılımlarından elde edilecek sonuçların önemli olduğu düşünülmektedir.

Türkiye'de bireyselleştirilmiş testler üzerine yapılmış çalışmalar genellikle BBT üzerine yoğunlaşmıştır (Aybek & Çıkrıkçı, 2018; Boztunç-Öztürk & Doğan, 2015; Bulut & Kan, 2012; Eroğlu & Kelecioğlu, 2015; Kalender, 2012; Kezer & Koç, 2014; Şenel, 2017; Sulak & Kelecioğlu, 2019; Şahin Kürşad, 2020; Şahin & Gelbal, 2020; Alkan, 2021; Arzu & Doğan, 2021). BÇAT üzerine yapılmış çalışmalar ise daha sınırlıdır (Doğruöz, 2018; Çoban, 2020; Erdem Kara & Doğan, 2022). A-BÇAT üzerine ise ülkemiz literatüründe yapılmış bir çalışmaya rastlanmamıştır. Bu açıdan A-BÇAT'ın ülkemiz literatürüne kazandırılması açısından da yapılacak araştırma önemli görülmektedir.

A-BÇAT yaklaşımı yeni bir yaklaşım olup farklı simülasyon koşullarında bu yaklaşımların etkililiğinin araştırıldığı çalışmaların uluslararası literatürde sınırlı olduğu görülmektedir (Zheng & Chang, 2015; Han & Guo, 2016; Choi & van der Linden, 2018). Bu çalışmada ele alınan olan farklı test uzunluğu ve yetenek dağılımının etkilerine daha önce herhangi bir çalışmada yer verilmediğinden; bu çalışmanın yeni yaklaşımların gelişimi açısından katkı sunacağı düşünülmektedir.

Araştırma Problemi

Belirlenen farklı benzetim koşulları altında S-BÇAT ve A-BÇAT yaklaşımlarının ölçme kesinliği ve madde güvenliği ne düzeyde değişim göstermektedir?

Alt Problemler. Problem cümlesi doğrultusunda aşağıda yer alan 6 alt problem şu şekilde oluşturulmuştur:

1) Test uzunluğuna (20-30-40) göre S-BÇAT ve A-BÇAT yaklaşımlarının RMSE, MAB ve BIAS değerleri nasıl değişim göstermektedir?

2) Yetenek dağılımına (normal dağılım, sağa çarpık, sola çarpık, uniform) göre S-BÇAT ve A-BÇAT yaklaşımlarının RMSE, MAB ve BIAS değerleri nasıl değişim göstermektedir?

3) Modül/test uzunluğu oranına (U-K-K, O-O-O, K-K-U) göre S-BÇAT ve A-BÇAT yaklaşımlarının RMSE, MAB ve BIAS değerleri nasıl değişim göstermektedir?

4) Farklı test uzunluklarında (20-30-40) S-BÇAT ve A-BÇAT yaklaşımlarının madde kullanım sıklığına ve kullanılan madde sayılarına göre madde güvenliği nasıl değişim göstermektedir?

5) Farklı yetenek dağılımlarında (normal dağılım, sağa çarpık, sola çarpık, uniform) S-BÇAT ve A-BÇAT yaklaşımlarının madde kullanım sıklığına ve kullanılan madde sayılarına göre madde güvenliği nasıl değişim göstermektedir?

6) Farklı modül/test uzunluğu oranlarında (U-K-K, O-O-O, K-K-U) S-BÇAT ve A-BÇAT yaklaşımlarının madde kullanım sıklığına ve kullanılan madde sayılarına göre madde güvenliği nasıl değişim göstermektedir?

Sayıltılar

Veri üretiminde kullanılan madde parametreleri, birey yetenek parametreleri ve yanıt örüntülerinin gerçek durumu yansıttığı varsayılmıştır.

Sınırlılıklar

1. Araştırma, benzetim koşulları altında üretilen veri seti ile sınırlıdır.

2. Araştırma 1-0 verisi altında 3 parametreli lojistik model ile sınırlıdır.

Tanımlar

Anında Bireyselleştirilmiş Çok Aşamalı Test (A-BÇAT): Modüllerin sınav anında katılımcının geçici yetenek düzeyine göre birleştirilerek katılımcıya sunulduğu bireyselleştirilmiş çok aşamalı test türü.

Aşama: İçerisinde en az bir modül bulunan S-BÇAT tasarımının basamağı.

Bireyselleştirilmiş Bilgisayarlı Test (BBT): Bireyin her bir maddeye verdiği cevap sonrası yeteneğinin kestirildiği ve kestirilen yetenek düzeyine uygun olarak yeni maddenin yönlendirildiği uyarlanabilir test modeli.

Bireyselleştirilmiş Çok Aşamalı Test (BÇAT): Sınav öncesi oluşturulan panel üzerinde sınav katılımcısının modüllere verdiği yanıtlara göre aşamalar arasında yönlendirildiği uyarlanabilir test yaklaşımı.

Doğrusal Test: Tüm sınav katılımcılarına sabit uzunlukta ve aynı (veya eşdeğer) formun uygulandığı test yaklaşımı.

Modül: Birden fazla maddenin istenen test özelliklerine göre bir araya getirildiği madde grubu.

Panel: Birkaç aşamanın bir araya gelerek oluşturduğu desendir.

Rota: Katılımcının panel içerisinde aşama ve modüller üzerinde izlediği yol.

Test Birleştirme: Belirlenen test özellikleri ve kısıtlamaları karşılayacak şekilde bir madde havuzundan maddelerin seçilerek bir test formunun oluşturulması.

Otomatik Test Birleştirme: Test özellikleri ve kısıtlamalarına göre madde havuzundan madde seçiminin bilgisayar algoritmaları tarafından yapılarak test birleştirme işleminin gerçekleştirilmesi.

Bölüm 2

Araştırmanın Kuramsal Temeli ve İlgili Araştırmalar

Araştırmanın bu bölümünde öncelikle BÇAT uygulamalarının temelini oluşturan Madde Tepki Kuramı (MTK) hakkında bilgiler verilmiştir. Ardından BÇAT ve Test Birleştirme üzerinde durulmuştur.

Madde Tepki Kuramı

Eğitimde gerçekleştirilen ölçmelerin hepsinde ölçme işlemine konu olan bir değişken vardır. Bu değişken genellikle doğrudan gözlenemeyen ve sezgisel olarak anlaşılabilen bir özelliktir. Örneğin bireyin zekâ düzeyinin ölçülmeye çalışıldığı düşünüldüğünde zekâ, tanımlanabilir, bu değişkene ilişkin özellikler sıralanabilir, fakat doğrudan gözlenemez. Dolayısıyla bireyin gözlenemeyen değişkenlerine örtük değişken (latent variable) adı verilmektedir (Baker, 2001). Örtük değişkenler, bireyin boy uzunluğu ya da ağırlığı gibi doğrudan gözlenebilir değişkenler olmayıp tanımlanabilen kuramsal ve kavramsal yapılardır. Eğitimde yapılan ölçme çalışmalarının temel amacı bireylerin sahip oldukları örtük değişkenlerin düzeyinin belirlenmesidir (Baykul, 2015).

Tarihsel gelişim içerisinde örtük değişkenlerin ölçülmesinde pek çok kuram ortaya konulmuştur. Bu kuramlardan Klasik Test Kuramı (KTK) ve Madde Tepki Kuramı (MTK) ölçmede üzerine en çok yoğunlaşılan iki kuramdır. KTK, bireyin maddelere verdiği yanıtlardan bireyin gerçek puanını belirlemeyi hedefler. KTK, bireyin gerçek puanını gözlenen puanına karışan hatalardan arındırılması ile elde edilebileceğini varsaymaktadır (Baykul, 2015). KTK'nın en büyük avantajı, varsayımlarının kolay karşılanması, madde parametrelerinin kolaylıkla hesaplanabilmesidir. Diğer taraftan, KTK bazı sınırlılıklar içermektedir. Lord (1953), KTK'daki gerçek puanın testin güçlüğüne göre değiştiğini belirtmektedir. Örneğin zor bir testte birey düşük gözlenen puan alırken – dolayısıyla düşük gerçek puan – kolay bir testte yüksek gözlenen puan – dolayısıyla yüksek gerçek puan – alacaktır (alıntı Sünbül & Erkuş, 2013). Bireyin yetenek düzeyi değişmemiş iken gerçek

puanının farklı değerler alması KTK'da bireyin gerçek puanının teste veya içerisinde bulunduğu gruba bağımlı olduğunu göstermektedir. KTK'nın sınırlılıklarını aşmak için Madde Tepki Kuramı (MTK) 1930'lu yıllarda geliştirilmeye başlanmıştır. Lord'un 1950 ve 1960'lı yıllardaki "normal ogive model" üzerine çalışmaları, Rasch'ın 1960'larda "Rasch model" adı ile yaptığı çalışmalar ve Birnbaum'un "lojistik model" üzerine yaptığı çalışmalarla MTK gelişim göstermiştir (Baker, 2001; Himelfarb, 2019).

Uyarlanabilir testin başarılı bir şekilde sürdürülebilmesi için maddelerin özelliklerini ve sınav katılımcılarının yeteneklerini yansıtacak parametrelere sahip iyi bir model seçimi önemlidir. Madde Tepki Kuramı (MTK), yaklaşık yüzyıllık tarihsel gelişiminde istatistiksel olarak başarılı bir şekilde işlemektedir. MTK'da örtük değişkenleri tanımlamak için yetenek (ability) terimi ϑ (theta) ile gösterilmektedir. Matematiksel olarak ϑ yetenek düzeyi $-\infty$ ile $+\infty$ arasında değerler alabilir. Bireyin yeteneğinin ölçülmesinde önceden geliştirilmiş ve her biri örtük özelliğin bazı yönlerini ölçmeye yönelik bir dizi maddeden oluşan test/ölçek bireye uygulanır. KTK'daki gibi katılımcının testten aldığı toplam puanın aksine, MTK'da bireyin her bir maddeye verdiği yanıtın doğru olma olasılığı ile ilgilenilir. MTK, testten alınan toplam puana değil, testin her bir maddesine odaklanır ve madde kalitesini ön plana çıkarır. Bireylerin maddeye verdiği yanıtların yetenek ölçeği üzerindeki regresyonu ile Madde Karakteristik Eğrisi (MKE) oluşur (Hambleton ve Swaminathan, 1985). MKE, MTK'nın temel taşı niteliğindedir. MKE bize maddenin kalitesi, hangi güçlük düzeylerinde daha iyi ölçüm yaptığı ve madde parametreleri hakkında bilgiler verir.

Testlerde yer alan maddelerin puanlama şekillerinin birbirine benzer olması beklenir. Testlerde kullanılan maddeler genellikle çoktan seçmeli sorulardan oluştuğundan doğru yanıtlar 1, yanlış yanıtlar 0 olarak kodlanır ve bu şekilde kodlanan puanlara iki kategorili (dichotomous) puanlanan maddeler olarak adlandırılır. Puanlama ikiden fazla kategori içeriyorsa (örneğin yapılandırılmış yanıtlı sorular 0-1-2 veya ölçekler 1-2-3-4-5) çok kategorili (polytomous) maddeler olarak adlandırılır. MTK'nın tarihsel gelişiminde pek çok MTK modeli geliştirilmiştir. Maddelerin özelliklerine göre iki kategorili (0-1, dichotomous) ya

da çok kategorili (polytomous) maddelere yönelik farklı MTK modelleri geliştirilmiştir. Rasch model, İki Parametreli Lojistik Model (2PLM), Üç Parametreli Lojistik Model (3PLM), Dört Parametreli Lojistik Model (4PLM), vd. iki kategorili (0-1, dichotomous) maddeler için geliştirilmiş MTK modelleri olarak sıralanabilir. Kısmi Puan Modeli (KPM), Aşamalı Tepki Modeli (ATM), Dereceli Tepki Modeli (DTM), Genelleştirilmiş Kısmi Puan Modeli (GKPM) çok kategorili (polytomous) maddeler için geliştirilmiş MTK modellerine örnek gösterilebilir. Bu araştırma 3PLM üzerine temellendiğinden araştırmada yalnızca 3PLM kuramsal olarak ele alınmıştır.

Üç Parametreli Lojistik Model (3PLM)

Birnbaum'un (1968) Üç Parametreli Lojistik Modeli (3PLM) yaygın olarak kullanılan bir MTK modelidir. Bireylerin bir maddeyi doğru yanıtlama olasılığını veren modelin genel formülü aşağıda yer alan denklem ile sunulmuştur:

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

θ yetenek düzeyinde bireyin bir maddeyi doğru yanıtlama olasılığı $P(\theta)$ ile gösterilir. Denklemde yer alan a madde ayırtediciliği, b madde güçlüğü, c ise şans parametresidir. Madde ayırtediciliği (a), MKE'nin $\theta = b$ noktasındaki eğimi olarak ifade edilebilir. MKE'nin daha dik olması bu maddenin ayırt ediciliğinin daha yüksek olduğunu gösterir. Madde ayırt ediciliği, $\theta = b$ noktasının üstündeki ve altındaki bireyleri ne düzeyde iyi ayırt edebildiğini gösterir. Madde ayırt ediciliğinin düşük olması, maddenin ilgili örtük özellik ile ilgili olarak bilen ve bilmeyenleri iyi ayırt edemediğini, dolayısıyla maddenin iyi çalışmadığının göstergesi olarak da kabul edilebilir (Reckase, 2009). Madde güçlüğü (b), maddenin kolaylığının veya zorluğunun belirlenmesini sağlar. Teorik olarak madde güçlüğü (b) parametreleri -3 < b < +3 aralığında değerler alır. Madde güçlüğü parametresinin azalması maddenin kolaylaştığı, artması ise maddenin zorlaştığı anlamını taşır. Aynı zamanda madde güçlüğü parametresi, maddenin ϑ yetenek ölçeğinin hangi noktasında verimli ölçüm yapabileceğini gösterir. Madde güçlüğünün bu özelliği, uyarlanabilir test yaklaşımlarında

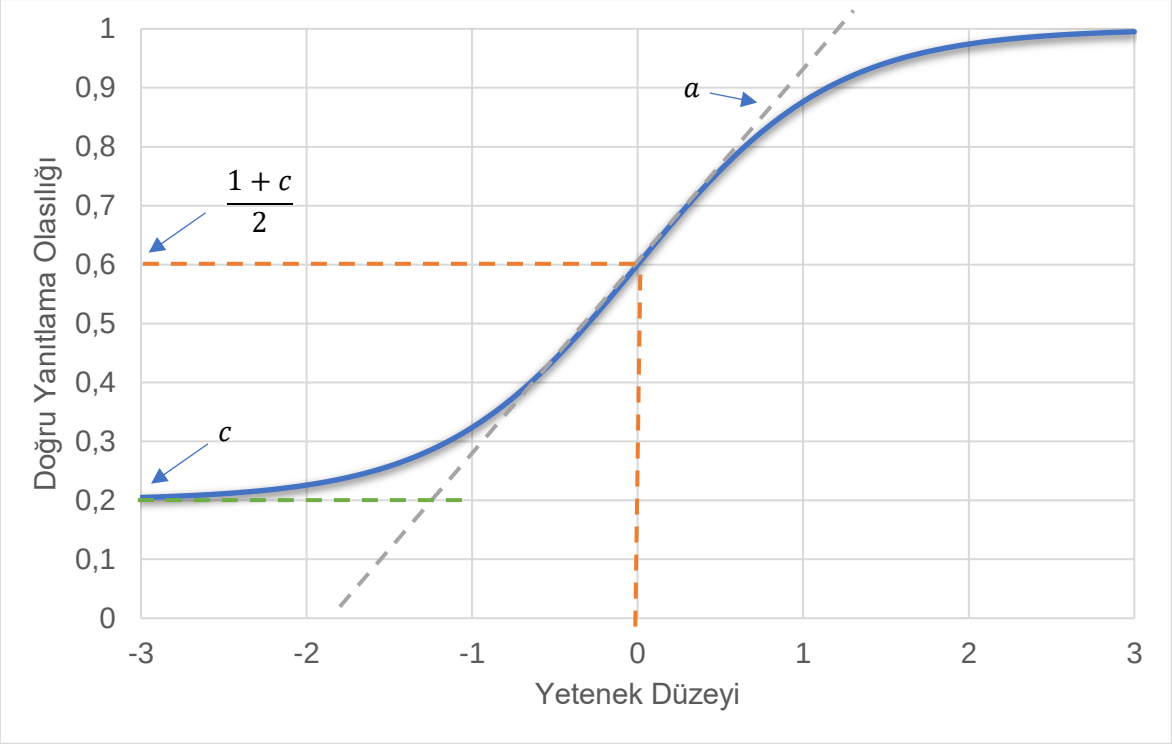
bireye sunulacak madde veya modülün belirlenmesinde önemli rol oynamaktadır. Madde şans parametresi (c) ise alt asimptot olarak da bilinir ve düşük yetenek düzeyindeki bireylerin maddeyi doğru yanıtlama olasılığını gösterir. Diğer bir deyişle, düşük yetenek düzeyindeki bireyler bir maddeyi daha yüksek oranda doğru yanıtlıyorsa maddenin c parametresi artar, aksi durumda azalır. D ise ölçekleme sabitidir ve genellikle 1,000 ya da 1,702 alınmaktadır.

Madde Karakteristik Eğrisi (MKE)

3PL modelin genel formülü ile ifade edilen denklem ile her ϑ yetenek düzeyindeki maddeyi doğru yanıtlama olasılıkları hesaplanarak MKE çizilebilir. Şekil 1'de madde parametreleri a=1, b=0, c=0.2 olacak şekilde bir maddenin MKE'si sunulmuştur. MKE'de işaretlendiği gibi eğrinin dikliği madde ayırt ediciliği ile ilişkilidir. Eğrinin alt asimptotunun y eksenini kestiği nokta şans parametresi olan c = 0.2 değerine karşılık gelir. $\frac{1+c}{2} = 0.6$ doğru yanıtlama olasılığına karşılık gelen yetenek düzeyi ise $\vartheta_b = 0{,}0$ madde güçlük düzeyini gösterir.

Şekil 1

Madde Karakteristik Eğrisi



Tablo 1'de madde parametreleri (a, b, c) verilen üç farklı maddenin Madde Karakteristik Eğrileri Şekil 2'de sunulmuştur.

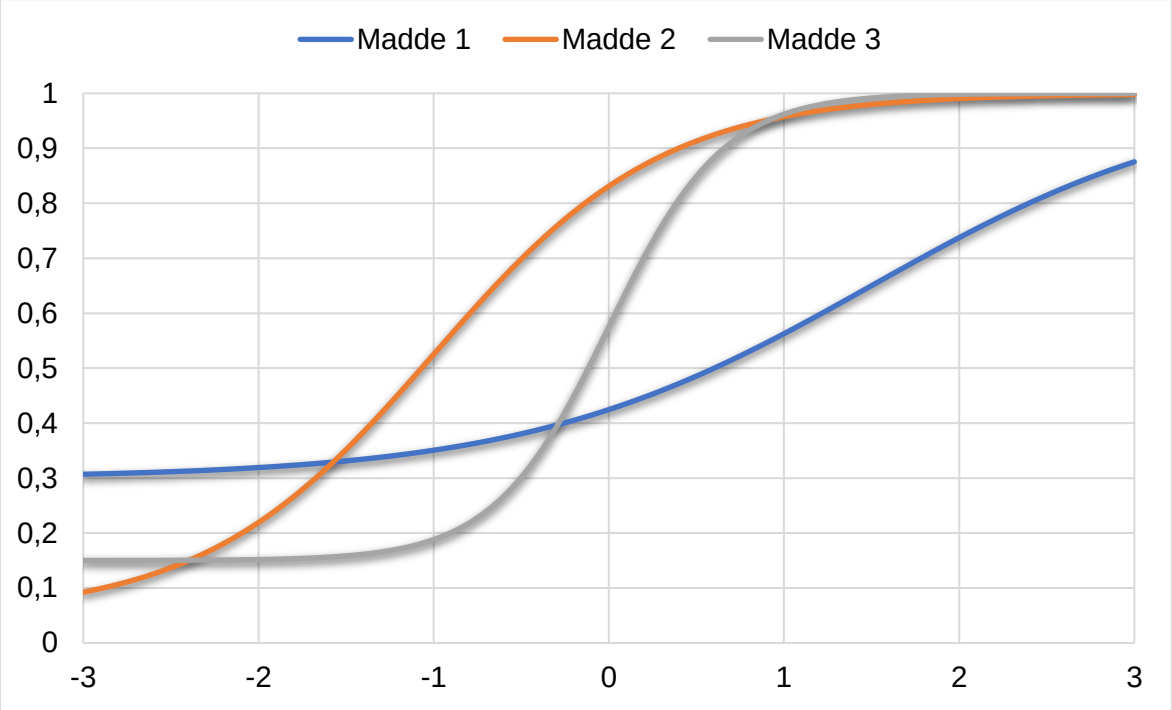
Tablo 1

Madde Parametreleri

Maddeler	Madde Ayırtediciliği (a)	Madde Güçlüğü (b)	Şans Parametresi (c)
Madde 1	0,6	1,5	0,30
Madde 2	0,9	-1	0,05
Madde 3	1,8	0	0,15

Şekil 2

Madde Karakteristik Eğrileri



Madde 1, madde ayırtediciliği düşük bir maddedir ve diğer MKE'lere göre daha eğik görünmektedir. Madde 1'in MKE'sinin diğer maddelere göre daha sağda yer alması zor bir madde olduğunu göstermektedir ve bu durum madde güçlüğü parametreleri incelendiğinde anlaşılabilir. Madde 2'nin MKE'si ise madde ayırt ediciliğinin yüksek olmasından dolayı daha dik bir eğim ile yükselmektedir. Madde 3'ün MKE'si yetenek ölçeğinin en solunda yer almaktadır. Bu durum maddenin diğer maddelere göre kolay olduğunu gösterir. Her üç maddenin MKE'lerinin alt asimptotlarına göre c şans parametreleri yorumlanabilir.

Madde Bilgi Fonksiyonu ve Test Bilgi Fonksiyonu

KTK'da tüm yetenek ölçeği için tek bir güvenirlik katsayısı kestirilmektedir. Halbuki bir maddenin/testin yetenek ölçeğinin her noktasında aynı hassasiyet ile kestirim gerçekleştiremeyeceği açıktır. MTK'da ise yetenek ölçeğinin her noktasında ne düzeyde hassas kestirim yapıldığını sunan madde/test bilgi fonksiyonu kavramı bulunmaktadır. Bilgi fonksiyonu, bir maddenin yetenek ölçeğinin hangi noktalarında etkili ve duyarlı kestirim

yaptığı bilgisini sunar. Örneğin bir maddenin yetenek ölçeğinin bir ϑ noktasında sunduğu bilgi miktarı büyük ise madde o θ düzeyinde etkili kestirimler gerçekleştirir. Bilgi miktarının küçük olması ise yeteneğin duyarlı bir şekilde kestirilemediğini belirtir. Bu durumu bir örnek ile açıklamak gerekirse, düşük düzeyli bir öğrenciye sunulan çok zor bir soru, öğrencinin ölçülen özelliği için bir ayrım yapamayacak olduğundan öğrencinin soruyu yanlış cevaplaması beklenir. Çünkü çok zor soru, düşük yetenek düzeylerinde bir bilgi sunamamakta ve etkili kestirim yapabilmesi için düşük yetenek düzeylerindeki öğrenciler arasındaki farkı ayırt edememektedir. Bir maddenin en duyarlı ve etkili kestirim yapabildiği yetenek düzeyi, madde güçlüğüne (b) karşılık gelir. Maddenin güçlük düzeyinden uzaklaştıkça bilgi miktarı azalmaktadır. Bilgi miktarı uç yetenek düzeylerinde ise sıfıra doğru yaklaşacaktır.

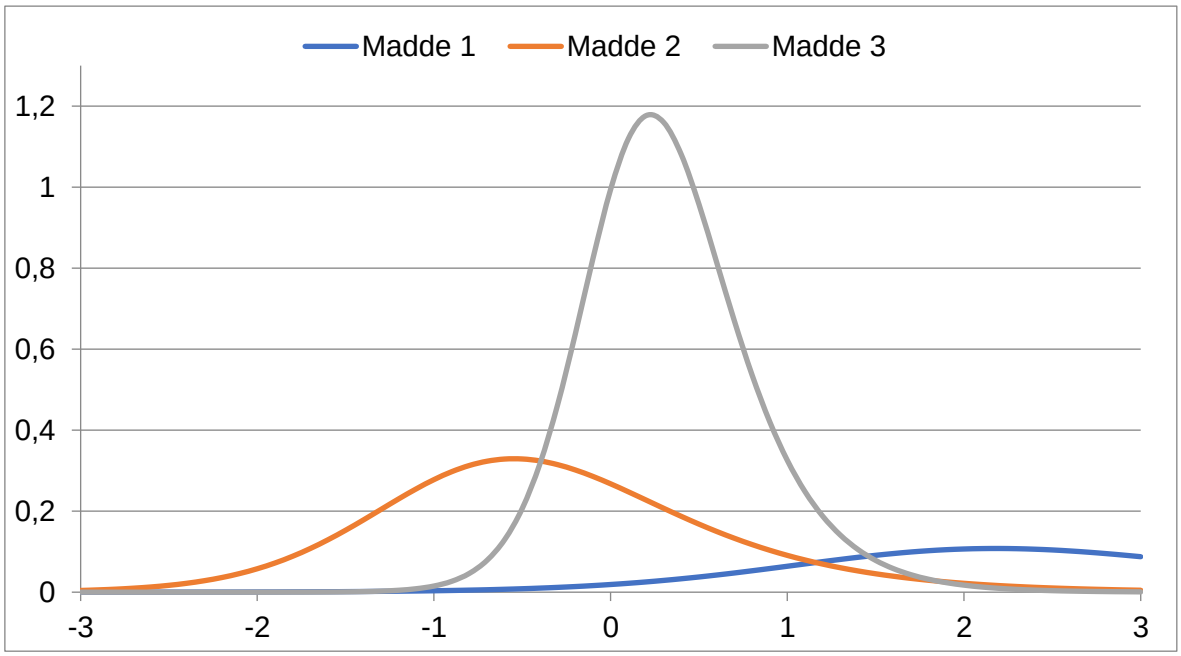
3PLM'nin madde bilgi fonksiyonu formülü aşağıda yer alan denklemde görülmektedir:

$$I(\theta) = a^2 \left[\frac{1 - P(\theta)}{P(\theta)}\right]\left[\frac{P(\theta) - c^2}{1 - c^2}\right]$$

$I(\theta)$, maddenin θ yetenek düzeyinde sunduğu bilgi miktarını belirtmektedir. Tablo 2'de parametreleri verilen üç farklı maddenin Madde Bilgi Fonksiyonları Şekil 3'te sunulmuştur.

Şekil 3

Madde Bilgi Fonksiyonları



Şekil 3'te yer alan Madde Bilgi Fonksiyonları incelendiğinde, her maddenin madde güçlüğü düzeyinde maksimum bilgi düzeyine ulaştığı görülmektedir. Ayrıca maddelerin bilgi miktarının a madde ayırtedicilik parametresi ile orantılı olarak arttığı hem grafikten hem de denklemden anlaşılabilmektedir.

Test Bilgi Fonksiyonu ise bir testin herhangi bir yetenek düzeyinde testin sağladığı bilgi miktarı olarak ifade edilebilir. Test Bilgi Fonksiyonu, herhangi bir yetenek düzeyinde testte yer alan maddelerin bilgi fonksiyonlarının toplamına eşittir. Test Bilgi Fonksiyonunun denklemi şu şekilde ifade edilebilir:

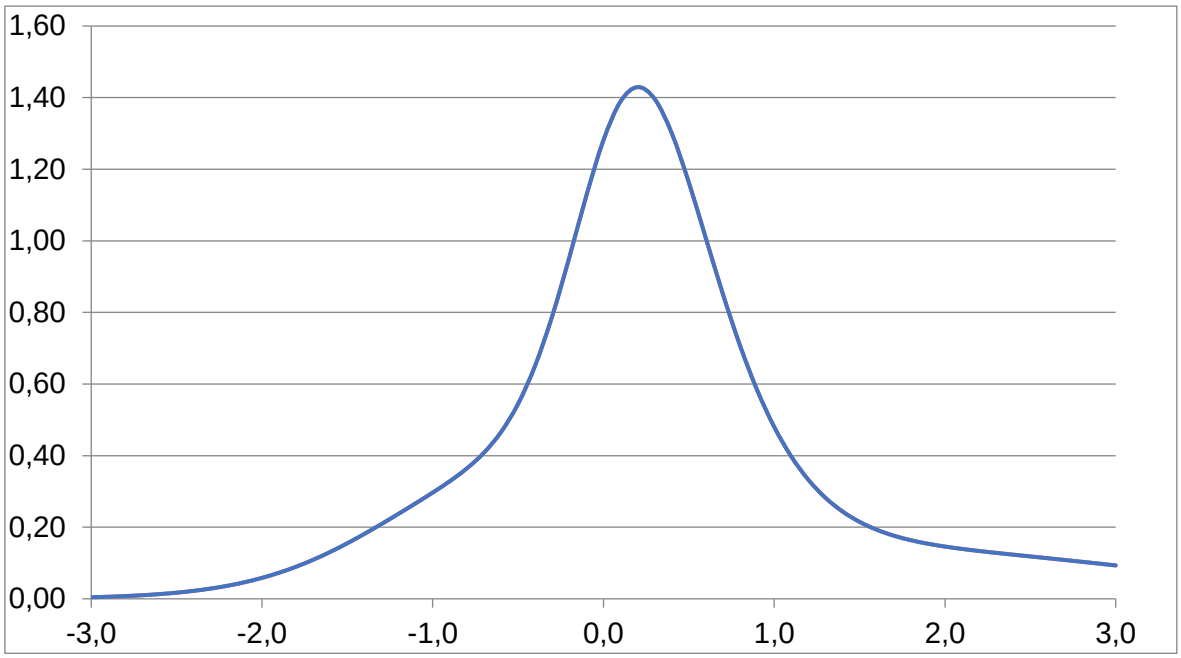
$$I_{Test}(\theta) = \sum_{i=1}^{N} I_i(\theta)$$

Denklemde yer alan i=1,2,3,..,N olmak üzere testte yer alan maddeleri ifade eder. $I_i(\theta)$, i. madde için θ yetenek düzeyindeki bilgi miktarıdır.

Tablo 2'de parametreleri verilen üç maddenin bir test oluşturabileceği varsayılırsa, bu testin Test Bilgi Fonksiyonu Şekil 4'teki gibi oluşacaktır.

Şekil 4

Test Bilgi Fonksiyonu



Madde Tepki Kuramının Varsayımları

MTK'nın sunduğu avantajlardan faydalanmak için modelin varsayımlarının sınanması ve sağlanması gerekir. Hambleton ve Swaminathan (1985), MTK'nın varsayımlarını (a) boyutluluk, (b) yerel bağımsızlık, (c) madde karakteristik eğrisi uyumu, (d) hız testi olmaması olmak üzere dört başlık altında ele almaktadır. MTK varsayımlarının sağlanamaması durumunda psikometrik olarak bazı sorunlar yaşanacaktır. Örneğin tek boyutluluk ihlal edilirse, örtük yetenek uzayının çok boyutlu yapısı ile tek boyutlu MTK modeli birebir eşleme yapmayacaktır. Dolayısıyla tek boyutlu MTK ile elde edilen sonuçların bireyler açısından yanlı olabileceği anlamına gelir (Reckase, 2009).

Sabit Bireyselleştirilmiş Çok Aşamalı Testler (S-BÇAT)

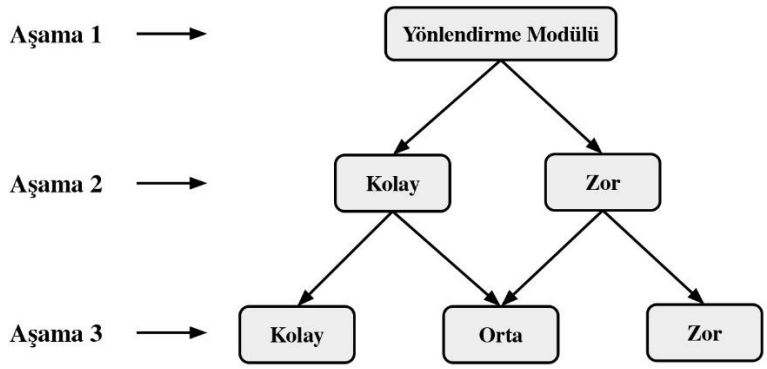
S-BÇAT, önceden birleştirilen madde gruplarının (modüller) algoritma tarafından seçilerek testin aşamalar halinde sınav katılımcılarına yöneltildiği algoritma tabanlı test yaklaşımıdır. S-BÇAT yaklaşımının literatürde farklı şekilde kısaltmalara sahip olduğu görülmektedir. BÇAT'ın önünde yer alan "sabit (fixed)" kelimesi literatüde "standart

(standard)" ve "geleneksel (conventional)" olarak da isimlendirilmektedir. Bu çalışmada "sabit" kelimesi tercih edilmiştir.

S-BÇAT'ta önceden birleştirilen bu madde gruplarının her biri modül olarak adlandırılmaktadır. Modüller, S-BÇAT tasarımının temel birimleridir. İlk aşamada genellikle yönlendirme modülü (routing module) adı verilen ve orta güçlük düzeyinde bir modül uygulanır. Birey yetenek düzeyine göre modüller arasında aşağı doğru yönlendirilerek sınav tamamlanır. Şekil 5'te, birinci aşamada bir modül, ikinci aşamada iki modül ve üçüncü aşamada üç modüle sahip 1-2-3 dizaynında S-BÇAT tasarımının örneği görülmektedir. Bu tasarımda yer alan modüllerin güçlüğü kolay, orta veya zor olarak isimlendirlebilir. S-BÇAT'ın ilk aşamasında tüm test katılımcılarına yönlendirme (routing) modülü adı verilen ve genellikle orta düzey madde güçlüğünde birleştirilen ilk modül uygulanır. Test katılımcılarının yönlendirme modülündeki performansına göre ikinci aşamada kolay ya da zor modüle yönlendirilirler. İkinci aşamadaki test performansına göre üçüncü aşamada yer alan kolay, orta veya zor modülden birine yönlendirilirler. Üçüncü aşamadaki performansı da göz önüne alınarak bireyin nihai yeteneği kestirilir ve test sonlandırılır.

Şekil 5

1-2-3 S-BÇAT Tasarımı



Aşama sayısı ve modül sayısı S-BÇAT'ın tasarımına göre değişebilir. Şekil 5'te yer alan oklar modüller arası yönlendirme kurallarını göstermektedir. Bu tasarımda yönlendirme kuralları sınırlandırılmıştır. Örneğin ikinci aşamada en kolay aşamadan üçüncü aşamadaki

en zor modüle geçilememektedir. Araştırmanın bu bölümünde bir S-BÇAT tasarımında yer alan temel bileşenler başlıklar halinde açıklanmıştır.

Test Tasarımı

S-BÇAT'da test tasarımının gerçekleştirilmesi için testin yapısını oluşturacak bir dizi kararlar almak gerekir. Bu kararlar, test birleştirme için panel, aşama, modül ve kısıtlamalarla ilgili soruları içerirken, yönlendirme kuralları ve yetenek kestirimi de bu kararlara uygun olarak uyarlanmalıdır.

Bir S-BÇAT tasarımında alınacak kararlar DT veya BBT'ye benzer şekilde işler. Test birleştirme ve yetenek kestirimi için alınması gereken kararlara dair sorular şunlardır:

- Test uzunluğu ne kadar olmalıdır?
- Kaç panel oluşturulmalıdır?
- Kaç aşama yeterlidir?
- Her aşamada kaç modül kullanılmalıdır?
- Modüller arasındaki yönlendirme kuralları nelerdir?
- Tüm test için istenen güçlük düzeyi nedir ve bu nasıl sağlanacaktır?
- Test nasıl puanlanmalıdır?

Modül düzeyinde dikkate alınması gereken sorular da vardır:

- Her bir modül uzunluğu ne kadar olmalıdır?
- Her modül için istenen güçlük düzeyi nedir?
- Her modül içindeki maddelerin güçlük dağılımı nasıl olmalıdır?
- Bir modülü oluşturmak için maddeler nasıl ve hangi kapsamdan seçilecek?
- Bir modülün özellikleri nelerdir?

Test tasarımında diğer önemli konular arasında ise kapsam dengelenmesi ve madde kullanım oranının kontrol edilmesi yer alır. Bu bölümde BÇAT tasarımının temel bileşenleri başlıklar halinde ele alınmıştır.

Aşama ve Modül

Modül, bir birim olarak yönetilen ve puanlanan bir madde grubunu ifade eder. Bir modüldeki maddeler genellikle kapsam özellikleri gibi belirli gereksinimlerin yanı sıra belirli güçlük gereksinimine de uygun olacak şekilde bir araya getirilmektedir. Modüle bazı kaynaklarda "test takımı" da denilmektedir, fakat modülde yer alan maddeler, Wainer, Bradlow ve Wang'ın (2007) çalışmasında olduğu gibi, birbirine bağımlı olmak zorunda değildir.

Aşama ise bir ya da daha fazla modülün bir araya gelmesi ile oluşur. Genellikle ilk aşamada tek modül yer alırken, ileriki aşamalarda modül sayısı artar. Birinci aşamadaki modülü alan sınav katılımcısı, yönlendirme kurallarına göre ikinci aşamadaki modüllerden birine yönlendirilir. Süreç, aşamalar arasında bu şekilde devam eder.

Test geliştiricisi, her modül için ortalama madde güçlüğünü ve modülün kapsayacağı güçlük aralığını belirlemelidir. Ayrıca modüller arasında ortalama madde ayrım aralığını da belirlemelidir. Genel olarak, modüllerin geliştirilmesinde madde güçlüğü ve madde ayırt ediciliği için testin amacına ve sınava girenlerin dağılımına göre maddelerin seçilmesi esastır. Kim ve Plake (1993), çok farklı madde güçlük parametrelerine sahip bir modülün özellikle uç yetenek dağılımlarında iyi ölçüm sağlarken, belirli bir düzeyde yoğunlaşan bir modülün yetenek dağılımının orta noktasında daha iyi ölçüm sağladığını belirmektedir.

Bir BÇAT tasarımında aşama ve modüllerin düzenlenmesi, BÇAT'ın geliştirilmesinde kritik derecede önemli bir husus olmaya devam etmektedir. Çoğu BÇAT araştırması iki, üç veya dört aşamalı tasarımlar kullanmıştır. Genel olarak bakıldığında, aşama sayısının artması daha fazla uyarlanmış ve esnek testlerin oluşmasına izin verir. Diğer taraftan, teste daha fazla aşama eklemek, nihai test formunun ölçüm hassasiyetine

fazla bir katkıda bulunmadan test düzeneğinin karmaşıklığını da arttırır (Luecht & Nungester, 1998; Patsula & Hambleton, 1999).

İki aşamalı BÇAT tasarımı basittir, fakat yalnızca bir yönlendirme işlemi olması nedeniyle daha yüksek bir yönlendirme hatası olasılığını içermesi dezavantajı olarak görülmektedir. Bu yönlendirme hatası olasılığı, özellikle yönlendirme için kesme puanına yakın olan sınav katılımcıları için yüksektir. Bu nedenle iki aşamadan daha fazla aşamaya sahip olmak bu tür hataların telafi edilmesini sağlayabilir (Zenisky, Hambleton & Luecth, 2009). Daha fazla aşama, daha fazla esnekliğe ve hassas ölçüme izin vermektedir.

Aşama sayısıyla ilgili bir diğer husus, sınava girenler için test uygulaması sırasında iyileşme fikridir. İki aşamalı bir BÇAT, nesnel anlamda oldukça bilgilendirici olabilir. Ancak sadece iki aşamanın operasyonel kullanımına ilişkin bir sorun, herhangi bir nedenle aşamalar arasındaki yönlendirme uygun değilse, sınava girenlerin yetenek kestiriminin ne düzeyde başarılı olacağına dair bir endişe bulunmaktadır (Wang, Fluegge & Luecht, 2012). Bu nedenle test geliştiricileri, testler için minimum üç aşamalı bir BÇAT uygulamayı tercih etme eğiliminde olmuştur. Aynı zamanda aşamalara eklenen bir ya da birkaç modül, sınava girenler için psikolojik bir adalet güvencesi işlevi de görmektedir.

Daha fazla aşama ve modül, testin sınava girenin yetenek düzeyine uyum sağlaması için daha fazla yönlendirme noktası sağlar. Bir BÇAT, sınava giren kişinin performansına uyum sağlamak için birçok aşama ve birçok farklı modül ile oluşturulabilir, bu da aslında bir BÇAT'ın uygun madde düzeyindeki bir BBT'nin uyarlanabilirliğine yaklaşmasına veya hatta bir alt uygun BBT aracılığıyla elde edilen ölçüm kalitesini aşmasını sağlar, ancak böyle bir seçim, ölçüm kesinliğini önemli ölçüde iyileştirmeden bir değerlendirmenin karmaşıklığını artırabilir (Jodoin, Zenisky & Hambleton, 2006). Dolayısıyla aşama ve modül sayısının artması, etkili kestirim sunabilir, fakat tasarımın karmaşıklığını arttırır ve daha fazla kaliteli madde ihtiyacı doğurur. Dolayısıyla bir BÇAT tasarımında aşama ve modül sayısının arttırılmasında avantaj ve dezavantajlar dikkate alınarak en uygun noktanın tespit edilmesi gerekir.

Literatürde yer alan pek çok araştırma ve uygulama, ilk aşamada bir modül kullanmaktadır. Modül sayısı ileriki aşamalarda arttırılmaktadır. Aşama sayılarında olduğu gibi daha fazla çeşitli güçlüklerde modüller eklemek daha fazla esnekliğe ve hassas kestirime imkân vermektedir. Patsula ve Hambleton (1999), ikinci ve üçüncü aşamada modül sayısını 5'e çıkarmanın yetenek kestiriminin doğruluğunu ve BÇAT tasarımının verimliliğini en üst düzeye çıkardığını belirtmektedir. Fakat aynı zamanda BÇAT tasarımının karmaşıklığının arttığını da ifade etmişlerdir. Armstrong ve Edmonds (2004), modül sayısı üzerine yaptığı çalışmada genel olarak bir aşamada en fazla dört modülün istendiğini ve üç aşamanın da yeterli olduğunu belirtmektedir.

BÇAT literatüründe en çok araştırmacı ilgisini çeken konuların arasında BÇAT tasarımının nasıl göründüğü yer almaktadır. Literatürde çalışılmış pek çok farklı aşama ve modül tasarımına sahip BÇAT tasarımı yer almaktadır. Bu tasarımlar Tablo 2'de sunulmuştur.

Tablo 2

BÇAT Desenleri Üzerine Yapılan Araştırmalar

BÇAT Tasarımı	Yapılan Araştırmalar
1-3	Reese, Schnipke & Luebke 1999; Schnipke & Reese, 1999; Wang, Fluegge & Luecht, 2012; Luo & Kim, 2018
1-2-2	Breithaupt & Hare, 2007; van der Linden ve diğerleri, 2007; Zenisky, 2004; Wang ve diğerleri, 2012; Park, 2015
1-3-3	Dallas ve diğerleri, 2012; Edwards, Flora & Thissen 2012; Hambleton & Xing, 2006; Jodoin, Zenisky & Hambleton, 2006; Keng & Dodd, 2009; Luecht, Brumfield & Breithaupt, 2006; Zenisky, 2004; Cetin-Berber, Sari & Huggins-Manley, 2019
1-2-3	Armstrong & Roussos, 2005; Zenisky, 2004; Yan, Lewis & von Davier, 2016; Wang ve diğerleri, 2012; Svetina ve diğerleri, 2019
1-3-2	Zenisky, 2004
1-3-4	Wang ve diğerleri, 2020
1-1-2-3	Belov & Armstrong, 2008; Weissman, Belov & Armstrong, 2007
1-5-5-5-5	Davey & Lee, 2011
5-5-5-5-5-5	Crotts, Zenisky & Sireci, 2012

Test ve Modül Uzunluğu

Test tasarımının yapısında olduğu gibi, test ve modül uzunluğuna ilişkin literatürde S-BÇAT uygulamalarına dair kapsamlı bir rehberlik sağlayan pek çok araştırma vardır. Basit bir geçerlik perspektifinden bakıldığında, bir testin uzunluğu, testin amacının gerektirdiği düzeyde kapsamı sağlayacak ve yeterli ölçme kesinliğini içerecek kadar uzun olmalıdır. Test uzunluğu, aynı zamanda madde havuzunun kalitesinden de yakından etkilenir (Hambleton & Xing, 2006). Daha kaliteli maddeler (yüksek bilgi veren, yüksek ayırt edici parametreli maddeler) yüksek verimli ölçmeyi destekler, ancak böyle maddelere madde havuzundan her zaman erişmek zor olabilir. Benzer şekilde bir S-BÇAT'ın ölçme hassasiyeti, aşamaların ve modüllerin düzenlenmesi ile yakından ilişkilidir. Daha az yönlendirme noktası olan bir test, istenen ölçme kesinliğini elde etmek için daha fazla yönlendirme noktası olan bir S-BÇAT'a göre daha fazla madde uygulanmasını gerektirebilir.

S-BÇAT üzerine test uzunluğunu inceleyen Stark ve Chernyshenko (2006), 40 veya 60 maddelik test uzunluğunu incelemişlerdir. 60 maddelik test uzunluğu, 40 maddelik test uzunluğundan daha yüksek ölçme kesinliği sunsa da bu iki uzunluğun karşılaştırılabilir düzeyde olduğunu belirtmektedirler. Modül uzunluğu ise araştırmalarda ilgilenilen başka bir husustur. Araştırmacılar, modül uzunluğunun aşamalara göre değişiminin etkilerini incelemişlerdir. Luecht ve Nungester (1998), 1-3-3 gibi tasarımlarda 2. ve 3. aşamalarda fazla modül olmasından dolayı ilk aşamada madde sayısının fazla olmasının 2 ve 3. aşamalardaki madde sayısını azaltacağını, dolayısıyla madde havuzunun daha etkili kullanılacağını belirtmektedir. Kim ve Plake (1993) ise ilk aşamada fazla madde kullanımının ölçme kesinliğini arttırdığını ifade etmektedir.

Yönlendirme Kuralı

Bir S-BÇAT geliştirirken verilmesi gereken önemli kararlar arasında, modüller arasında testin akışını yönlendirmek için hangi mekanizmanın kullanılacağı yer alır. Yönlendirme kuralı, S-BÇAT'ın amacına ve tasarımına bağlı olarak oldukça farklı olabilen, seçilen kuralları kullanarak önceki modül(ler)deki performanslarına dayalı olarak, sınava

girenleri farklı yollara veya sonraki aşama modüllerine yönlendiren veya sınıflandıran süreçtir.

Yönlendirme kuralı üzerine yapılan araştırmalar, her bir sınava giren kişi test boyunca ilerlerken, bir aşamada sınava girenlere modüllerin atanmasının nasıl optimize edileceği konusunda rehberlik sağlamaya çalışmıştır. Bir birey için yeni modülün atanması yetenek kestirimi ve madde (modül) kullanım oranı için çok büyük etkilere sahiptir. Uygulanan yönlendirme kuralları, sınava giren ve modül arasındaki eşleşmeyi belirler ve bu nedenle bu karar, bir S-BÇAT'tan elde edilen sonuçların kullanışlılığı için önemlidir.

S-BÇAT literatüründe temel olarak iki farklı yönlendirme kuralı öne çıkmaktadır. Bunların ilki norm referanslı yönlendirmeye dayanır. Bu yöntemde önceden belirlenmiş oransal dağılıma göre modüller arasında bireylerin yetenek kestirimlerine göre yeni modül ataması yapılır. Sınava giren bireyler yetenek tahminlerine göre sıraya koyulduğunda norm referanslı olarak bir karar verme süreci işler. Örneğin üç modüllü bir aşamada önceden belirlenen 30-40-30 oranına göre sınava giren bireylerin yüzde 30'una kolay güçlükte modül, %40'ı orta güçlükte modül, %30'u zor güçlükte modül atanır. Pek çok seçenek mevcuttur ve katılımcıların modüller arasında simetrik veya eşit olarak bölünmesi de gerekmez. Bu yöntemin avantajı modül (madde) kullanım oranlarını belli bir düzeyde tutmaya imkân sunmasıdır. Diğer taraftan bu yöntemin dezavantajı ise sınava katılanların yetenek dağılımı hakkında herhangi bir bilgi bulunmadığında kullanılacak oransal dağılımın bireyleri uygun olmayan modüllere yönlendirebilmesidir. Dolayısıyla hem yetenek kestirimleri hatalı olabilmekte hem de BBT ve S-BÇAT'in temel çıkış noktası olan bireyin yetenek düzeyine uygun test uygulanması özelliğinden faydalanılamamaktadır.

Bir diğer yönlendirme kuralı olan kriter referanslı bir yönlendirme yönteminde ise bilgiyi maksimize eden bir yaklaşım benimsenmektedir. Sınava giren kişinin yetenek düzeyinin bazı göstergeleri (yetenek kestirimi veya doğru sayısı) bir kural ile karşılaştırılır ve sınava giren kişiye bu karşılaştırmaya göre bir modül atanır. Kriter referanslı yönlendirme kuralı temel olarak iki yöntem altında incelenmektedir. Birincisi MTK'ya dayalı olan ve bireye

uygulanacak modül seçiminde bilgiyi maksimize edecek yöntemdir. Bu yöntemde bireyin yönlendirme modülünden sonra yeteneği hesaplanır ve birey yetenek düzeyine göre maksimum bilgiyi içeren modüle yönlendirilir. İkinci yöntem ise önceden belirlenen kesme puan aralıklarına ya da doğru sayısına göre yönlendirme yapar. Bu yöntemde bireyin yeteneği kestirilir ya da doğru sayıları toplanır ve bu yetenek düzeyi (veya puanı) modüllere atamada kriter olarak kullanılır. Örneğin yetenek düzeyine göre üç modülden oluşan bir aşama için -0.5 ve 0.5 yetenek düzeyleri kesme puanı olarak belirlenmişse, yönlendirme modülünden kestirilen geçici yetenek düzeyi -1.5 olan sınav katılımcısı ikinci aşamada -1.15 <-0.5 olduğundan kolay modüle yönlendirilecektir. Toplam puana örnek olarak ise 10 soruluk bir yönlendirme modülünün uygulandığını düşünürsek, 0-3 doğrusu olan katılımcılar kolay güçlükteki modüle, 4-7 doğrusu olanlar orta güçlükteki modüle, 8-10 doğrusu olanlar zor güçlükteki modüle yönlendirilir. Bilgiye dayalı yaklaşım, uygulanan maddelerin özelliklerini de dikkate alarak daha iyi bir katılımcı-modül eşleşmesi sunarken, doğru sayısı yaklaşımı daha kolay anlaşılan ve karşılaştırılabilir sonuçlar sağlar (Armstrong, 2002; Davey & Lee, 2011).

Madde Havuzu ve Hedefler

Madde havuzunda yer alan maddelerin kalitesi ölçmeden elde edilecek kestirimlerin doğruluğunu ve tutarlılığını etkilemektedir. Özellikle S-BÇAT'ta spesifik, istatistiksel ve istatistiksel olmayan hedeflerin karşılanması ve otomatik test birleştirme süreçlerinin iyi işlemesi için oldukça iyi hazırlanmış bir madde havuzu gerekmektedir. İstatistiksel hedefler, her modül içindeki tüm maddelerin olası yetenek aralıklarının belirlenmesini ve hedeflenen modül bilgi fonksiyonlarını ve genel hedef test bilgi fonksiyonlarını içerebilir. İstatistiksel olmayan hedefler ise modüllerde yer alan maddelerin sınavın kapsamına göre dağılım kurallarını ve modüllerde yer alan farklı madde türü dağılımı kurallarını içerir (Melican, Breithaupt & Zhang 2010). Madde havuzu geliştirme üzerine pek çok çalışma, S-BÇAT'tan elde edilen ölçüm sonuçlarında bir değişken olarak madde havuzunun istatistiksel kalitesine odaklanmıştır. Xing ve Hambleton (2004), madde havuzundaki maddelerin ayırt edicilik

değerlerinin etkisini zayıf ve güçlü madde havuzları ile incelemiştir. Daha iyi madde havuzlarının doğruluk ve tutarlılık açısından çok net bir şekilde daha avantajlı olduğu sonucuna ulaşmıştır. Benzer bir başka çalışmada, Wang, Fluegge ve Luecht (2012), iyileştirilmiş madde havuzunun daha iyi sonuçlar sağladığını tespit etmiştir.

Kapsam Dengelemesi

Bir test oluşturulurken, madde havuzunda yer alan maddelerin her birinin temsil ettiği içerik (content) alanındaki maddelerin dengeli bir şekilde testte temsil edilmesini sağlamak önemlidir. S-BÇAT uygulamalarında kapsam dengelemesi gibi istatistiksel olmayan hedefi sağlamak için madde özelliklerinin dağılımını kontrol etmek amacıyla çok sayıda kısıtlamaya gitmek gerekmektedir. Kısıtlamaların sayısı arttıkça karmaşa artabillir ve bir takım algoritmalara ihtiyaç duyulur.

S-BÇAT'ta kapsam dengelemesi otomatik test birleştirme işlemi aşamasında gerçekleştirilir. Modül, aşama ve panellerin oluşturulması ile test formları sabit (fixed) hale gelmekte, dolayısıyla kapsam dengelemesi sağlanmış olmaktadır. Bir aday, hangi panelde ilerlerse ilerlesin, hangi modüle geçiş yaparsa yapsın, sınavdan önce belirtilen oranlarda farklı kapsamlardan maddeleri yanıtlamış olacaktır (Diao & van der Linden, 2011).

Madde Kullanım Sıklığı (Item Exposure Rate) ve Madde Güvenliği

Ölçmede ve bilgisayar tabanlı test uygulamalarda ilgilenilen tüm konular arasında test ve madde güvenliği en önemli hususlardan biridir. Test ve madde güvenliği için maddelerin çok sık kullanılmaması önemlidir. Bu manada S-BÇAT'ta da bireylerin madde ve modül bazında madde kullanım sıklığı riskini en aza indirmeye yönelik stratejiler geliştirilmeye çalışılmaktadır (Georgiadou, Triantafilou & Economides, 2007; Stocking & Lewis, 2000). Aslında madde kullanım sıklığı kontrolü, S-BÇAT için BBT'den çok daha basit bir problemdir. Çünkü S-BÇAT'ta önceden ayarlanmış seviyelerde modüller bulunmakta ve katılımcılara modüller sunulmaktadır. S-BÇAT'ın yapısal birimi olarak modüllerin kullanılması madde kullanım oranı açısından farklı bir şekilde düşünmeyi gerektirmektedir.

Davey ve Lee (2011), S-BÇAT'ta madde güvenliğini sağlamak ve madde kullanım oranlarını düşük düzeyde tutmak için üç unsura odaklanmaktadır. Birincisi, S-BÇAT'da çoklu paneller içerisinde çok aşamalı yapılar aracılığıyla test uygulaması esnasında yanıtların kopyalanması veya soruların açığa çıkması olasılığı en aza indirgenebilir. İkincisi, S-BÇAT'ta otomatik test birleştirme özellikleri değiştirilerek madde kullanım oranlarını üst düzeye çıkaran maddelerin aşırı kullanımını sınırlamak için maddelerin yeniden kullanımına sınır getirilebilir. Üçüncüsü, madde bankasına periyodik olarak yeni maddeler eklenmeli, madde kullanım oranı yüksek maddeler gözden geçirilmeli veya bankadan çıkarılmalıdır. Böylece madde bankası zamanla yenilenecek ve güvenlik arttırılacaktır.

Yetenek Kestirimi

Doğrusal testlerde, sınav katılımcısının yetenek düzeyi testin sonunda kestirilirken, BBT ve S-BÇAT uygulamalarında bireye her uygulanan madde veya modülden sonra ve testin sonunda olacak şekilde birden fazla yetenek kestirimi yapılması gerekir. Her modül veya maddeden sonra kestirilen geçici yetenek düzeyi, bir sonraki modül veya maddenin belirlenmesi için kullanılmaktadır.

BBT ve S-BÇAT'ta yetenek kestirimi için MTK'ya dayalı olarak yapılmaktadır. MTK'da madde ve yetenek ölçeğinin aynı ölçek üzerinde ölçekleniyor olması uygulamalarda avantaj sağlamaktadır (Reckase, 1989).

MTK'ya dayalı yöntemlerden, Maksimum Olabilirlik Kestirim Yöntemi (ML- Maximum Likelihood), Maksimum Sonsal Kestirim Yöntemi (MAP) ve Beklenen Sonsal Kestirim Yöntemi (EAP) alan yazında sıklıkla kullanılmaktadır. ML yönteminde sınav katılımcısının n maddeye verdiği yanıtlara göre olabilirlik fonksiyonu hesaplanmakdır. Fonksiyonun maksimum olduğu nokta, sınav katılımcısının yeteneği olarak belirlenir. Olabilirlik fonksiyonun maksimum olduğu noktanın belirlenmesi için fonksiyonunun birinci türevi alınır. Fonksiyonun birinci türevinin sıfır olduğu değer katılımcının yetenek düzeyi (θ) olarak belirlenmektedir. ML yöntemi ile katılımcının yeteneğinin kestirilebilmesi için yanıt örüntüsünde en az bir yanlış ve doğru yanıtın bulunması gerekmektedir. Bireyin tüm

yanıtları doğru veya yanlış ise olabilirlik fonksiyonunu maksimum yapan nokta belirlenememektedir (Embreston & Reise, 2000).

MAP yöntemi, Samejima tarafından ML yöntemine alternatif olarak sunulmuştur. MAP yönteminde sınav katılımcısının yeteneği, n maddeye verdiği yanıtlara göre hesaplanan sonsal dağılım fonksiyonun maksimuma ulaştığı değerdir. Sonsal dağılım fonksiyonunu maksimum yapan değeri bulabilmek amacıyla fonksiyonun birinci türevi alınır. Fonksiyonun türevinin sıfıra eşit olduğu değer katılımcının θ değeri olarak belirlenir. Farklı bir deyişle, sonsal dağılım fonksiyonunun ortancası (modu) bireyin yetenek düzeyi olarak belirlenmektedir (Reckase, 2009).

EAP yönteminde ise sınav katılımcısının yetenek düzeyi sonsal dağılım fonksiyonunun ortalamasıdır. Sonsal dağılım fonksiyonunun ortalamasını hesaplamak için (- ∞ + ∞) aralığında dağılımın integrali alınır. Elde edilen değer bireyin yetenek düzeyi olarak belirlenir. ML ve MAP yöntemi iteratif yöntemlerdir. EAP ise iteratif olmadığından doğrudan bireyin yeteneğini kestirebilmektedir. Dolayısıyla EAP ile yapılan kestirimler daha hızlı yapılabilmektedir. Ayrıca, eğer sonsal dağılım fonksiyonu, tek modlu ve simetrik bir fonksiyon ise MAP ve EAP yöntemleri aynı yetenek düzeyini kestirmektedir (Reckase, 2009).

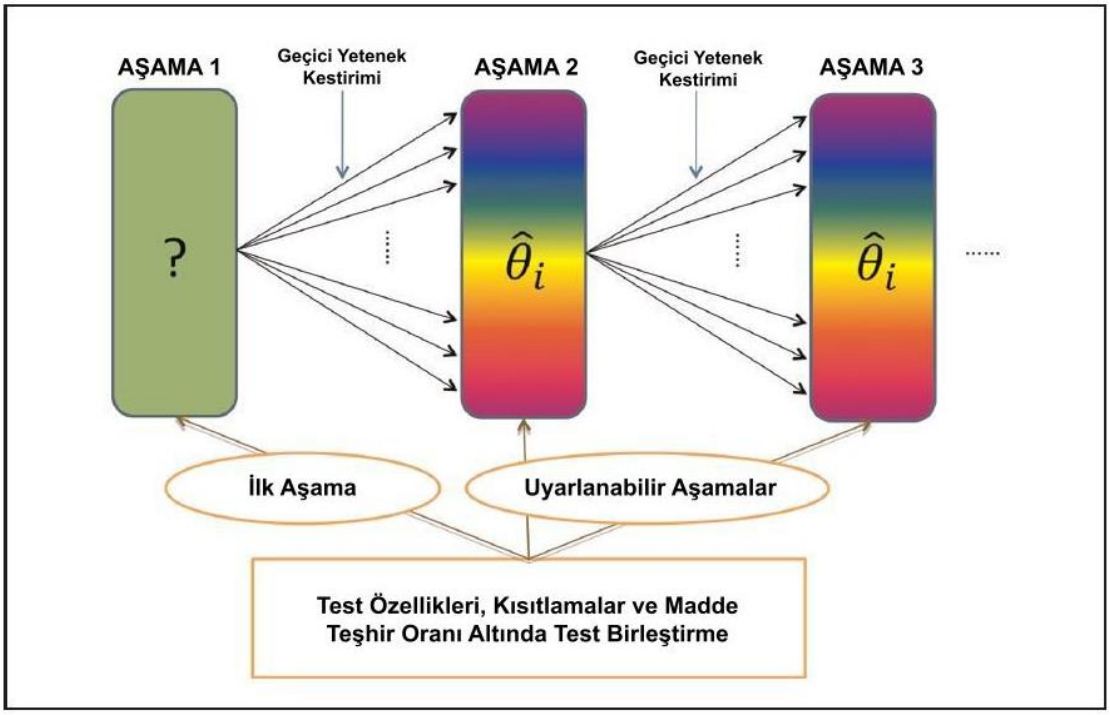
Anında Bireyselleştirilmiş Çok Aşamalı Testler (A-BÇAT)

S-BÇAT'a benzer şekilde, Anında Bireyselleştirilmiş Çok Aşamalı Testler'de (A-BÇAT) de test uygulaması aşamalar halinde uygulanır. A-BÇAT'ta, S-BÇAT'ın aksine, modüller sınav esnasında bireyin yetenek düzeyine göre anında birleştirilmektedir (Tay, 2015). A-BÇAT uygulamasında genellikle ilk aşamada tüm test uzunluğunda ve orta güçlük düzeyinde bir test birleştirilir ve ilk modül uzunluğundaki kısmı sınav katılımcısına uygulanarak katılımcının geçici yetenek tahmini hesaplanır. Ardından bireyin geçici yetenek düzeyine göre yeniden test birleştirmesi yapılır ve ikinci modül uzunluğunda madde sınav katılımcısına uygulanır. Bu durum, her bir sınava giren kişinin, her aşamadan sonra geçici yetenek parametresine dayalı olarak ikinci ve üçüncü aşamalarda farklı bir madde seti aldığı

anlamına gelir. Başka bir deyişle, A-BÇAT testleri, her sınava giren kişi için benzersiz bir şekilde uyarlanır. Bu işlem test sonlanıncaya kadar devam eder. Zheng ve Chang (2015), gölge testler ile A-BÇAT uygulamasının genel bir çerçevesini Şekil 6 üzerinde göstermektedir.

Şekil 6

A-BÇAT Genel Çerçevesi



Şekil 6'da üç aşamalı A-BÇAT uygulaması görülmektedir. İlk aşamada sınav katılımcısına orta güçlükte bir modül birleştirilerek uygulanır, ardından ikinci ve sonraki aşamalarda katılımcının geçici yetenek düzeyinde anında modüller birleştirilerek uygulanır. Şekil 6'da her ne kadar üç aşamalı bir A-BÇAT tasarımı sunulsa da, A-BÇAT'ta modül sayısı, modül uzunluğu istenildiği gibi kolayca değiştirilebilir. Özetle, A-BÇAT, her modüldeki madde sayısı ve aşama sayısı da dahil olmak üzere pek çok açıdan esnektir (Zheng & Chang, 2015). A-BÇAT'ta aşamalardaki maddelerin seçimi, maksimum test bilgisi, madde kullanım sıklığı kontrolü ve kapsam dengeleme gibi test özelliklerine dayalı olarak yapılabilir.

Madde Kullanım Sıklığı (Item Exposure Rate) ve Madde Güvenliği

Maddelerin madde havuzundan teste seçiminde Maksimum Bilgi Kriteri (MBK) yüksek düzeyde bilgi veren maddeleri daha fazla seçmeye meyillidir. Dolayısıyla belli bir grup madde yüksek düzeyde açığa çıkarken, diğer maddeler düşük düzeyde kullanılmaktadır. Bu durum yüksek riskli testler için, maddelerin aşırı kullanım sıklığına ulaşmalarına ve dolayısıyla madde güvenliği endişelerine yol açmaktadır. Düşük riskli testler için dahi maddelerin kullanımının dengesizliği madde geliştirme maliyeti açısından sürdürülemez bir hale gelebilir (Choi ve diğerleri, 2021).

Gölge testi yaklaşımında, madde kullanım sıklığının kontrolü teste uygunsuzluk (ineligibility) kısıtlaması eklenmesi ile sağlanabilir. Testten önce, mevcut sınav katılımcısının yetenek düzeyine uygun olan ve uygun olmayan maddeleri belirlenmesinde maddenin anlık kullanım sıklığına (Pr$\{A_i|\theta\}$) bağlı olarak maddenin uygunluk oranını (Pr$\{E_i|\vartheta\}$) belirlemek için Bernoulli deneyleri yapılır. Anlık olarak uygun olmayan madde grubu V_{inel} test birleştirmede gölge testinde şu şekilde kısıtlanır (Choi ve diğerleri, 2021):

$$\sum_{i \in V_{inel}} x_i = 0$$

Uygunluk olasılıkları, aşağıdaki yineleme fonksiyonu aracılığıyla sürekli olarak güncellenir:

$$Pr^{(I+1)}\{E_i|\theta\} = min\left[\frac{r^{max}}{Pr^{(I)}\{A_i|\theta\}} Pr^{(I)}\{E_i|\theta\},\ 1\right]$$

Formülde I=1,2,…,I sınav katılımcılarının sırası olmak üzere; Pr$\{E_i|\theta\}$, bir i maddesi için θ yetenek düzeyinde koşullu uygunluk olasılığı; Pr$\{A_i|\theta\}$, i maddesi için I sınav katılımcısına kadar olan koşullu madde kullanım sıklığı ve r^{max} ise madde kullanım sıklığının üst sınırıdır (örneğin $r^{max} = 0.25$).

Fonksiyon, olumsuz bir geri besleme mekanizması içerir. Belirli bir madde için kullanım sıklığı oranı Pr$\{A_i|\theta\}$, r^{max}'tan fazla olduğunda, bir sonraki sınava giren için

$Pr^{(I+1)}\{E_i|\theta\}$ bu maddenin uygunluk olasılığını aşağı doğru ayarlar. Madde kullanım sıklığı r^{max}'tan küçükse, uygunluk olasılığı yukarı doğru ayarlanır (Choi ve diğerleri, 2021).

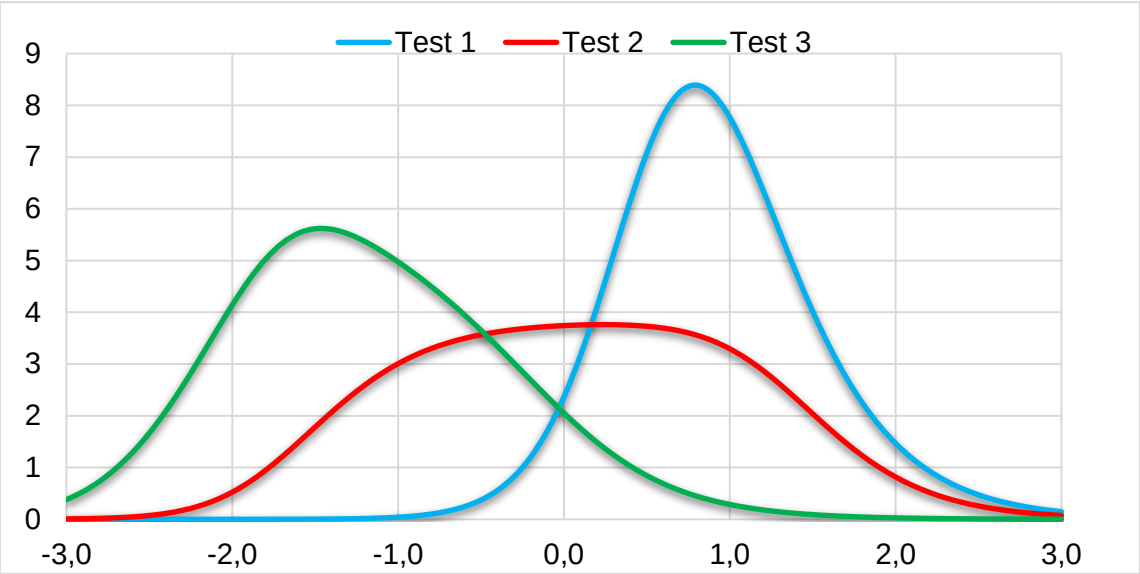
Test Birleştirme (Test Assembly)

Test birleştirme, istatistiksel olan ve istatistiksel olmayan kısıtlamaların doğrultusunda madde havuzundan kısıtlama şartlarını taşıyan maddelerin seçilerek testin oluşturulması işlemidir (Luo, 2020).

Test birleştirme işlemi; testin amacı, kapsamı ve değerlendirme türü birlikte ele alınarak gerçekleştirilmelidir. Şekil 7'de geniş bir madde havuzundan test uzunluğu 10 madde olacak şekilde üç farklı şekilde birleştirilmiş testlerin test bilgi fonksiyonu grafik üzerinde gösterilmektedir. Test 1, kesme puanı $\theta = 1$'nda maksimum bilgi düzeyine ulaşmış olup belli bir kesme puanına göre adayların minimum yeterlik düzeyini sağlaması istenen bir kabul sınavı için uygun görünmektedir. Test 2 ise öğrencilerin dönem içerisinde başarı seviyelerini belirlemek, başarı notu vermek ya da hedef/davranışları kazanıp kazanmadığını belirlemek için $\theta = [-1, 1]$ aralığında maksimum bilgi düzeyine ulaşan bir test olduğundan bu amaçlarla kullanılabilir. Diğer taraftan, tanılayıcı değerlendirme bağlamında düşük puanlı veya yetersiz bireylerin belirlenmesi amacıyla yetenek ölçeğinin alt ucunda maksimum bilgi miktarına ulaşan Test 3 testinin birleştirilmesi mantıklı görünmektedir.

Şekil 7

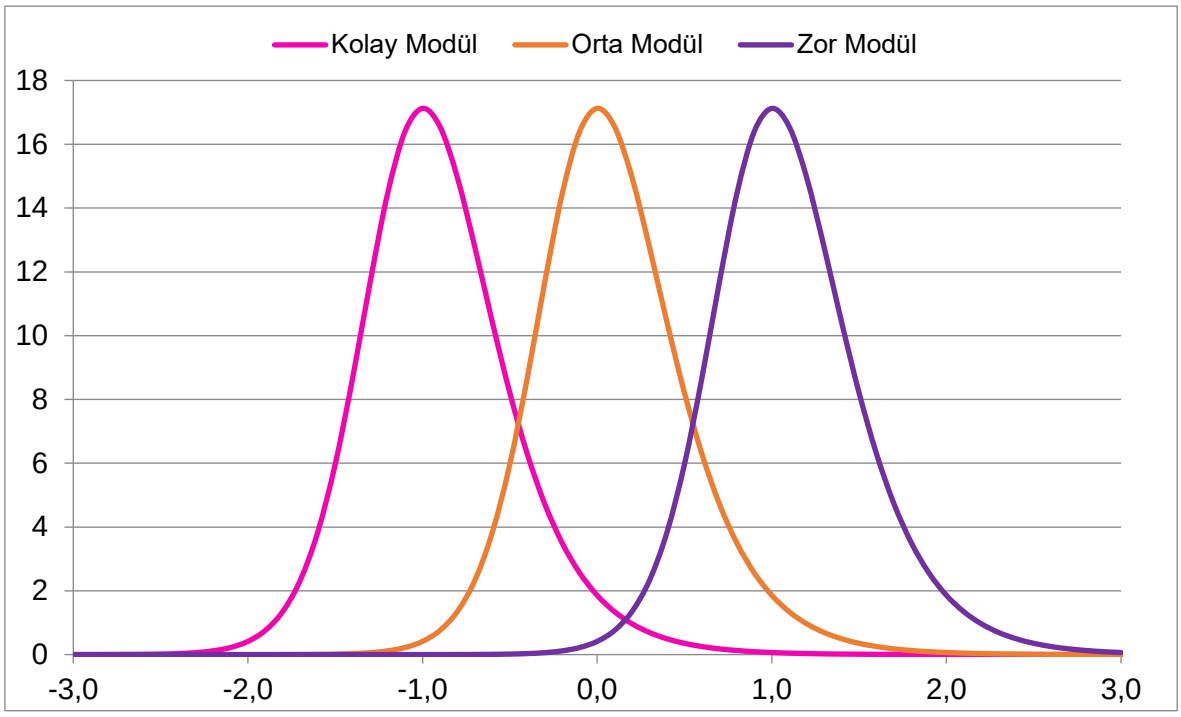
Üç Farklı Test Birleştirme Örneği



Test geliştiricileri, test uygulamasından önce kısıtlamaları dikkate alarak test formlarını oluşturabilir. Ancak madde kullanım sıklığının belli düzeyde tutulması, kapsam dengelemesi, paralel modül ve paralel panellerin aynı anda oluşturulması gibi kısıtlamaların da yer alması durumunda testin manuel olarak birleştirilmesinin güçleştiği görülür. Örneğin S-BÇAT uygulamalarında aynı aşamada yer alan modüllerin test özellikleri, kapsamı, kısıtlamaları ve bilgi düzeyleri açısından psikometrik olarak birbirine paralel olması gerekir. Şekil 8'de sırasıyla $\theta = -1$, $\theta = 0$ ve $\theta = 1$ yetenek düzeylerinde birbirine paralel olacak şekilde birleştirilmiş kolay, orta ve zor modülün bilgi fonksiyonları görülmektedir.

Şekil 8

Modül Bilgi Fonksiyonları



Şekil 8'de yer alan modüllerin elle (manuel) birleştirilmesi düşünüldüğünde hem belirli kısıtlamalar ve test özelliklerinin sağlaması hem de modüllerin psikometrik açıdan birbirine paralel olmasını sağlamak oldukça zordur. Dolayısıyla maddeleri belirli özelliklerine göre testlere seçen algoritmalara ihtiyaç duyulmaktadır. Kısıtlamaların getirdiği karmaşaların çözüme kavuşturulması için otomatik test birleştirme (OTB) yöntemleri kullanılmaktadır (Li & van der Linden, 2018).

Otomatik Test Birleştirme

Geçerlik perspektifinde, birden fazla birbirine paralel test formu birleştirmede test deneyimi ve uygulaması açısından temel kriterler; test formlarının adil olması, karşılaştırılabilir olması, psikometrik olarak eşdeğer olması, belirlenen hedef ve kısıtlamaları karşılamasıdır. Geleneksel test birleştirme yöntemi, büyük ölçüde test geliştiricilerinin deneyimlerine ve içgörülerine dayanır; bu nedenle, test geliştiricilerinin kendi bilişsel süreçlerinin, deneyimlerinin (ya da deneyimsizliklerinin) teste olan etkilerinden

kaçınılamaz ve adil, eşit, güvenilir ve en uygun şekilde test formlarının birleştirilmesi pek olası değildir. Özellikle çoklu panel ve modül yapısının karmaşıklığı (örneğin 4 paralel panel içerisinde 36 paralel modül birleştirme işlemi) düşünüldüğünde test birleştirme işini bilgisayar algoritmaları ile gerçekleştiren Otomatik Test Birleştirme (OTB) yöntemlerinin önemi daha da ortaya çıkmaktadır (Luo, 2020).

OTB, özellikle çoklu paralel paneller açısından test programlarının karmaşık ihtiyaçları göz önüne alındığında, BÇAT için pek çok avantaj sunar. Bu nedenle, tüm bilgisayar tabanlı testlerin OTB kullanılarak oluşturulması bir dizi test spesifikasyonunu hızlı ve doğru bir şekilde karşılama potansiyeli nedeniyle yalnızca bir zaman meselesi gibi görünmektedir. Tabii ki çok fazla test kısıtlaması OTB için sorun yaratabilir, bu yüzden hedeflerin belirlenmesinde gerçekçi olmak gerekir. Örneğin kapsam dengelemesi, test bilgi hedefi, sözcük sayısı, madde türü gibi kısıtlamalarla birlikte uygun bir otomatik test birleştirme çözümünün elde edilmesi, madde havuzunun kısıtlamalarla uyumuna bağlıdır (Luecht, Brumfield & Breithaupt, 2006).

Metodolojik açıdan, OTB algoritmalarının iki alt başlıkta ele alınabilir: (i) buluşsal (heuristic) yöntemler ve karma tamsayılı programlama (mixed integer programming - MIP). Buluşsal yöntemler, test birleştirme işlemini test formuna maddeleri sıralı olarak ekleyecek şekilde yerel optimizasyon problemine böler. Bu yöntemler, maddeleri sırayla seçtiğinden genelde madde bilgisi yüksek olan maddeleri seçmeye daha meyillidir (Zheng ve diğerleri, 2016). Yani, ilk birleştirilen test, tüm madde havuzunda yer alan daha kıymetli maddelere erişirken, sonraki birleştirilen formlar azalmış madde havuzundan birleştirilmek durumunda kalır. Dolayısıyla bu yöntemlerin bu davranışını dengelemek için farklı stratejilere ihtiyaç vardır. Diğer taraftan, buluşsal yöntemler, tüm olasılıkları değerlendirmeden çözüm için bir kısayol bulmaya yöneliktir. Dolayısıyla buluşsal yöntemler, çözülecek bir problemin optimal (en uygun) çözüm yöntemini bulmaz, bu yöntemlerin temelinde en uygun çözümden ziyade çalışan bir çözüme ulaşmak hedeflenmektedir (Luo, 2020). MIP yöntemleri ise, buluşsal yöntemlerin aksine, yüksek boyutlu 0-1'li bir arama uzayını sistematik olarak tarar ve en

uygun çözüme ulaşmayı hedefler. MIP genellikle buluşsal algoritmalardan çok daha fazla hesaplama içermesine rağmen, bulduğu çözüm teorik olarak tüm arama uzayında mümkün olan en iyi çözümdür. Dolayısıyla bu araştırmada test birleştirme işlemlerinde MIP yöntemi kullanılmıştır.

Test Özellikleri ve Kısıtlamalar

Test birleştirmenin en uygun biçimde yapılmasında en önemli adım, test kısıtlamalarının kesin, eksiksiz ve birbiriyle çakışmayacak şekilde tanımlanmasına bağlıdır. Eksik test kısıtlamaları, muhtemelen verimsiz çözümler ile sonuçlanacaktır. Örneğin, aynı anda bir araya getirilecek paralel formlarda madde çakışmasını engellemenin unutulması halinde, aynı maddenin farklı formlarda yer alması bizi şaşırtmamalıdır. Benzer şekilde, gereksiz kısıtlamaların da modelde yer almaması çözüme ulaşmak açısından daha verimli olacaktır (van der Linden, 2018).

Test kısıtlamaları altında oluşacak test formunun karakteristik nitelikleri “test özellikleri (test specification)” olarak tanımlanır. Örneğin, testte yer alacak kapsamlardan madde sayıları, maddelerin özellikleri, testin güçlüğü vb. değişkenler test özellikleridir. Test özelliklerinin iki temel bileşeni vardır: a) nitelik b) gereklilik. Bu iki temel bileşenin her ikisi de test birleştirme probleminin bir parçası olarak test özelliklerini modellemek için matematiksel ifadeleri tanımlamaya yardımcı olurlar (van der Linden, 2005). Nitelikler, “nicel düzey” ve “tür” olarak ikiye ayrılır. Testte yer alacak nicel düzeylerin belirlenmesi için (i) madde, (ii) uyaran, (iii) madde kümesi, (iv) test ve (v) çoklu formlar düzeyinde test özellikleri belirlenmelidir. Maddelerin nicel düzeylerine örnek olarak kapsam dağılımları, madde formatı, test uzunluğu, TBF değerleri, toplam kelime sayıları, cevap anahtarı dizileri, maddelerin psikometrik parametreleri, beklenen toplam yanıtlama süreleri örnek olarak sıralanabilir. Test formlarında ortak uyaranlar (common stimulus) yer alabilir. Ortak uyaran, birkaç maddenin bir metin, grafik ya da tablo üzerinde birbirine bağlanan ortak köklü soruları ifade eder (örneğin okuma pasajları, vaka çalışmaları, veri kümeleri, grafik yorumlama soruları vb.). Ortak uyaranların nicel düzeyleri ise kapsam dağılımı, ortak uyaranların sayısı,

madde türleri, maddelerin güçlükleri olarak örneklendirilebilir. Aynı anda birden fazla form birleştirilecekse formların nitelikleri arasındaki benzerlik ve farklılıklar açıkça belirlenmelidir. Çoklu formlar için nicel düzey çoklu formların güçlük düzeyleri, maddeler veya uyaranlar arasındaki örtüşmeler olarak örnek verilebilir.

Test özelliklerinin türlerine gelince, (i) kategorik ve (ii) mantıksal nitelikler olarak sınıflandırılabilir (van der Linden, 2018). Testin oluştuğu madde havuzu, madde tipi, cevap formatı, madde yazarları, maddelerin bilişsel seviyeleri vb. özelliklerine göre kategorik olarak kolaylıkla gruplanabilir ve bu gruplamalardan oluşturulacak test kısıtlamaları kolayca modellenebilir. Fakat mantıksal (Boolean) türdeki test kısıtlamalarını modellemek daha zordur. Mantıksal nitelikler, maddelerin birbirleriyle ilişkileri içerisinde sahip olduğu niteliklerdir. Örneğin, düşman maddeler (içeriklerinin birbirine benzer olması sebebiyle bir maddenin diğer maddenin çözümüne işaret ettiği maddeler), ortak köklü maddeler, dışlanan maddeler (örneğin bpis<0.20 ise maddenin teste alınmaması), şartlı maddeler (örneğin hem Kapsam 2'de yer alan hem de ayırt ediciliği 0.9 üzerindeki maddeler) mantıksal türde test özellikleri olarak sıralanabilir.

Test özelliği formüle edilen "gereksinim"ler iki başlık altında ele alınabilir: (i) hedef ve (ii) kısıtlama. Bir test özelliği için hedef, havuz için mümkün olan bir niteliğin maksimum veya minimum değerini şart koşmasıdır. Örneğin, birleştirilecek bir testin $\vartheta = 1$ noktasında maksimum bilgiye ulaşmasının istenmesi bir hedeftir. Diğer taraftan bir test özelliğine alt ya da üst sınır getirilmesi bir kısıtlamadır. Hedef olan test özellikleri, "mümkün olduğunca en çok", "en büyük" ve "en küçük" gibi terimlerle kolayca tanınır. Kısıtlamalar ise tipik olarak "arasında", "belirli aralıklarda" veya "en fazla" gibi terimleri içerir (van der Linden, 2018).

Prensip olarak, test birleştirme modellerinde kısıtlama sayısı için bir sınır yoktur. Gerekli olmayan kısıtlamaların modele eklenmesi, modelin çözümüne bir zarar vermeyebilir, fakat madde havuzundan elde edilecek çözümler kümesini daraltır. Gerekli olmadıkça modele kısıtlama eklenmemelidir. Diğer bir mevzu, eklenen kısıtlamaların kendi aralarında çelişmemesi veya madde havuzunun karşılayamayacağı özellikleri içermemesi

gerektiğidir. Kendi aralarında çelişen (örneğin iki kapsamdan alınacak madde sayısının test uzunluğundan fazla olması vb.) veya madde havuzunun karşılayamayacağı kısıtlamalar çözümü imkânsız hale getirir. Her ne kadar kısıtlama ve hedef sayısına dair bir sınır olmasa da, her eklenen hedef fonksiyonu, maddelerin seçiminde farklı bir optimizasyon yüklediğinden aynı anda birden fazla fonksiyonu optimize etmek işleri güçleştirmektedir.

Araştırmanın bu bölümünde ifade edilen hedefler ve kısıtlamalar altında bir test birleştirme örneği aşağıda sunulmuştur.

Test Birleştirmenin Yapısı

Testin verimli ölçümler sunması için madde havuzundan seçilerek testi oluşturan maddelerin hem test özelliklerini ve kısıtlamalarını sağlaması hem de madde havuzunda yer alan maddelerden en uygun (optimal) olanlarının seçilmesi gerekir. Test birleştirme işleminde en uygun maddenin seçilmesi sorunsalı her ne kadar özel görünse de, benzer sorunlarla sanayide, endüstride, iş hayatında karşılaşılmaktadır. Örneğin üretim planlamasında, havayolu ekipman planlamasında, elektronik çip tasarımında, paketleme ve kargolamada, iş atamasında ve nakliyede benzer sorunlar yer almaktadır. Bu sorunların ortak noktası, eldeki nesnelerin, faaliyetlerin veya seçeneklerin farklı kombinasyonlarının bazı hedefler ve kısıtlamalar kümesi altında en uygun çözümüne ulaşarak çözülebilmesine dayanır. Bu sorunlardan bazıları en basit haliyle, bir birey alışverişe çıktığında harcayacağı para miktarı ile maksimum fayda sağlayacak ürünleri almaya çalışırken, bir öğretmen bir ders süresinin verimliliğini en üst düzeye çıkarmak için öğrencilerin konuyu daha rahat anlayabileceği etkinlikleri ve örnekleri seçerken, bir kargo çalışanı, elindeki paketleri dağıtmak için yoldan ve zamandan tasarruf etmek için en uygun yolu araştırırken benzer sorunlarla karşı karşıyadır. Bu tarz sorunlar, kombinatoryal optimizasyon problemleri (combinatorial optimization problems) olarak bilinir (van der Linden, 2018). Bu problemlerin her biri aşağıdaki dört unsurdan oluşur: (i) sonlu bir nesneler, faaliyetler veya seçenekler havuzu; (ii) bunların bir veya daha fazla kombinasyonunu seçme görevi; (iii) yerine getirilmesi gereken bir dizi kısıtlama; ve (iv) seçimin maksimum veya minimum olması

gereken bir hedefi. Bu dört unsurla ilgili optimizasyon problemleri, tipik olarak karma tamsayılı programlama (mixed-integer programming - MIP) metodolojisi kullanılarak çözüme ulaşılmaktadır. MIP uygulamaları her zaman aşağıdaki üç adımı izler: (i) genellikle karar değişkenleri olarak adlandırılan, problemin çözümünü kontrol eden değişkenlerin tanımı; (ii) özellik ve kısıtlamaların her birini modellemek için değişkenleri kullanmak; ve (iii) amaç fonksiyonunu optimize eden ve tüm kısıtlamaları karşılayan karar değişkenleri için değerlerin kombinasyonu için modelin çözülmesi.

Bir test birleştirme probleminin nasıl çözülebileceği aşağıda bir örnek ile incelenmiştir. i=1,2,…,100 maddelerine sahip ve kalibre edilmiş bir madde havuzu olduğunu ve bu maddelerin V1, V2, ve V3 olmak üzere farklı kategorilere ait olduğunu varsayalım.

İlk adım, bu test birleştirme problemi için karar değişkenlerini tanımlamaktır. Karar değişkeni ile bir maddenin test formuna seçimi, şu şekilde tanımlanan ikili değişkendir:

$$x_i = \begin{cases} 1, & eğer\ madde\ teste\ seçilirse \\ 0, & aksi\ takdirde \end{cases} \qquad (1)$$

$$i = 1, \ldots, I$$

Tablo 3, madde havuzundaki 100 maddeden elde edilebilecek tüm olası test formlarının nasıl tanımlandığını gösterir. Tablodaki satırların her biri farklı bir olası test formunu temsil eder. 100 maddelik bir madde havuzundan 2^{100} farklı test formu oluşturulabilmektedir. Test birleştirme sorununun çözümü, bu olası farklı test formlarından hedefler ve kısıtlamalar göz önüne alınarak en uygun (optimal) olanı belirlemektir.

Tablo 3

100 Maddelik Havuzdan Oluşturulabilecek Tüm Olası Test Formları

100 Maddelik Bir Havuzdan Olası Tüm Test Formlarının Seçimi için 0–1 Karar Değişkenleri							
Madde	1	2	…	i	…	99	100
Değişken	x_1	x_2	…	x_i	…	x_{99}	x_{100}
Form 1	0	0	…	…	…	0	0
Form 2	0	0	…	…	…	0	1
Form 3	0	0	…	…	…	1	1
…	…	…	…	…	…	…	…
Form 2^{100}	1	1	…	…	…	1	1

Test birleştirme problemi, aşağıda yer alan yedi farklı hedef veya kısıtlama altında bir test birleştirme işlemi olsun. Bu yedi farklı hedef veya kısıtlamayı sağlayacak şekilde bir test birleştirme işleminin gerçekleştirilmesi için hedef ve kısıtlamalar birbirine bağlı olarak aşağıdaki gibi modellenecektir.

1) Birleştirilecek test, $\theta = \theta_c$ yetenek düzeyinde en az T_k değerinde bilgi sunmalı,

$$\sum_{i=1}^{n} I_i(\theta_c)x_i \geq T_k \quad (1)$$

I_i, madde bilgi fonksiyonu, θ_c, maksimum bilgi miktarına ulaşılacağı yetenek düzeyi, T_k, test formu için hedeflenen bilgi miktarı, x_i, maddenin teste alınma durumunu belirten karar değişkenidir.

2) Madde havuzundaki V1 ve V2 kategorilerinden sırasıyla n1 ve n2 sayıda madde içermeli ve V3 kategorisinden hiçbir madde içermemeli,

$$\sum_{x_i \in V_1} x_i = n_1 \quad (1)$$

(V_1 kapsamından n_1 sayıda madde seç)

$$\sum_{x_i \in V_2} x_i = n_2 \quad (1)$$

(V_2 kapsamından n_2 sayıda madde seç)

$$\sum_{x_i \in V_3} x_i = 0 \quad (1)$$

(V_3 kapsamından hiç madde seçme.)

$$x_i \in \{0,1\},\ i = 1, \ldots, I \quad (1)$$

(değişkenlerin tanımı)

3) Soru başına toplam ortalama yanıtlama süresi 70'den küçük olmalı,

$$\sum_{i=1}^{n} q_i x_i < 70 * n \quad (1)$$

n, test formunda yer alacak madde sayısı, q_i, her maddenin ortalama yanıt süresi değişkeni, x_i, karar değişkenidir. Bu hedef ile oluşturulacak testte yer alan maddelerin ortalama yanıt süreleri toplamının 70*n'den daha küçük olması amaçlanmaktadır.

3) Madde havuzunda yer alan i_{23} ve i_{24} maddeleri testte beraber yer almalı (ya da hiç yer almamalı),

$$x_{23_i} - x_{24_i} = 0 \quad (1)$$

Bu mantıksal kısıtlama ile maddelerin arasındaki koşullu ilişkilerle ilgilenilmiştir. Bu örnekte i_{23} ile i_{24} maddesinin herhangi birinin seçilmesi durumunda diğer maddenin de seçilmesi, aksi takdirde her ikisinin de seçilmemesi istenmiştir ($i_{23} \leftrightarrow i_{24}$; $\leftrightarrow$: ancak ve ancak koşullu önermesi). Denklemdeki karar değişkenleri incelendiğinde, x_{23_i} ve x_{24_i}'nin her ikisi de aynı anda 0 veya 1 olduğunda denklemdeki eşitliğin sağlanacağı görülmektedir.

4) Madde havuzunda yer alan i_{38} ve i_{76} maddeleri aynı testte yer almamalı (düşman maddeler),

$$x_{38_i} + x_{76_i} \leq 1 \quad (1)$$

Bu mantıksal kısıtlama ile i_{38} ile i_{76} maddeleri düşman maddeler (enemy items) olarak modellenmiştir. Düşman maddeler, doğrudan testin geçerliğini düşüreceğinden, denklemde görüleceği üzere testte bu maddelerden herhangi birinin olması ya da hiçbirinin olmaması istendiği anlaşılabilir.

5) Madde havuzundan ortak köklü maddeler testte yer almalı,

Ortak köklü (uyaranlı) maddelerin teste alınmasında temel fikir, maddelerin ve uyaranların seçimini ayrı karar değişkenleri ile ayrı mantıksal kısıtlamalar kullanarak koordine etmektir. Madde havuzunda yer alan ortak köklü madde gruplarını s=1,….,S ile temsil edildiğini düşünelim. Maddeler de ortak köklü madde gruplarına göre s ile ilişkili olacak şekilde $i_s = 1, \dots, I_s$ ile temsil edilsin. s ortak köklü madde grubunun seçilme durumunu ifade eden karar değişkeni z_s;

$$z_s = \begin{cases} 1, & s\ ortak\ köklü\ maddesi\ \ teste\ seçilirse, \\ 0, & aksi\ takdirde \end{cases} \quad (1)$$

Maddelerin teste seçilmesi için kullanılan karar değişkeni;

$$x_{i_s} = \begin{cases} 1, & i_s\ maddesi\ teste\ seçilirse, \\ 0, & aksi\ takdirde \end{cases} \quad (1)$$

Yukarıdaki karar değişkenleri kullanılarak elde edilecek amaç fonksiyonunun testte yer alacak maddelerin ve ortak madde gruplarının seçiminde iki gereksinimi karşılayacak şekilde modellenmelidir: (i) madde havuzundaki bir ortak madde grubu seçilirse, o kümeden bir madde seçimi yapılmalıdır ve (ii) bir ortak madde grubu seçilmişse bu kümeden sınırlı sayıda madde seçilmelidir. Bu iki gereksinimi yerine getirmek için seçilecek madde sayılarını da içerecek şekilde amaç fonksiyonu aşağıdaki gibi oluşturulabilir:

$$\sum_{i=1}^{I_s} x_{i_s} \leq {n_s}^{(maksimum)} z_s,\ \ tüm\ s\ ortak\ madde\ grupları\ ve\ i_s\ için \quad (1)$$

$${n_s}^{(minimum)} z_s\ \ \leq \sum_{i=1}^{I_s} x_{i_s},\ \ tüm\ s\ ortak\ madde\ grupları\ ve\ i_s\ için \quad (1)$$

Yukarıda belirtilen gereksinimleri karşılamak için formülde yer alan z_s=1 ise s ortak köklü madde grubundan n_s değerleri aralığında madde seçilecektir. z_s=0 ise s ortak köklü madde grubundan madde seçilmeyecektir. Eğer ki ortak köklü madde grubundan seçilecek madde sayısı sabitlenecekse denklemde eşittir (=) işareti kullanılarak ve n_s kısmına bu sabit değer atanarak gerçekleştirilebilir. Yalnızca bir üst sınır durumu yalnızca ${n_s}^{(minimum)} = 0$'a ayarlanarak gerçekleştirilebilirken, yalnızca bir alt sınır belirlenmesi halinde ${n_s}^{(maksimum)}$

değeri havuzda mevcut olan ortak köklü maddelerin boyutundan daha büyük keyfi bir değer atanabilir.

6) Birbirine paralel olacak şekilde birden fazla test formu birleştirilmeli,

Birinci maddede yer alan hedef modellemesi tekli form içindir. Bu maddenin modeli üzerinde yapılacak modifikasyon ile çoklu form birleştirme işlemi gerçekleştirilebilir. Bunun için; (i) karar değişkenlerinin tanımı değiştirilmeli, (ii) amaç fonksiyonu ayarlanmalı ve (iii) paralel formlar arasındaki gerekli ilişkileri kontrol etmek için kısıtlamalar eklenmelidir. Burada yapılacak değişikliklere göre yukarıda yer alan diğer modeller de güncellenmelidir, bu çalışmada basit bir anlatım hedeflendiğinden diğer modeller güncellenmemiştir. f=1,2,….F olmak üzere paralel formları belirtsin. Karar değişkenleri şu şekilde yeniden tanımlanmalıdır:

$$x_{if} = \begin{cases} 1, & i\ maddesi\ f\ formuna\ seçilirse, \\ 0, & aksi\ takdirde \end{cases} \qquad (1)$$

Amaç fonksiyonu ise birinci maddede yer alan denklem 1'in şu şekilde değiştirilmesi ile basitçe elde edilebilir:

$$\sum_{i=1}^{I} I_i(\theta_c) x_{if} \leq T_k + y, \qquad tüm\ f\ formları\ için \qquad (1)$$

$$\sum_{i=1}^{I} I_i(\theta_c) x_{if} \leq T_k - y, \qquad tüm\ f\ formları\ için \qquad (1)$$

$$y > 0$$

Bu çözümde y değişkeni minimize edilerek formların bilgi miktarı farklarını minimuma indirmek amaçlanmaktadır. Dolayısıyla tüm olası test formları arasından hem $\vartheta = 1$ düzeyinde maksimum değeri veren ve hem de y değişkeninin minimum değerine ulaşması ile T_k düzeyinde bilgi sunan ve birbirine paralel formlar elde edilecektir.

7) Her madde paralel formlarda en fazla bir kez kullanılmalı (maksimum kullanım sayısı=1 olmalı),

$$\sum_{f=1}^{F} x_{if} \leq 1, \qquad tüm\ i\ maddeleri\ için \qquad (1)$$

Formlar arasında madde çakışması istenmiyorsa (her madde en fazla bir kez kullanılmak isteniyorsa) denklemdeki gibi maddelerin bir sefer kullanılması modellenebilir.

Yukarıdaki denklemler ile modellenen test birleştirme sorunun çözümü için MIP çözücüler kullanılmaktadır. MIP çözücüler, genellikle özel bir modelleme dili aracılığıyla modellerin girilmesinin ardından çözümleri döndürmeye izin veren yazılım programları tarafından bulunur. Çözümler, problem ön işleme, özel yapılardan yararlanma ve optimize edilmiş algoritmaların (genellikle dal ve kesme olarak bilinen algoritma versiyonları) bir karışımı kullanılarak hesaplanır (Chen ve diğerleri, 2010).

Test birleştirme uygulaması için kullanılabilen güçlü ticari çözücüler arasında IBM CPLEX (International Business Machine Corporation, 2010), Gurobi (Gurobi Optimization, Inc., 2017) yer almaktadır. Microsoft Excel ise kullanıcılarının elektronik tablo biçimini kullanarak test birleştirme modellerini formüle etmelerini sağlayan bir Premium Solver Platform eklentisine sahiptir (Cor ve diğerleri, 2008). Ayrıca ücretsiz MIP çözücüler de mevcuttur. R dilinde yazılmış için “lpSolveAPI” (Konis, 2016), “Rglpk” (Theussl ve diğerleri, 2019) ve “Rsymphony” (Harter ve diğerleri, 2021) paketleri örnek olarak sıralanabilir.

S-BÇAT için Test Birleştirme

Sabit Bireyselleştirilmiş Çok Aşamalı Testlerde (S-BÇAT) sınav katılımcıları, yetenek düzeylerine göre sınav uygulamasından önce birleştirilen testlerden farklı madde kümeleri alırlar. Panel ve modüllerden oluşan bir S-BÇAT modelinin verimli sonuçlar üretmesi için her aşamada yer alan modüllerin yetenek ölçeğini kapsayacak şekilde farklı noktalarda yüksek düzeyde bilgi sunan maddelerden oluşmalıdır. Ayrıca adilliği sağlamak için farklı panellerde yer alan aynı düzeydeki modüller psikometrik özellikleri açısından birbirine paralel olmalıdır. Diğer taraftan, panel ve modüller, birçok farklı olası yol boyunca tüm istatistiksel ve istatistiksel olmayan hedefleri ve kısıtlamaları karşılamalıdır. Otomatik Test Birleştirme (OTB) algoritmaları, test geliştiricileri üzerindeki yükün çoğunu azaltabilse de, bu algoritmalar, S-BÇAT tasarımının artan karmaşıklığına göre uyarlanmalıdır.

S-BÇAT'ın temel yapısı incelendiğinde, paneller ve modüllere dayalı bir test birleştirme prosedürü izlenmektedir. Panel, her biri birden fazla modülden oluşan birkaç uyarlanabilir aşamaya bölünmüştür. Aynı aşamadaki modüller farklı zorluk seviyelerinde sabitlenir. Test sırasında, sınava girenler önceki aşamalardaki performanslarına göre her aşamada en uygun modüle yönlendirilir. S-BÇAT geliştiricileri, genellikle madde ve test güvenliği veya madde havuzundaki maddelerin verimli kullanımı için birden çok paralel panel oluşturmak isterler. Eğer ki paralel panellerde yer alan modüller, test bilgi fonksiyonları veya diğer hedef ve kısıtlama kriterleri açısından yeterince benzerse paralel olarak kabul edilebilir (Zheng, Wang, Culbertson & Chang, 2016).

S-BÇAT için test birleştirme işlemi, panel tasarımının oluşturulması ile başlar. Panel tasarımında modüllerde yer alacak maddelerin özellikleri, test bilgi fonksiyonlarının tepe yapacağı yetenek düzeyleri ve bilgi miktarları, diğer hedefler ve kısıtlamalar belirlenmelidir. Panel tasarımı oluşturulduktan sonra S-BÇAT panellerinin birleştirilmesi genellikle iki adımda ilerler: (i) madde havuzundaki maddeler ile modüller birleştirilir, (ii) Elde edilen modüllerden paneller birleştirilir. Panellerin birbirine paralel olması için iki ana yöntem vardır (Luecht & Nungester 1998): (a) Aşağıdan Yukarıya (Bottom Up), (b) Yukarıdan Aşağıya (Top Down).

Aşağıdan Yukarıya Yöntemi. Bu yöntemde paneller arası paralellik, her modül için paralel formlar birleştirilerek sağlanır. Paralel modüller daha sonra karıştırılır ve çok sayıda paralel panel oluşturmak için eşleştirilir. Her modülün alternatif formları paralel olduğundan, ortaya çıkan panellerde birbirine karşılık gelen yollar da otomatik olarak paralel olacaktır. Aşağıdan yukarıya yaklaşımı, madde havuzunun paralel panel yapılarındaki hedefler ve kısıtlamalarla uyumlu bir şekilde karşılaması durumunda daha kolaydır.

Yukarıdan Aşağıya Yöntemi. Yukarıdan aşağıya yöntemi, sınav katılımcısının izlediği yola göre testlerin birbirine paralel olmasını amaçlar. Bu yöntemde öncelikle modüller paralel olacak şekilde birleştirilir. Ardından paneller arasında paralellik elde etmek ve istatistiksel olmayan kısıtlamaları karşılamak için panel düzeyinde ek bir optimizasyon

turu gerçekleştirilir. Bu yöntem, kısıtlamaların her modül için eşit olarak bölünemediği ve bu nedenle yalnızca yol düzeyinde belirlenebildiği kısa testler için faydalı olabilir, fakat bir aşamadaki modüllerin birbirine paralel olmasını garanti etmemektedir. Aşağıdan yukarıya yönteminde paralel modüller oluşturulabilse bile, yukarıdan aşağıya yöntemi, modüller arasında ileri düzey kısıtlamaların uygulanabilmesi (düşman maddeler, ortak köklü maddeler vb.) için test tasarımı üzerinde daha fazla kontrole izin vermektedir (Breithaupt & Hare 2007).

A-BÇAT için Test Birleştirme

A-BÇAT'ta test birleştirme Gölge Test yaklaşımı ile gerçekleştirilmektedir (Choi & van der Linden, 2018). Gölge Test yaklaşımı, uyarlanabilir bir testin uzunluğunda testin tüm özelliklerini ve kısıtlamalarını karşılayacak şekilde birleştirilen testlere dayanır. Gölge Test yaklaşımında birleştirilen testteki en yüksek düzeyde bilgi sunan madde sınav katılımcısına uygulanır, diğer maddeler ise katılımcıya gösterilmez. Bu nedenle gölge test adı verilmiştir. Bu bölümde gölge testinin temel yapısı ele alınmıştır.

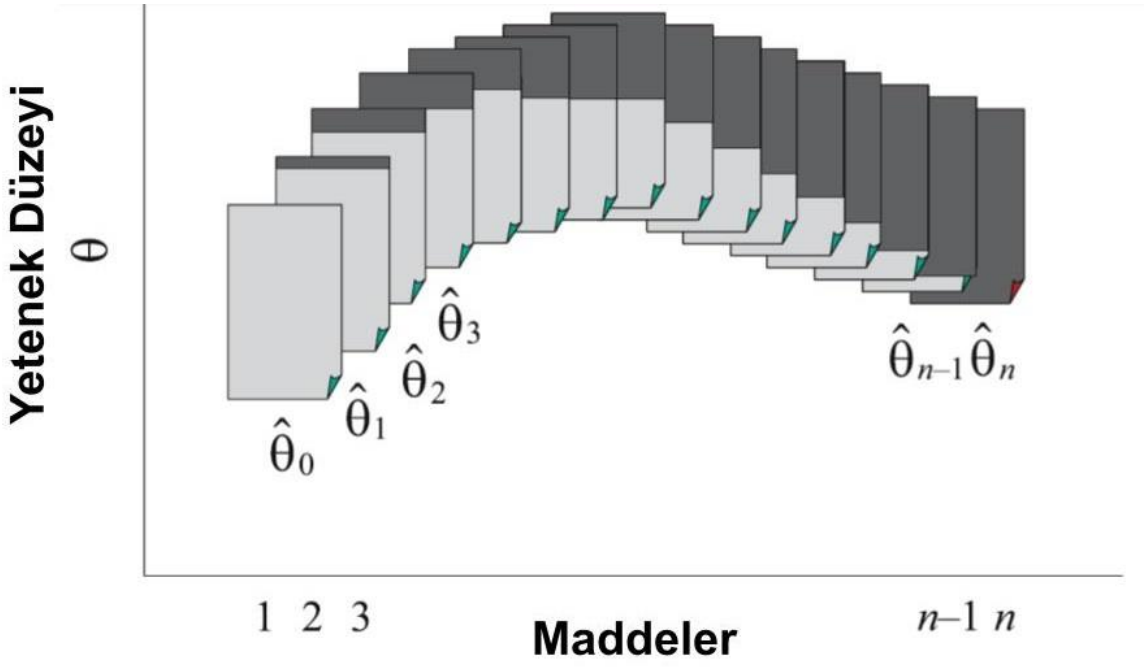
Gölge Test Yaklaşımı. Gölge test ile BBT uygulamalarında her yeni madde bireye sunulmadan önce gerçek zamanlı bir test birleştirmesi yapılır. Gölge testin üç temel özelliği vardır: a) Gölge test tüm test kısıtlamalarını karşılayacak şekilde birleştirilir. b) Sınav katılımcısına uygulanan maddeleri de içerir. c) Bireyin geçici yetenek düzeyinde maksimum bilgi sunan tam uzunlukta (n) bir testtir. Gölge test birleştirildiğinde sınav katılımcısına uygulanacak olan madde, en fazla bilgiye sahip olan maddedir.

Gölge testi yaklaşımının temel yapısı Şekil 9'da gösterilmektedir. Şekil 9 incelendiğinde grafiğin yatay ekseni, testteki maddelerin konumunu gösterir; dikey eksen, maddeler tarafından ölçülen yeteneği temsil eder. Gölge testlerinin dikey konumu ne kadar yüksekse, mevcut yetenek kestiriminin yüksek olduğunu ifade eder. Sonlara doğru ise gölge testlerin konumunun birbirine yakınlaşması, nihai yetenek kestiriminin istikrarlı bir hale geldiğini gösterir. Gölge testlerinin daha koyu kısmı, sınav katılımcısı tarafından yanıtlanan maddeleri temsil eder. Daha açık kısım ise, yetenek kestiriminin yeni bir güncellemesinden

sonra yeniden birleştirilen kısmını temsil eder. En son gölge testi, sınava giren tarafından fiilen alınan maddelerin tümünü içerir.

Şekil 9

Gölge Testin Temel Yapısı



Gölge testlerin çalışma mantığının algoritması aşağıda yedi adımda özetlenmiştir (van der Linden, 2009):

Adım 1: Sınavı başlatın.

Adım 2: Tüm kısıtlamaları karşılayan ve belirlenen başlangıç yetenek düzeyinde maksimum bilgi sunan bir gölge testi oluşturun.

Adım 3: Gölge testindeki maksimum bilgi veren maddeyi test katılımcısına uygulayın.

Adım 4: Geçici yetenek kestirimini gerçekleştirin.

Adım 5: Uygulanan maddeyi bir sonraki gölge testine dahil etmek için test birleştirme modelini güncelleyin.

Adım 6: Kullanılmayan tüm maddeleri madde havuzuna iade edin.

Adım 7: Sınav katılımcısına test uzunluğu (n) kadar madde uygulanana kadar Adım 2-6'yı tekrarlayın.

Gölge test yaklaşımının altında yatan fikir hem tüm test kısıtlamalarını karşılayan hem de sınav katılımcısının gerçek yeteneği hakkında maksimum bilgi sunan bir test oluşturmaktır. Ancak, gerçek yetenek her zaman bilinmediğinden, tek umulabilecek şey, bu ideale mümkün olduğunca yakın olan madde seçimidir. Gölge testi yaklaşımının şu özelliği vardır; sınav katılımcısının gerçek yetenek düzeyinde bilgi fonksiyonu açısından en uygun değere yaklaşan test oluşturmaktır. Algoritma, her gölge testi için tüm kısıtlamaları aynı anda karşılayacak test birleştirir. Sonraki her bir gölge testi, sınava giren kişiye halihazırda uygulanmış olan tüm maddeleri içerir. Bu nedenle, son gölge testi gerçek uyarlamalı testtir ve her zaman tüm kısıtlamaları karşılar. Ayrıca, her gölge testi, geçici yetenek düzeyinde bilgi fonksiyonu maksimum değere sahip olacak şekilde birleştirilir ve gölge testinden seçilerek sınav katılımcısına uygulanacak maddenin bu fonksiyona maksimum katkısı vardır. Bununla birlikte, test birleştirme işleminin hızlı ve başarılı bir şekilde gerçekleştirilmesi için iyi tasarlanmış bir madde havuzu gerekmektedir. Diğer taraftan, testin oluşturulmasında eklenen kısıtlamalar da test birleştirme işleminin hızını etkileyecektir. Dolayısıyla küçük bir madde havuzundan ciddi kısıtlamalar ile oluşturulan bir uyarlanabilir testin sadece birkaç kısıtlama içeren büyük bir madde havuzundan yapılan uyarlanabilir testten daha yavaş olması beklenir (van der Linden, 2009).

Dondur - Yenile Mekanizması. Orijinal gölge testi yaklaşımı, her θ yetenek güncellemesinde gölge testinin yeniden birleştirilmesine dayanmaktadır. Bununla birlikte, her maddeden sonra yeni test birleştirmesi yerine, testin önceden belirlenen konumlarında test birleştirme işleminin yapılabileceği fikri öne çıkmıştır. Choi vd. (2016), test birleştirmenin yalnızca belirli aralıklarla (örneğin her 4 maddede bir) yeniden birleştirilmesinin, her maddede yeniden test birleştirme ile neredeyse eşdeğer sonuçlar verebileceğini belirtmektedir. Özellikle iki geçici yetenek kestirimi arasındaki fark çok küçük olduğunda (örneğin $\theta_{k-1} - \theta_{k-2} < 0.1$) biraraya getirilen gölge testin önceki gölge testle aynı

olabileceği ifade edilmektedir. Aynı durum madde havuzunda geçici yetenek düzeyinde yüksek bilgi veren madde sayısı az olduğunda ya da test kısıtlamaları aşırı derecede olduğunda da gözlemlenebilir (Choi ve diğerleri, 2016). İlk olarak van der Linden ve Diao (2014) tarafından tanıtılan bu dondurma-yenileme mekanizması, test için tercih edilen uyarlama seviyelerini belirlemek için de kullanılabilir. Çeşitli uyarlama seviyelerine sahip genelleştirilmiş gölge testi derleme çerçevesinin birkaç örneği şunlardır:

– Anında Doğrusal Test (On-The-Fly Linear Test): Belirlenen yetenek düzeyinde bir sefer test birleştirilmesi yapılarak oluşturulan benzersiz testtir. Daha fazla yeniden birleştirme yapılmadan uygulanan tek bir gölge testi ifade eder (her sınava giren için oluşturulmuş benzersiz bir doğrusal sabit biçimli test),

- Anında Bireyselleştirilmiş Çok Aşamalı Test (On-The-Fly Multistage Testing) : Sınava giren kişinin geçici yetenek düzeylerine göre yalnızca önceden belirlenmiş madde konumlarında birleştirilen uyarlanabilir test,

– Hibrit Uyarlanabilir Testler (Hybrid Adaptive Testing): Doğrusal bir sabit formdan sonra tam uyarlanabilir bir gölge testin izlediği test (veya tersi, önce uyarlanabilir maddeler, ardından anında doğrusal test).

Son örnekte gösterildiği gibi, dondur - yenileme mekanizması, sabit maddelerin bir bölümünden sonra uyarlamalı testler yapmak için de kullanılabilir. Örneğin, uyarlanabilir bir testte ortak köke dayalı dört sorudan oluşan bir matematik sorusu sınav katılımcısına yöneltilebilir ve katılımcı bu dört soruyu yanıtlarken yeniden birleştirme uygulaması dondurulur, katılımcının sorular arasında serbestçe ileri ve geri hareket etmesine izin verilebilir. Ortak köke dayalı tüm sorular yanıtlandıktan sonra geçici yetenek düzeyi kestirilerek, bu yetenek düzeyine göre yeniden test birleştirilerek sınav uygulaması devam ettirilebilir.

Bu çalışmada farklı BÇAT modellerinin karşılaştırılması amaçlandığından dondur-yenile mekanizması ile yukarıda verilenlerden ikinci seçenek olan A-BÇAT'a

yoğunlaşılmıştır. A-BÇAT, sadece önceden belirlenmiş bazı noktalarda sınava giren kişinin yeteneğinin kestirildiği ve yetenek düzeyine uygun olacak şekilde yeniden birleştirilen bireysel bir gölge testi olarak ifade edilebilir (Choi & van der Linden, 2018). Dondur-yenile mekanizması kullanılarak A-BÇAT'ın nasıl oluşturulacağına dair 12 maddelik bir test örneği Şekil 10'da sunulmuştur.

Şekil 10

A-BÇAT ile Dondur Yenile Mekanizması

Madde Pozisyonu	1	2	3	4	5	6	7	8	9	10	11	12
2 Aşamalı A-BÇAT	1	0	0	0	0	0	1	0	0	0	0	0
3 Aşamalı A-BÇAT	1	0	0	0	1	0	0	0	1	0	0	0

Şekil 10'da test uzunluğu 12 maddeden oluşan iki farklı A-BÇAT tasarımının yeniden birleştirme mekanizması görülmektedir. Kırmızı kare içerisinde 1 yazan madde pozisyonlarında katılımcının geçici yetenek düzeyine göre yeniden birleştirme ile gölge test oluşturulmaktadır. 2 Aşamalı A-BÇAT tasarımında üç aşamanın her birinde 4 madde yer aldığı görülmektedir. Bu tasarımda sınav katılımcısına ilk aşamada θ = 0 düzeyinde 12 maddelik bir gölge test birleştirilir. Sınav katılımcısı 6. maddeyi yanıtladıktan sonra geçici yetenek düzeyi kestirilir ve kestirilen geçici yetenek düzeyinde yeniden gölge test birleştirilir. Katılımcı, yeniden birleştirilen gölge testteki 6 maddeyi de yanıtlar ve sınav nihai yeteneği kestirilerek sonlandırılır. Sonuç olarak bu tasarımda 1. ve 7. madde pozisyonlarında iki kez gölge test birleştirildiğinden bu tasarım 2 Aşamalı A-BÇAT olarak isimlendirilir. 3 Aşamalı A-BÇAT tasarımında ise 1, 5, ve 9. madde pozisyonlarında yeniden birleştirme yapılarak gölge test yaklaşımı ile 3 Aşamalı A-BÇAT uygulaması gerçekleştirilmiş olur.

Geniş Ölçekli Değerlendirmeler [International Large-Scale Assessment - ILSA]

Öğrenci becerilerini ölçmeye yönelik günümüz uluslararası değerlendirmelerinin kökeni 1960'ların başında Uluslararası Eğitim Başarılarını Değerlendirme Kuruluşu

(International Association for the Evaluation of Educational Achievement - IEA) tarafından yürütülen Birinci Uluslararası Matematik Çalışması'na (First International Mathematics Study - FIMS) dayanmaktadır. FIMS'ın iki temel amacı, çoktan seçmeli maddelerden zihinsel fonksiyonlar hakkında çıkarımlarda bulunmak ve geniş ölçekli bir uluslararası çalışmanın uygulanabilirliğini test etmekti (Purves, 1987). O dönemde kısıtlı bir bütçe ve çok az modern teknolojik olanaklarla FIMS gibi uluslararası bir projenin gerçekleştirilmesi ILSA'ların gelecek yüzyıl için önemini ve vizyonunu ortaya koymaktadır. ILSA'lar zaman içerisinde modern eğitim araştırmalarında, eğitim politikaları tartışmalarında ve kamusal söylemlerde önemli konuma yükselmiştir.

Günümüz eğitim araştırmalarında ve politika tartışmalarında PISA (Programme for International Student Assessment), TIMSS (Trends in International Mathematics and Science Study), PIRLS (Progress in International Reading Literacy Study), PIAAC (Programme for the International Assessment of Adult Competencies) gibi modern ILSA'lar yıllar içerisinde gittikçe önemini arttırmıştır. Özellikle katılımcı ülkelerin performanslarını karşılaştıran başarı sıralamaları basında, gazetelerde, sosyal medyada ve politikacıların ülkenin eğitim sistemi hakkında söylemlerinde yoğun bir şekilde yer almaktadır. ILSA'lar, ülkeler arası başarı sıralamalarının haricinde, eğitimde hesap verebilirlik, adillik (fairness) ve eğitim sisteminin izlenmesi konularına da vurgu yapmaktadır. Ayrıca, dünya çapında bir dizi alanda veya bir öğretim düzeyinde öğrenme ve başarı arasındaki ilişkilerle ilgilenen araştırmacılar ve politika yapıcılar için ILSA'lar başlıca kaynaklar haline gelmektedir.

ILSA'lar, okuma, fen ve matematik gibi bir dizi alandaki kesitsel başarı kestirimlerinin yanında, genellikle bu çalışmaların politika tartışmalarında yer alan katılımcı öğrencilerden, evlerinden, öğretmenlerinden, velilerinden ve okullarından elde edilen çok sayıda yardımcı arka plan bilgileri, öğrenci geçmişi anketleri, öğrenmeye yönelik tutumları, ev ortamı kaynakları, çalışma ve boş zaman alışkanlıkları ve okul zorbalığı gibi pek çok bileşeni içerir. Ayrıca, öğretmenlerden pedogojik uygulamalara, okul iklimine, çalışma koşullarında dair bilgiler yer alabilmektedir. Benzer şekilde okul yöneticilerinden eğitim sürecine ve okulun

imkanlarına dair bilgiler toplanabilmektedir. ILSA'ların veri tabanında yer alan bu bilgiler uluslararası düzeyde öğrenmenin bağlamı ve ilişkileri ile ilgilenen araştırmacılara mükemmel bir kaynak sunmaktadır.

Bu bölümde ILSA'lardan bu çalışma kapsamında madde parametrelerinin üretilmesi ve sınav kapsamının oluşturulmasında faydalanılan TIMSS uygulamasına dair bilgiler yer almaktadır.

Uluslararası Matematik ve Fen Eğilimleri Araştırması (TIMSS - Trends In Mathematics And Science Study)

Uluslararası Matematik ve Fen Eğilimleri Araştırması (TIMSS), dünyanın pek çok ülkesinden öğrencilerin matematik ve fen başarıları hakkında uluslararası ve güvenilir ölçümler yapmayı hedefleyen bir değerlendirme çerçevesidir. TIMSS, 1995 yılından bu yana, genellikle her 4 yılda bir, 4. ve 8. sınıf öğrencilerinden toplamaktadır. Matematik ve fen bilimleri değerlendirmelerine ek olarak, öğrencilere, öğretmenlerine ve okul yöneticilerine öğrenme için arka plan bağlamları hakkında bilgi toplamak amacıyla anketler uygulanmaktadır. TIMSS, IEA tarafından desteklenmekte ve yürütülmektedir. Türkiye TIMSS uygulamasında 1999 yılından itibaren katılmaktadır. TIMSS 2019 uygulamasında döndüncü sınıflarda 58, sekizinci sınıflarda 38 farklı ülkeden olmak üzere toplamda yaklaşık 580 bin öğrencinin katılımı ile gerçekleştirilmiştir.

TIMSS Değerlendirme Çerçevesi. TIMSS uygulamasında öğrencilerin matematik ve fen alanlarındaki becerilerini değerlendirmek için değerlendirme çerçevesi kitapçıklar (booklet) halinde oluşturulmuştur. Soruların yer aldığı değerlendirme havuzundaki maddeler, daha önceki döngülerde kullanılan maddeler ile güncel döngü için geliştirilen maddelerden oluşur. Değerlendirme havuzunda yer alan maddeler ile içerisinde 10-18 aralığında soru yer alan bloklar oluşturulur. Kitapçıklar ise içerisinde iki matematik ve iki fen olmak üzere dört bloktan oluşturulur. Değerlendirme havuzunda yer alan sorular ile her öğrencinin bir kitapçığı yanıtlamasını gerektiren 14 öğrenci başarı kitapçığı (student achievement booklet) oluşturulur. Değerlendirme havuzundan oluşturulan her blok, iki

kitapçıkta yer alır ve tüm kitapçıklardan elde edilen veriler bir araya getirildiğinde hem çeşitli kitapçıklardan alınan yanıtları hem de önceki döngülerden elde edilen yanıtları birbirine bağlamak için bir mekanizma sağlanmış olur. Oluşturulan 14 kitapçık, yazılım tarafından önceden belirlenen atamalara göre sınıfta yer alan katılımcı öğrencilere sunulur. Bloklardan oluşan kitapçık modelinin daha iyi anlaşılması için Tablo 4'te yer alan TIMSS 2019 Kitapçık Deseni incelenebilir.

Tablo 4

TIMSS 2019 Kitapçık Deseni

Öğrenci Başarı Kitapçıkları	Değerlendirme Blokları			
	Matematik Blokları		Fen Blokları	
Kitapçık 1	MP01/ME01	MP02/ME02	SP01/SE01	SP02/SE02
Kitapçık 2	SP02/SE02	SP03/SE02	MP02/ME02	MP03/ME03
Kitapçık 3	MP03/ME03	MP04/ME04	SP03/SE03	SP04/SE04
Kitapçık 4	SP04/SE04	SP05/SE05	MP04/ME04	MP05/ME05
Kitapçık 5	MP05/ME05	MP06/ME06	SP05/SE05	SP06/SE06
Kitapçık 6	SP06/SE07	SP07/SE07	MP06/ME06	MP07/ME07
Kitapçık 7	MP07/ME07	MP08/ME08	SP07/SE07	SP08/SE08
Kitapçık 8	SP08/SE08	SP09/SE09	MP08/ME08	MP09/ME09
Kitapçık 9	MP09/ME09	MP10/ME10	SP09/SE09	SP10/SE10
Kitapçık 10	SP10/SE10	SP11/SE11	MP10/ME10	MP11/ME11
Kitapçık 11	MP11/ME11	MP12/ME12	SP11/SE11	SP12/SE12
Kitapçık 12	SP12/SE12	SP13/SE13	MP12/ME12	MP13/ME13
Kitapçık 13	MP13/ME13	MP14/ME14	SP13/SE13	SP14/SE14
Kitapçık 14	SP14/SE14	SP01/SE14	MP14/ME14	MP01/ME01

Tablo 4'te görüldüğü üzere bir kitapçıkta iki matematik ve iki fen olmak üzere dört blok yer almaktadır. Aynı zamanda her bloğun farklı iki kitapçıkta yer aldığı görülmektedir.

TIMSS 2019 uygulamasında 14 matematik ve 14 fen olmak üzere 28 blok kullanıldığı ve bu 28 blok ile 14 farklı kitapçık (booklet) oluşturulduğu belirtilebilir. Daha fazla ayrıntılı bilgi için Mullis ve Martin (2017) incelenebilir.

Örnekleme Deseni. TIMSS uygulamasında ulusal örnekleme tasarımı, iki aşamalı örnekleme tasarımına (ilk aşama olarak okullar ve ikinci aşama olarak okullardaki sınıflar) göre yapılmaktadır. İlk aşamada evrendeki uygun öğrencileri içeren tüm okullardan oluşan listeden büyüklükleriyle orantılı olasılıkla okullar örneklenir. Bu aşamada her okul için iki yedek okul da belirlenir. Bu aşamalar IEA'nın yardımı ile ülkenin Ulusal Araştırma Koordinatörü tarafından belirlenen örnekleme çerçevesine göre gerçekleştirilir. Ardından ikinci aşamada belirlenen okullardan sınıf örneklemesi gerçekleştirilir. Bu aşamada belirlenen okullardaki bir (veya daha fazla) sınıfta yer alan tüm öğrenciler örnekleme dahil edilir. Bu sınıfların belirlenmesinde ise okullardan sınıfların seçimi Ulusal Araştırma Koordinatörü'nün desteği ile IEA'nın geliştirdiği yazılım tarafından gerçekleştirilir. Bu örnekleme aşamalarının gerçekleştirilmesinde ülkelerin ulusal farklılıkları göz önüne alınarak bölge, kıta, ada, nufüs gibi değişkenlere göre tabakalandırmalar yapılabilmektedir (LaRoche, Joncas & Foy, 2020).

Ölçekleme Yöntemi (Scaling Method). TIMSS uygulamalarında katılımcı öğrencilerin maddelere verdikleri yanıtları ve maddeleri değerlendirmek için Madde Tepki Kuramı (MTK) kullanılmaktadır. TIMSS 1995 yılında Rasch modelini kullanmıştır ve 1999 yılından bu yana çoktan seçmeli maddeler için üç parametreli lojistik (3PL) modeli, 0-1 şeklinde puanlanan açık uçlu yanıtı yapılandırılmış maddeler için iki parametreli lojistik (2PL) modeli ve 0-1-2 şeklinde puanlanan açık uçlu yanıtı yapılandırılmış maddeler için genelleştirilmiş kısmi puan modelini (GPCM) (Muraki, 1992) kullanmaktadır (von Davier, 2020). Ayrıca her TIMSS döngüsünde önceki döngülerde kullanılan maddelerin ortak maddeler olduğu göz önüne alınarak madde parametreleri Stocking-Lord yöntemi ile önceki yıllara göre ölçeklenmektedir.

Olası değerler (Plausible Values – PV). Geniş ölçekli değerlendirmelerde katılımcı öğrencilere sunulan sorular değerlendirme havuzunda yer alan maddelerin belirli bir kısmını oluşturmaktadır. Dolayısıyla katılımcı öğrenciler soruların yalnızca bir kısımını yanıtlamaktadır. Aynı zamanda katılımcı öğrenci grubu değerlendirme havuzunda yer alan tüm maddeleri yanıtlamış olmaktadır. Her bir katılımcı öğrencinin yanıtlamadığı soruların olması sebebiyle yetenek düzeyi kestirimi yerine eksik yanıt matrisinden öğrenci başarısını ifade eden olası değerler raporlanmaktadır. Olası değerlerin kestiriminde öğrenci başarısı kayıp değer olarak belirlenir ve çoklu atama (multiple imputation) yöntemi ile bu değer kestirilir. Aynı zamanda ölçme hatalarını azaltmak amacıyla her katılımcı öğrenci için birden fazla olası değer kestirilmektedir. TIMSS uygulamasında her öğrenci için 5 farklı olası değer kestirilir. Geniş ölçekli değerlendirmelerden elde edilen verilerle yapılacak analizlerde bu olası değerlerin herhangi birini veya ortalamasını kullanmak yerine tüm olası değerlerin kullanılması gerektiği belirtilmektedir (Laukaityte & Wiberg, 2017; Arıkan ve diğerleri, 2020).

Olası değerler yöntemi öğrenci başarısını kayıp değer olarak kabul eder (Rutkowski ve diğerleri, 2010). Öğrenci başarı dağılımları Rubin'in (1987) çoklu atama (multiple imputation) yöntemi ile kestirilir. Bu dağılımların içinden rastgele seçimler yapılır ve atanan çoklu verilere olası değerler denir (Rutkowski ve diğerleri, 2010). Olası değerler gözlemlenemeyen gizil değişkenler için emsal değerlerdir (Wu, 2005). Her öğrencinin gözlemlenemeyen gizil başarı değişkeni vardır ve bu değişkene ait çoklu veriler atanmaktadır (Laukaityte & Wiberg, 2017; Wu, 2005). OECD (2017) olası değerleri, puan dağılımlarından bireylere atamak için rastgele seçilen sayılar olarak tanımlamaktadır. Bu dağılıma ise marjinal sonsal dağılım (marginal posterior distribution) denmektedir. Seçkisiz hata varyans bileşenleri içeren olası değerler test puanları olarak ele alınmadan, evrenin performansını tanımlamak amaçlı kullanılmaktadır (OECD, 2017). Kısacası ölçüm hatalarını azaltmak için her kişiye birden fazla değer atanmaktadır (Laukaityte & Wiberg, 2017). Eğer ölçüm hatası küçükse, kişi için atanan çoklu değerler birbirine yakın; ölçüm hatası büyük ise de atanan değerler birbirine uzak olmaktadır (Wu, 2005). Atanan olası

değerler sayesinde geniş ölçekli testlerdeki çıkarımlar daha geçerli hale gelmekte ve sonuçlarının pratiğe katkısı daha verimli olmaktadır (Laukaityte & Wiberg, 2017). PISA ve TIMSS gibi bir çok geniş ölçekli testlerin verilerinde 5 olası değer kullanmaktadır (Laukaityte & Wiberg, 2017, OECD, 2017). PISA, 2015 uygulamasından beri 10 tane olası değer raporlamaya başlamıştır. PIAAC içerisinde de 10 tane olası değer raporlanmaktadır. National Assessment of Educational Assessment (NAEP) veritabanı ise 20 tane olası değer kullanmaktadır. Laukaityte ve Wiberg (2017)'in yaptıkları simülasyon çalışmaları birden fazla olası değerin kullanılmasının yapılan tahminin ve ölçüm hatasının doğruluğunu arttırdığını göstermiştir. PISA, PIAAC ve TIMSS gibi geniş ölçekli testlerin verilerinde sonuçlar olası değerler olarak raporlandığı için yapılacak tüm analizlerde bu olası değerleri dikkate alan yöntem ve yazılımların kullanılması zorunludur. Olası değerleri dikkate almadan elde edilen sonuçların hatalı olacağının farkında olunması gerekmektedir (LaRoche & Foy, 2015; OECD, 2017, Rutkowski ve diğerleri, 2010).

TIMSS 2023 ve Yeni Yönelimler. TIMSS, her değerlendirme döngüsünde yeniliklerin gelişimini desteklemektedir. TIMSS 2019'da ülkelerin neredeyse yarısı bilgisayar tabanlı dijital değerlendirme formatını seçmiştir ve kağıt-kalem formatını seçen diğer yarısı da dijital değerlendirme formatına geçişi başlatmıştır (Yin & Foy, 2021). TIMSS 2023 için ülkelerin büyük çoğunluğu dijital değerlendirmeye geçmiştir veya geçmektedir. Ayrıca TIMSS 2023 uygulamasında öğrenci yeteneğinin daha iyi kestirilmesi için daha geniş bir değerlendirme güçlüğünü sağlayacak şekilde yeni bir grup uyarlamalı değerlendirme tasarımını (group adaptive assessment design) benimsemektedir. Aynı zamanda bu yeni tasarım PIRLS 2021'de de tanıtılmaktadır.

Yeni grup uyarlamalı değerlendirme tasarımında maddelerden oluşan bloklar kolay, orta ve zor (hard, medium, easy) olmak üzere üçe ayrılmıştır. 14 bloklu ve 14 kitapçıklı geleneksel tasarım yerini korurken, kitapçıkların oluşturulması ve isimlendirilmesinde bazı farklılıklar bulunmaktadır. Kitapçık oluşturulurken içerisinde yer alan bloklara göre isimlendirilmektedir. Kitapçığı oluşturan iki blok zor-zor veya orta-zor güçlükte bloklar ise

kitapçık daha zor kitapçık (more difficult booklet) olarak adlandırılmaktadır. Eğer ki iki blok, kolay-orta veya kolay-kolay güçlükte bloklardan oluşuyorsa kitapçık daha kolay kitapçık (less difficult booklet) olarak adlandırılmaktadır. Katılımcı öğrencilerin bu kitapçıklardan hangisini alacağı, ülkesinin önceki döngülerdeki başarı durumuna göre belirlenecektir. Daha fazla bilgi için Yin ve Foy (2021) incelenebilir.

Görüldüğü üzere, TIMSS 2023 uygulamasında grup uyarlamalı değerlendirme tasarımına geçiş yapılmıştır. Uygulanacak olan bu yaklaşım tam manada bireyselleştirilmiş bir test yaklaşımı mantığı içermese de BBT ve BÇAT'ın temel mantığına benzer şekilde katılımcılara uygulanan maddelerin katılımcının yetenek düzeyine göre belirlenmesi açısından benzerlik taşıdığı ifade edilebilir.

İlgili Araştırmalar

Bu bölümde literatürde yer alan ve S-BÇAT ve A-BÇAT üzerine ilgili araştırmalara ve sonuçlarına yer verilmiştir.

Patsula (1999), kağıt-kalem testi, BBT ve S-BÇAT'ı farklı simülasyon koşullarında karşılaştırdığı doktora tezinde, yetenek kestiriminin kesinliğinin en yüksek BBT'de olduğu ve ardından sırasıyla S-BÇAT ve kağıt kalem testi geldiği sonucuna ulaşmıştır. Üç aşamalı BÇAT deseninde ilk aşamalarda madde sayısının az olmasının yetenek kestiriminin doğruluğunu arttırdığı bulgusu yer almaktadır. İki aşamalı S-BÇAT deseninden üç aşamalı S-BÇAT desenine geçildiğinde yetenek kestiriminin doğruluğunun arttığı ifade edilmiştir.

Schnipke ve Reese (1999), çeşitli S-BÇAT tasarımlarının performanslarını BBT ve doğrusal test tasarımları ile karşılaştırdı. Hukuk Okulu Kabul Testinden (LSAT) alınan maddelere dayalı olarak gerçekleştirilen çalışmada, yetenek kestiriminin doğruluğu en yüksek BBT'de iken onu sırasıyla S-BÇAT ve doğrusal test izlemiştir. Araştırmacılar, S-BÇAT'ın doğrusal teste göre testi kısalttığı ve daha etkili kestirim yaptığı sonucuna ulaşmışlardır.

Xing ve Hambleton (2004), madde bankası boyutu ve madde kalitesinin S-BÇAT tasarımı ve kağıt-kalem testi tasarımı üzerindeki etkisini araştırdığı çalışmasında, daha düşük kaliteli maddelere sahip büyük bir madde havuzunun, daha kaliteli maddelere sahip daha küçük bir madde havuzundan daha kötü sonuçlar verebileceğini göstermektedir. Bu sonuç, bir S-BÇAT tasarımı oluştururken madde havuzunun madde kalitesi ve boyutunun önemli hususlar olduğunu göstermiştir.

Zenisky (2004), farklı BÇAT tasarımlarının performansını araştırmıştır. Toplam test bilgisi azaldıkça yetenek parametreleri arasındaki korelasyonun azaldığı bulgusuna erişmiştir. Benzer şekilde test bilgisi azaldıkça RMSE değerleri artış göstermiştir. Yönlendirme yöntemi için, doğru sayısı puanlama yöntemi, en yüksek karar doğruluğunu sağladı, ardından biraz daha düşük olan tanımlanmış popülasyon aralıkları yönlendirme yöntemi geliyordu. Son olarak ise rastgele yönlendirme yöntemi biraz daha kötüydü.

Jodoin, Zenisky ve Hambleton (2006), sabit uzunluklu ve BÇAT tasarımını karşılaştırdığı çalışmalarında 60 maddelik test uzunluğunun 40 maddelik test uzunluğuna göre daha iyi sonuç verdiğini, fakat aradaki farkın kabul edilebilir düzeyde olduğunu belirtmektedir. Ayrıca sabit uzunluklu ve S-BÇAT ile yapılan sınıflama doğruluğunun üst düzeyde olduğu belirtilmektedir. İki aşamalı S-BÇAT tasarımı, sabit uzunluklu ve üç aşamalı S-BÇAT tasarımından biraz daha kötü sonuçlar vermiş olsa da, maliyet ve test süresini azaltması açısından iki aşamalı testlerin övgüye değer olduğu belirtilmiştir.

Keng (2008), Madde Takımı Tepki Modeli (Testlet Response Theory - MTTM) ile BBT ve BÇAT tasarımlarının etkililiğini ele almıştır. MTTM temelinde gerçekleştirilen BBT ve S-BÇAT simülasyonları sonucunda, BBT'nin daha iyi ölçme kesinliği sunduğu görülmüştür. Çarpık yetenek dağılımlarında ise S-BÇAT'ın ölçme kesinliğinin daha fazla azaldığı sonucuna ulaşılmıştır.

Kim, Chung, Dodd ve Park (2012), karma formatlı gerçek veriden elde edilmiş madde havuzu ile BÇAT ile farklı BÇAT desenlerini karşılaştırmıştır. İlk aşamada yer alan

modülün test bilgi düzeyi arttığında S-BÇAT'ın doğru sınıflandırma oranının arttığı görülmüştür.

Andrew (2014), S-BÇAT'ta yönlendirme (routing) ve puanlama yöntemlerinin etkilerini araştırdığı çalışmasında, 1-3, 1-2-3, 1-2-3-4 S-BÇAT tasarımlarını ele almıştır. Araştırma sonuçlarına göre, Madde Tepki Kuramı'ndan elde edilen sonuçların geleneksel yöntemlere göre daha doğru yönlendirme yaptığı sonucuna ulaşılmıştır.

Hembry (2014), 3 Parametreli Madde Takımı Tepki Modeli (Testlet Response Theory) ile karma formatlı madde takımların dayalı madde havuzu üzerinde S-BÇAT'ın işlevselliğini incelemiştir. S-BÇAT'ın, 3PL Madde Takımı Modeli ile yeterli ölçme kesinliği sunduğu görülmüştür. Yönlendirme yöntemlerinin etkisinin düşük olduğu belirtilmiş olup, çeşitli koşullar arasında ciddi bir yanlılık görülmediği ifade edilmektedir.

Park, Kim, Chung ve Dodd (2014), genelleştirilmiş kısmi puan modeli (GPCM) ile karma formatlı madde havuzu kullanılarak S-BÇAT tasarımlarının etkililiğini karşılaştırmıştır. Araştırmacılar farklı BÇAT tasarımlarının iyi düzeyde sınıflandırma kesinliği sunduğunu belirtmektedirler.

Choi, Moellering, Li ve van der Linden (2016), test birleştirmede dondur-yenile mekanizmasını kullanarak BBT, A-BÇAT ve bu iki yöntemin birleşimi ile oluşan hibrit yöntemi karşılaştırmıştır. Sonuçlar her üç yaklaşım için oldukça benzerdir. Araştırmacılar, özellikle ortak köklü maddeler olduğunda ve test kısıtlamaların karşılanması gerektiğinde dondur-yenile mekanizmasının oldukça başarılı şekilde çalıştığı ve testin ölçüm kesinliğinde ciddi bir düşüşün olmadığı sonucuna ulaşmışlardır.

Sarı ve Huggins-Manley (2017), farklı bireyselleştirilmiş testlerde içerik alanları sayısı ile test uzunluğunun farklılaştığı durumlarda, BBT ile S-BÇAT modellerini karşılaştırmıştır. Araştırma sonuçlarına göre BBT, S-BÇAT'a göre içerik alanı sayısı değişiminden daha fazla etkilenmiştir.

Zheng, Nozawa, Zhu ve Gao (2016), S-BÇAT'ta panellerin oluşturulmasında kullanılan aşağıdan-yukarıya (top-down) test birleştirme yöntemini incelemiştir. 1-2-4 ve 1-2-3-4 tasarımlarının ele alındığı çalışmada, aşağıdan-yukarıya yöntemi ile oluşturulan panellerin bilgi eğrilerinin benzer olduğu ve S-BÇAT'ın doğrusal testlere göre avantajlı olduğu sonucuna ulaşılmıştır.

Zheng ve Chang (2015), S-BÇAT'ın her aşamasında sınava girenlerin geçici yetenek kestirimine dayalı olarak anında modül birleştiren ve testten önce birleştirilmemiş yenilikçi bir BÇAT tasarımını tanıtmaktadır. Bu çalışmanın sonuçları, A-BÇAT ile BBT'nin ölçme kesinliği sonuçlarının karşılaştırılabilir olduğunu ve bu iki yöntemin S-BÇAT'tan daha iyi ölçme kesinliği ve test güvenliği sunduğunu belirtmektedir.

Tay (2015), sınıflama doğruluğu ve sınıflama tutarlılığı açısından S-BÇAT ve A-BÇAT'ı 12,18,24 ve 30 test uzunluklarında karşılaştırmıştır. Tüm koşullarda A-BÇAT'ın S-BÇAT'a göre daha yüksek düzeyde sınıflama doğruluğu ve sınıflama tutarlılığı sonuçları ürettiği görülmektedir.

van der Linden ve Diao (2016), kalibre edilmiş gerçek veri seti ile BBT, hibrit BBT, A-BÇAT, S-BÇAT ve DT olmak üzere beş farklı test yaklaşımını simülasyon ile karşılaştırmışlardır. Bu çalışmanın sonuçları DT'nin en az verimli olduğunu, DT'nin ardından S-BÇAT'ın geldiğini göstermektedir. Diğer yaklaşımlar olan BBT, hibrit BBT ve A-BÇAT'ın ise yaklaşık olarak eşit verimlilikte çalıştığı belirtilmektedir.

Han ve Guo (2016), önceden birleştirilmiş test modüllerini içermeyen ve her aşamada anında yeni bir modülü birleştiren bir A-BÇAT tasarımı önermektedir. Bu yöntemde, sınav katılımcısının kestirilen bir θ yetenek düzeyine göre anında yeni bir modül birleştirilir. Birleştirilen modül, test geliştiricisi tarafından belirlenen test bilgi fonksiyonunda (TIF) yer alan bilgi miktarına ulaşana kadar iteratif olarak yeniden birleştirilir. Ardından birleştirilen modül, sınav katılımcısına sunulur ve bu işlemler kullanılan modül sayısı sonlanıncaya kadar devam eder. Araştırmacılar, BBT, S-BÇAT ve A-BÇAT tasarımlarını karşılaştırdılar. 1-3-3 BÇAT tasarımı, yeni geliştirilen A-BÇAT tasarımı ile düşük iterasyonlu

şekillendirmelerde benzer ölçüm kesinliği sonuçları verirken, iterasyon sayısı 100'e çıktığında yeni A-BÇAT tasarımı S-BÇAT'tan daha iyi ölçüm kesinliği sonuçları üretmiştir. BBT ise her iki yöntemden de daha iyi ölçüm kesinliği üretmektedir.

Choi ve van der Linden (2018), gölge testi yaklaşımını kullanarak oluşturdukları A-BÇAT ile BBT ve doğrusal test tasarımlarını karşılaştırmıştır. A-BÇAT'tan elde edilen ölçüm kesinliği BBT'den çok az miktarda düşük olsa da doğrusal testten daha iyi olduğu sonucuna ulaşılmıştır.

Tian (2018), BBT ve A-BÇAT'ı dört farklı durdurma kuralı ile karşılaştırdığı tezinde BBT ve A-BÇAT'ın benzer sonuçlar ürettiği ve BBT için geliştirilen durdurma kurallarının A-BÇAT için de kullanılabileceği sonucuna ulaşmıştır.

van der Linden (2021), gerçek veri setinden elde edilmiş 300 maddelik bir madde havuzu ile anında DT, S-BÇAT, A-BÇAT ve BBT'yi karşılaştırmıştır. Simülasyon çalışmasında test uzunluğu 30 madde ve her yetenek düzeyinde 250 katılımcı olacak şekilde $\theta = -2.0, -1.5, \ldots, 2.0$ noktalarında üretilmiştir. Anında DT için $\theta = -1.5,\ 0,\ 1.5$ düzeyleri boyunca maksimuma ulaşacak şekilde test birleştirmesi gerçekleştirildi. S-BÇAT ve A-BÇAT için 1-3-3 tasarımı altında test birleştirmeleri yapıldı. Sonuçlar, anında DT'nin en kötü sonuçları ürettiğini göstermektedir. A-BÇAT ile BBT'nin birbirine benzer ve gayet iyi ölçüm kesinliğine ulaştığı çalışmada, S-BÇAT her ne kadar anında DT yaklaşımından iyi olsa da, BBT ve A-BÇAT'tan belirgin bir şekilde daha düşük ölçüm kesinliği sonuçları ürettiği görülmüştür.

Bölüm 3

Yöntem

Bu bölümde araştırmanın türü, verilerin elde edilmesi, araştırma deseni ve verilerin analizi başlıkları ele alınmıştır.

Araştırmanın Türü

Araştırmada, farklı BÇAT yaklaşımlarının çeşitli benzetim koşulları altında etkililiğinin incelenmesi amaçlanmıştır. Araştırmada yer alan veriler simülasyon yöntemi ile üretilmiştir, farklı test desenleri ve farklı koşullar altında karşılaştırmalar yapılmıştır. Simülasyon çalışmaları, olasılık dağılımları üzerinden rastgele örnekleme yapılarak veri üretilmesini ve üretilen verilerin analiz edilmesini içeren bilgisayar deneyleridir. Bu deneylerde belirli senaryolar üzerinde istatistiksel yöntemlerin performansları hakkında sonuçlar elde etmek için kullanılmaktadır. Psikometride simülasyon çalışmaları, yeni yöntemlerin değerlendirilmesinde ve alternatif yöntemlerin değerlendirilmesinde önemli bir yer tutmaktadır (Feinberg & Rubright, 2016; Morris, White & Crowther, 2019). Bu çalışma, verilerin ilgili olasılık dağılımları ve simülasyon koşulları üzerinden simüle edildiği bir Monte Carlo simülasyon çalışmasıdır. Monte Carlo çalışmaları, pek çok yönüyle deneysel çalışmaları yansıtmakta olup veriler bilgisayar aracılığı ile üretilmektedir (Harwell, Stone, Hsu & Kirisci, 1996). Çalışmada ele alınan tüm koşulların gerçek veri üzerinde aynı anda ele alınmasının çok zor olması nedeniyle simülatif veri tercih edilmiştir. Bu araştırma, farklı BÇAT yöntemlerini karşılaştırmayı amaçlamakta olup hangi yöntemin daha uygun sonuçlar verdiğini ortaya koyacağından betimsel araştırma olduğu söylenebilir (Büyüköztürk, Kılıç-Çakmak, Akgün, Karadeniz & Demirel, 2013; Fraenkel, Wallen & Hyun, 2012).

Verilerin Elde Edilmesi

Verilerin üretilmesinde R programlama dilinden faydalanılmıştır. R, açık kaynak kodlu ve ücretsiz bir istatistik programlama dilidir. Verilerin üretilmesinde öncelikle 400 maddeden oluşan madde havuzu parametreleri ve 1000 bireyden oluşan yetenek

parametreleri üretilmiştir. Test birleştirme için “Rmst” (Luo & Kim, 2018) ve “TestDesign” (Choi, Lim & van der Linden, 2021) paketleri kullanılmıştır. Ardından S-BÇAT analizleri birleştirilen testler ile “mstR” (Magis, Yan & Von Davier, 2017) paketi, A-BÇAT analizleri ise “TestDesign” (Choi, Lim & van der Linden, 2021) paketi ile aynı madde havuzu ve yetenek parametreleri kullanılarak gerçekleştirilmiştir.

Madde Havuzunun Üretilmesi

Araştırma kapsamında 3 parametreli lojistik model (3PL) esas alınarak 400 maddeden oluşan madde havuzu üretilmiştir. Madde havuzunun üretilmesinde TIMSS 8. sınıf düzeyinde matematik dersi için 2003, 2007, 2011, 2015 ve 2019 uygulamalarında kullanılan ve 3PL’ye dayalı parametreleri kestirilen maddelerin a, b ve c parametrelerinin dağılımları incelenmiştir. Bu maddelerin parametrelerinin oluşturduğu dağılımlar dikkate alınarak; a parametreleri log-normal dağılımdan $a \sim lnN(0.2, 0.3)$, b parametreleri normal dağılımdan $b \sim N(0, 0.7)$, c parametreleri beta dağılımından $c \sim Beta(5, 16)$ elde edilmiştir. Lognormal dağılım, Log-normal dağılım, logaritması normal dağılım sergileyen bir rastgele değişkenin tek-kuyruklu bir olasılık dağılımıdır. Log-normal dağılımın iki temel girdi parametresi bulunur. Bu parametreler log-ortalama (ln(x)’in ortalaması) ve log-standart sapma (ln(x)’in standart sapması) parametreleridir. Normal dağılım, Gauss dağılımı veya çan eğrisi olarak da adlandırılan dağılımdır. Psikolojik veya fiziksel ölçümlerin daha iyi açıklanmasından dolayı normal dağılım pek çok istatistiksel modelin temellerinde kendine yer edinmiştir. Normal dağılımın iki temel girdi parametresi bulunur, bunlar ortalama ve standart sapma değerleridir. Beta dağılımı ise [0,1] aralığında iki pozitif şekil parametresi (genellikle α ve β olarak bilinir) ile normalize edilmiş bir sürekli olasılık dağılımıdır. Beta dağılımının α ve β parametreleri ile dağılımın çarpıklığı, basıklığı ve ranjı ayarlanır. Madde havuzunda yer alan 400 maddenin parametrelerinin betimsel istatistikleri Tablo 5’te Tablo 5’te sunulmuştur.

Tablo 5

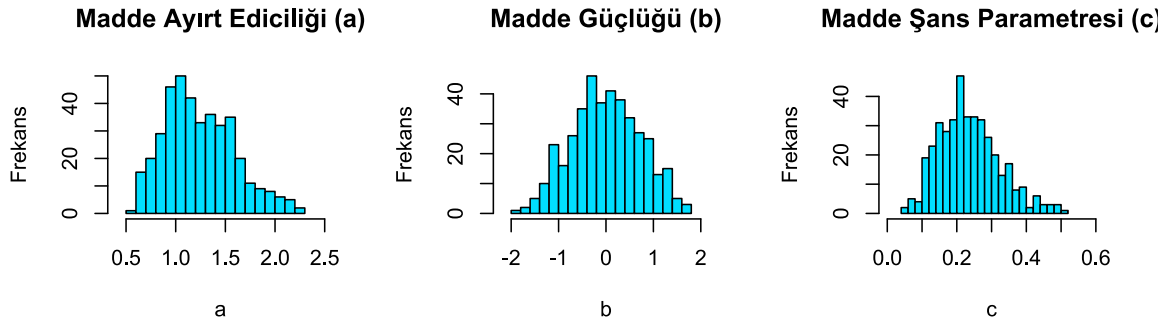
Madde Parametrelerinin Betimsel İstatistikleri

Parametre	K	Minimum	Maksimum	Ortalama	Standart Sapma
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

Üretilen madde parametrelerinin dağılımları Şekil 11'de yer alan histogram grafiği ile gösterilmektedir.

Şekil 11

Madde Parametrelerinin Histogram Grafiği



Ayrıca madde havuzunun dört farklı kapsamdan oluştuğu varsayılarak Kapsam 1 (%30), Kapsam 2 (%30), Kapsam 3 (%20) ve Kapsam 4(%20) olacak şekilde maddeler sırasıyla 120, 120, 80, 80 madde olacak şekilde tüm maddeler rastgele olacak şekilde kapsamlara atanmıştır.

Yetenek Dağılımlarının Üretilmesi

Araştırmada dört farklı normal, sağa çarpık, sola çarpık ve tekdüze (uniform) yetenek dağılımı ele alınmıştır. Tüm yetenek dağılımlarında katılımcı sayısı N=1000 olarak belirlenmiştir. Sağa çarpık ve sola çarpık yetenek dağılımları Fleishman (1978) tarafından önerilen güç yöntemi (power method) ile normal dağılımdan elde edilmiştir. Fleishman'ın güç yöntemi denklemi aşağıdaki denklemde sunulmuştur:

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

Denklemde a,b,c,d olmak üzere dört sabit yer almaktadır. X ise kullanılan normal dağılımla elde edilen parametreleri ifade eder. Normal, sağa çarpık ve sola çarpık yetenek dağılımlarını üretmek için kullanılan X dağılımları ve a, b, c, d dağılımları Tablo 6'da yer almaktadır. Tekdüze yetenek dağılımı ise U~U(-3, +3), uniform dağılımdan elde edilmiştir.

Tablo 6

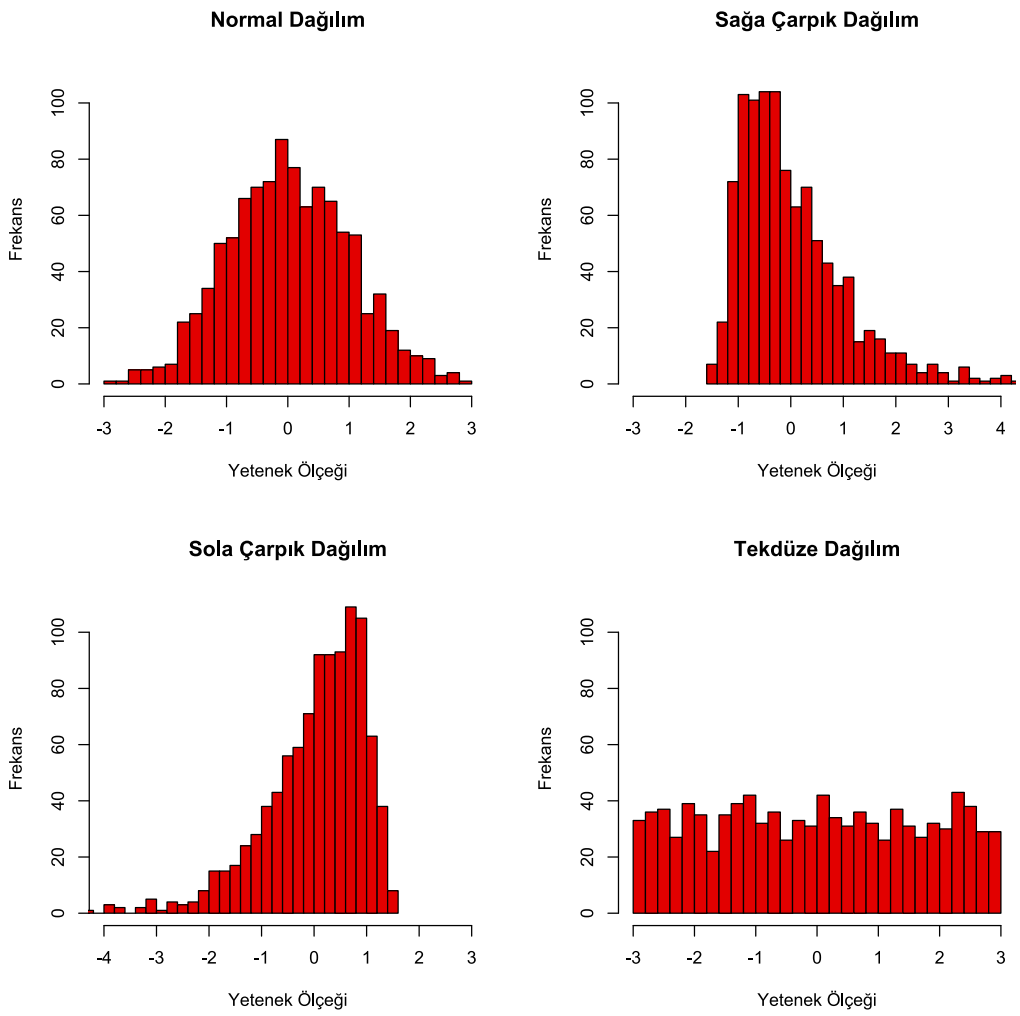
Çarpık Yetenek Dağılımlarının Üretilmesi

Dağılım	N	X	Çarpıklık	Basıklık	a	b	c	d
Normal	1000	N(0, 1)	0.00	0.00	0.00	1.00	0.00	0.00
Sağa Çarpık	1000	N(0, 1)	1.50	4.00	-0.21	0.85	-0.21	0.04
Sola Çarpık	1000	N(0, 1)	-1.50	4.00	0.21	0.85	-0.21	0.04
Uniform	1000	U(-3, +3)	-	-	-	-	-	-

Yetenek dağılımlarının histogram grafiği Şekil 12'de sunulmuştur.

Şekil 12

Yetenek Dağılımlarının Histogram Grafikleri



Araştırma Deseni

Çalışma kapsamında yürütülen simülasyon çalışmasına dair ayrıntılar aşağıda başlıklar halinde sunulmuştur.

Bu çalışmada farklı BÇAT yaklaşımları (S-BÇAT ve A-BÇAT), farklı benzetim koşullarında karşılaştırılmıştır. Değişimlenecek koşullar Tablo 7'de sunulmuştur.

Tablo 7

Manipüle Edilen Koşullar

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

Not: U: Uzun, K: Kısa

Tablo 7'de görüldüğü üzere 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 tekdüze) ve üç farklı modül/test uzunluğu oranı dağılımı (U-K-K, O-O-O, K-K-U) koşulları değişimlenerek simülasyon çalışması gerçekleştirilmiştir. Tüm koşullar birbirleri ile çaprazlanmıştır. Dolayısıyla birinci simülasyon çalışmasında 2x3x4x3=72 koşul incelenmiş ve her koşul için 100 replikasyon yapılarak analizler gerçekleştirilmiştir.

BÇAT Desenlerinin Oluşturulması

Araştırmada BÇAT deseni olarak 1-2-3 S-BÇAT ve A-BÇAT desenleri ele alınmıştır. Bu desenlerin oluşturulmasında "Rmst" ve "TestDesign" paketleri kullanılmıştır. İki farklı BÇAT deseni, dört farklı yetenek dağılımı, üç farklı test uzunluğu ve üç farklı modül/test uzunluğu oranı koşulları altında incelenmiştir.

1-2-3 S-BÇAT deseninde başlangıç aşamasında tek modül yer alırken, ikinci aşamada kolay ve zor olmak üzere iki modül, üçüncü aşamada ise kolay-orta-zor olmak üzere üç modül yer almaktadır. Toplamda 6 modülün yer aldığı desende, ilk modül (başlangıç modülü) $\theta = 0$ noktasında maksimum bilgi değerine ulaşacak şekilde birleştirilmiştir. Bu modülü alan katılımcılar geçici yetenek düzeyine göre kolay veya zor

modülden birine atanmakta, ardından üçüncü aşamada kestirilen geçici yetenek düzeyine göre kolay-orta-zor modülden birine atanarak uygulama tamamlamaktadır.

Bu çalışmada 1-2-3 BÇAT tasarımı tercih edilmiştir. Bu tasarımın tercih edilmesinin nedeni, 1-2-3 BÇAT tasarımının ikinci aşamasındaki modüllerin modül bilgi fonksiyonlarının maksimum olduğu noktalar ile diğer aşamaların modül bilgi fonksiyonunun maksimum olduğu noktalar ile çakışmamasıdır. Bir diğer deyişle, 1-3-3 BÇAT tasarımında ikinci ve üçüncü aşamadaki modüllerin modül bilgi fonksiyonunu maksimum yapan değerlerin çakışması, modüllerin kümülatif bilgi fonksiyonunu azaltıcı etkide bulunmaktadır. Cetin-Berber, Sari ve Huggins-Manley (2018) gerçekleştirdikleri çalışmada 1-2-3 ve 1-3-3 BÇAT tasarımlarının oldukça benzer ölçme kesinliği sonuçlarına sahip olduğunu raporlamaktadırlar. Benzer şekilde, Yiğiter ve Doğan (2023), 1-3, 1-2-3 ve 1-3-3 BÇAT tasarımlarını inceledikleri çalışmalarında 1-2-3 ve 1-3-3 BÇAT tasarımının oldukça benzer ölçme kesinliğine sahip olduğunu, hatta 1-2-3 BÇAT tasarımının az bir fark ile daha iyi ölçme kesinliği sunduğunu belirtmektedir. Dolayısıyla madde havuzundan daha efektif yararlanmak için 1-2-3 BÇAT tasarımı tercih edilmiştir.

1-2-3 BÇAT tasarımında modüllerin maksimum bilgi değerine ulaşacakları noktalar ve yönlendirme için kesme yetenek puanları ve modüllerde yer alacak madde sayıları Tablo 8’de sunulmuştur.

Tablo 8

Test Uzunluğuna Göre Modül Madde Sayıları ve Madde Kullanım Sıklığı Kontrolü

Desenler	Test Uzunluğu				Madde Kullanım Sıklığı	Panel Sayısı
	Adı	20	30	40		
		Modül Uzunluğu Dağılımı				
1-2-3 S-BÇAT	U-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-U	5-5-10	7-8-15	10-10-20		
A-BÇAT	U-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-U	5-5-10	7-8-15	10-10-20		

Not: *Anında BÇAT yönteminde madde kullanım sıklığı Uygunsuz (Ineligibility) yöntemi ile 0.33'e sabitlenerek kontrol edilmiştir (van der Linden & Weldkamp, 2004).

S-BÇAT yaklaşımında madde kullanım sıklıklarının kontrol altına alınması için panel sayısı kullanılmıştır. Her S-BÇAT tasarımı için madde kullanım sıklığının 0.33 olması için 3 panel oluşturulmuştur. A-BÇAT deseninde ise modüller anında oluşturulduğundan uygunsuzluk (ineligibility) (van der Linden & Weldkamp, 2004) yöntemi ile madde kullanım sıklığı oranı 0.33'e sabitlenerek kontrol edilmiştir.

A-BÇAT deseninde bireylerin başlangıç yetenek düzeyi (initial theta) $\theta = 0$ olarak belirlenmiştir. Bu yetenek düzeyine göre bir test birleştirilerek bireye başlangıç modülü olarak sunulmuştur. Ardından bireyin geçici yetenek düzeyine göre kısıtlamalar ve kapsam dengelemeleri de göz önüne alınarak bireyin geçici yetenek düzeyinde "anında" yeni bir test birleştirmesi yapılmış ve oluşan modül bireye sunulmuştur. Bu işlem tekrar gerçekleştirilmiş ve 1-2-3 S-BÇAT desenine benzer şekilde bireyin 3 modül alması ile test sonlandırılmış ve bireyin 3 modüle verdiği yanıtlara göre nihai yetenek düzeyi kestirilmiştir.

Modüllerin oluşturulmasında modül/test uzunluğu oranı dikkate alınmıştır. Örneğin 40 maddelik bir test uzunluğunda ve K-K-U (1/4-1/4-1/2) modül/test uzunluğu oranları için modül uzunlukları 10-10-20 olacak şekilde OTB işlemi gerçekleştirilmiştir. A-BÇAT için ise

bu oranları sağlamak amacıyla bu maddelerden hemen sonra yeni modül birleştirilerek sınav katılımcısına sunulacaktır. Örneğin 10-10-20 modül uzunluklarına sahip bir A-BÇAT tasarımı için 1-11-21. Madde konumlarında kısıtlamalar da göz önüne alınarak yeni modül birleştirilmiş ve sınav katılımcısına sunulmuştur.

Panellerin ve modüllerin test birleştirme ile oluşturulması için modül bilgilerinin maksimum hale getirildiği noktalar Tablo 9'da sunulmuştur.

Tablo 9

S- BÇAT Panel ve Modüllerinin Birleştirilmesi

Desenler	Aşama 1	Aşama 2	Aşama 3
1-2-3	ϑ = 0	ϑ = (-0.5, 0.5)	ϑ = (-1, 0, 1)
Anında BÇAT	ϑ = 0	$\vartheta = \vartheta^*$	$\vartheta = \vartheta^*$

Not: ϑ^* geçici yetenek düzeyi.

Her iki BÇAT tasarımında da başlangıç yetenek düzeyi ϑ = 0 olacak şekilde testler birleştirilmiştir. S-BÇAT yaklaşımında 1-2-3 tasarımı ile Tablo 9'da verilen yetenek düzeylerinde maksimum bilgi değerine ulaşacak şekilde ikinci ve üçüncü aşamaların modülleri oluşturulmuştur. 1-2-3 S-BÇAT tasarımının test birleştirme işleminde hibrit yöntemden (maddelerden bottom-up yöntemi ile modüllerin oluşturulduğu ve modüllerin ise top-down yöntemi ile bir araya getirilerek test tasarımının oluşturulduğu yöntemden) faydalanılmıştır. A-BÇAT'ta ise benzer şekilde ϑ = 0 noktasında ilk modül oluşturulmuş ve diğer modüller kestirilen geçici yetenek düzeyinin olduğu noktaya göre gölge test yaklaşımı ile birleştirilerek katılımcıya sunulmuştur. Hem S-BÇAT hem de A-BÇAT yaklaşımında test birleştirme için Rglpk (Makhorin, 2017) algoritmasından faydalanılmıştır.

Verilerin Analizi

Verilerin analizinden elde edilen sonuçların değerlendirilmesinde kestirilen ve gerçek yetenek parametreleri ile Hata Kareleri Ortalamasının Karekökü (Root Mean Square Error – RMSE), ortalama mutlak hata (mean absolute bias - MAB) ve yanlılık (BIAS)

değerleri kullanılmıştır. RMSE değerleri, n toplam birey sayısı, $\widehat{\theta_i}$ kestirilen yetenek düzeyi ve θ_i gerçek yetenek düzeyi olmak üzere aşağıda verilen formül ile hesaplanmıştır.

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

MAB değerinin hesaplanmasında kullanılan formül aşağıda yer almaktadır.

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

BIAS değerinin hesaplanmasında kullanılan formül aşağıda yer almaktadır.

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

Ayrıca, farklı benzetim koşulları altında iki farklı test tasarımının karşılaştırılması için elde edilen RMSE ve MAB değerleri ile etki büyüklüğü değerleri hesaplanmıştır. Cohen d değeri ile etki büyüklüğünün hesaplanmasında aşağıdaki formüller kullanılmıştır:

$$Harmanlanmış\ Standart\ Sapma = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2}} \quad (1)$$

$$Cohen\ d = \frac{Ortalama\ Farkları}{Harmanlanmış\ Standart\ Sapma} \quad (1)$$

Yukarıdaki formüller ile hesaplanan Cohen d değeri; d < 0.20 ise küçük etki, 0.20 < d < 0.50 ise orta etki, 0.80 < d ise büyük etki değeri olarak yorumlanmaktadır (Cohen, 1988).

Madde güvenliğinin incelenmesinde ise her madde için madde kullanım sıklığı hesaplanarak incelenmiştir. Madde kullanım sıklığı aşağıdaki formül ile hesaplanmıştır.

$$Madde\ Maruziyet\ Oranı = \frac{n_M}{n_T} \quad (1)$$

Formülde yer alan n_M, m maddesinin uygulandığı katılımcı sayısını ifade ederken, n_T ise toplam katılımcı sayısını ifade etmektedir.

Bölüm 4

Bulgular, Yorumlar ve Tartışma

Bu bölümde öncelikle araştırma problemine dair tüm koşullardan elde edilen bulgular sunulmuştur. Ardından sırasıyla alt problemlerine göre bulgulara yer verilmiştir.

Araştırma Probleminin Ölçme Kesinliğine İlişkin Bulgular

Araştırma problemi "Belirlenen farklı benzetim koşulları altında S-BÇAT ve A-BÇAT yaklaşımlarının ölçme kesinliği ve madde güvenliği ne düzeyde değişim göstermektedir?" şeklindeydi. Bu başlıkta bu araştırma probleminin ölçme kesinliği bölümüne yanıt vermek için aynı madde havuzu ve aynı yetenek dağılımları altında S-BÇAT ve A-BÇAT yöntemlerinden elde edilen yetenek kestirimleri karşılaştırılmıştır. İncelenen 72 koşuldan elde edilen sonuçlar Tablo 10'da sunulmuştur.

Tablo 10'da yer alan sütunlar sırasıyla test uzunluğu, modül/test uzunluğu oranı ve yetenek dağılımını göstermektedir. Ardından gelen sütunlarda S-BÇAT ve A-BÇAT yöntemlerine göre kestirilen yetenek düzeylerine göre RMSE, MAB ve BIAS değerleri yer almaktadır. Ayrıca replikasyonlardan elde edilen RMSE ve MABdeğerlerine göre Cohen d değeri ile etki büyüklüğü hesaplanarak raporlanmıştır.

Tablo 10

Farklı BÇAT Yaklaşımlarına Göre Tüm Koşullardan Elde Edilen Bulgular

Koşul	Test Uzunluğu	Modül/Test Uzunluğu Oranı	Yetenek Dağılımı	RMSE			MAB			BIAS	
				S-BÇAT	A-BÇAT	d_{RMSE}	S-BÇAT	A-BÇAT	d_{OMH}	S-BÇAT	A-BÇAT
1	20	U-K-K	Normal	0,382	0,371	1,657	0,300	0,291	1,587	0,008	0,007
2	20	U-K-K	Sağa Çarpık	0,423	0,404	2,369	0,315	0,303	1,604	0,050	0,040
3	20	U-K-K	Sola Çarpık	0,431	0,414	1,887	0,311	0,303	1,070	-0,028	-0,029
4	20	U-K-K	Tekdüze	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	Tekdüze	0,552	0,524	2,235	0,432	0,409	2,715	-0,058	-0,047
9	20	K-K-U	Normal	0,398	0,364	4,013	0,314	0,285	4,164	0,008	0,004
10	20	K-K-U	Sağa Çarpık	0,431	0,399	5,287	0,325	0,301	4,368	0,039	0,039
11	20	K-K-U	Sola Çarpık	0,438	0,405	4,146	0,322	0,297	4,928	-0,021	-0,022
12	20	K-K-U	Tekdüze	0,559	0,519	3,640	0,440	0,406	3,592	-0,047	-0,047
13	30	U-K-K	Normal	0,338	0,323	1,884	0,266	0,253	2,148	0,007	0,009
14	30	U-K-K	Sağa Çarpık	0,380	0,361	2,540	0,281	0,268	2,004	0,044	0,039
15	30	U-K-K	Sola Çarpık	0,388	0,367	2,808	0,277	0,266	1,843	-0,025	-0,018
16	30	U-K-K	Tekdüze	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	Tekdüze	0,476	0,456	2,233	0,370	0,355	1,870	-0,052	-0,044
21	30	K-K-U	Normal	0,34	0,317	2,568	0,268	0,251	2,441	0,006	0,008
22	30	K-K-U	Sağa Çarpık	0,372	0,344	4,317	0,279	0,259	3,640	0,037	0,032
23	30	K-K-U	Sola Çarpık	0,385	0,356	3,615	0,28	0,258	4,004	-0,021	-0,017
24	40	K-K-U	Tekdüze	0,477	0,444	3,127	0,372	0,346	2,886	-0,047	-0,043
25	40	U-K-K	Normal	0,306	0,295	1,579	0,24	0,232	1,456	0,005	0,002
26	40	U-K-K	Sağa Çarpık	0,337	0,328	1,638	0,251	0,244	1,554	0,036	0,036
27	40	U-K-K	Sola Çarpık	0,350	0,336	2,158	0,249	0,243	1,206	-0,021	-0,015
28	40	U-K-K	Tekdüze	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	Tekdüze	0,433	0,413	2,361	0,336	0,320	2,297	-0,045	-0,042
33	40	K-K-U	Normal	0,317	0,291	3,733	0,250	0,230	3,350	0,005	0,002
34	40	K-K-U	Sağa Çarpık	0,345	0,317	3,744	0,259	0,239	3,640	0,030	0,029
35	40	K-K-U	Sola Çarpık	0,357	0,330	3,610	0,259	0,240	3,183	-0,018	-0,014
36	40	K-K-U	Tekdüze	0,433	0,412	2,219	0,337	0,317	2,872	-0,039	-0,044

Not: U: Uzun, O: Orta, K: Kısa.

Tablo 10'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 üzere tüm koşullarda S-BÇAT yönteminden elde edilen RMSE değerleri, A-BÇAT'tan elde edilen RMSE değerinden daha yüksektir. Bu durum, A-BÇAT'ın tüm koşullarda S-BÇAT'a göre daha etkili yetenek kestirimi yaptığını göstermektedir. Ayrıca d_{RMSE} sütununda görüldüğü üzere tüm koşullarda Cohen d etki büyüklüğü değeri [Cohen d > .80] olup tüm koşullarda A-BÇAT yöntemi S-BÇAT yöntemine göre büyük etki göstermiştir.

S-BÇAT'tan elde edilen MAB değerleri [0.250, 0.443] aralığında değerler değişmektedir. A-BÇAT'tan elde edilen MAB değerleri ise [0.230, 0.419] aralığındadır. Tabloda görüldüğü üzere tüm koşullarda S-BÇAT'tan elde edilen MAB değeri A-BÇAT'tan elde edilen MAB değerinden daha yüksektir. MAB değerleri de A-BÇAT'ın S-BÇAT'a göre etkili kestirim yaptığını doğrulamaktadır. Yine benzer şekilde d_{BIAS} sütununda görüldüğü üzere tüm koşullarda Cohen d eti büyüklüğü değeri [d > .80] olup A-BÇAT yönteminin S-BÇAT yöntemine göre büyük etki gösterdiğini belirtmektedir.

Literatürde bu araştırmanın ölçme kesinliği bulgularını destekleyen çalışmalar bulunmaktadır. Zheng ve Chang (2015), BBT, A-BÇAT ve S-BÇAT'ı karşılaştırdığı çalışmasında bu araştırmanın bulgularına benzer şekilde A-BÇAT'ın S-BÇAT'tan daha iyi ölçme kesinliği sunduğunu belirtmektedir. Tay (2015) ise sınıflama doğruluğu ve sınıflama tutarlılığına göre A-BÇAT'ın S-BÇAT'tan daha iyi sonuçlar sunduğunu ifade etmektedir. Literatürde yer alan diğer çalışmalar da bu çalışmanın bulgularına benzer şekilde A-BÇAT'ın S-BÇAT'tan daha iyi ölçme kesinliği sunduğunu söylemektedir (Zheng & Chang, 2015; Tay, 2015; van der Linden & Diao, 2016; Han & Guo, 2016; van der Linden, 2021).

Hem RMSE hem de MAB sütunlarından anlaşılacağı üzerei test uzunluğu arttıkça her iki test yaklaşımının da ölçme kesinliğinin belirgin bir şekilde arttığı görülmektedir. Diğer taraftan, modül/test uzunluğu oranına göre S-BÇAT'ta K-K-U oranının U-K-K ve O-O-O oranına daha düşük ölçme kesinliğine sahip iken, A-BÇAT'ta ise her üç oranın benzer ölçme

kesinliğine sahip olduğu söylenebilir. Yetenek dağılımlarına göre normal dağılımlarda her iki yöntemin de en yüksek ölçme kesinliğine sahip olduğu görülürken, normal dağılımın ardından sağa ve sola çarpık dağılım geldiği görülmektedir. Her iki test yaklaşımının yetenek dağılımı açısından en düşük ölçme kesinliğine tekdüze dağılımın sahip olduğu belirtilebilir.

BIAS değerlerine göre ise her iki BÇAT yönteminin de yanlılık düzeylerinin oldukça düşük ve benzer sonuçlar ürettiği görülmektedir. Normal dağılımlarda her iki BÇAT yönteminin BIAS değerlerinin oldukça küçük olduğu görülmektedir. Sağa çarpık, sola çarpık ve tekdüze dağılımlarda ise BIAS değerinin normal dağılımdan fazla olduğu belirtilebilir.

Yetenek ölçeğine göre S-BÇAT ve A-BÇAT yöntemlerinden elde edilen ölçme kesinliği (RMSE ve MAB) sonuçları Ek-3'te yer alan grafiklerde sunulmuştur. Araştırmanın devamında test uzunluğu, modül/test uzunluğu oranı ve yetenek dağılımına göre S-BÇAT ve A-BÇAT yöntemlerinin ölçüm kesinliği sonuçları karşılaştırılmıştır.

Birinci Alt Araştırma Problemine İlişkin Bulgular

Birinci alt araştırma problemi "Test uzunluğuna (20-30-40) göre S-BÇAT ve A-BÇAT yaklaşımlarının RMSE, MAB ve BIAS değerleri nasıl değişim göstermektedir?" şeklindeydi. Test uzunluğuna göre Tablo 10'daki ilgili hücrelerin ortalamalarından elde edilen sonuçlar Tablo 11'de sunulmuştur.

Tablo 11

Test Uzunluğuna Göre RMSE, MAB ve d değerleri

Test Uzunluğu	RMSE			MAB			BIAS	
	S-BÇAT	A-BÇAT	d_{RMSE}	S-BÇAT	A-BÇAT	d_{BIAS}	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 11 incelendiğinde 20 test uzunluğu için ortalama RMSE değeri S-BÇAT yaklaşımı için 0.451 iken, A-BÇAT yaklaşımı için 0.425'tir. Benzer şekilde 20 test uzunluğu için ortalama MAB değeri S-BÇAT için 0.344 iken A-BÇAT için 0.325'tir. Ayrıca ortalama Cohen d etki büyüklüğü değerinin de RMSE için 3.013, MAB için 2.809 olduğu görülmektedir. 20 test uzunluğu için A-BÇAT yaklaşımının S-BÇAT'a göre daha ölçme kesinliğine sahip olduğu görülmektedir.

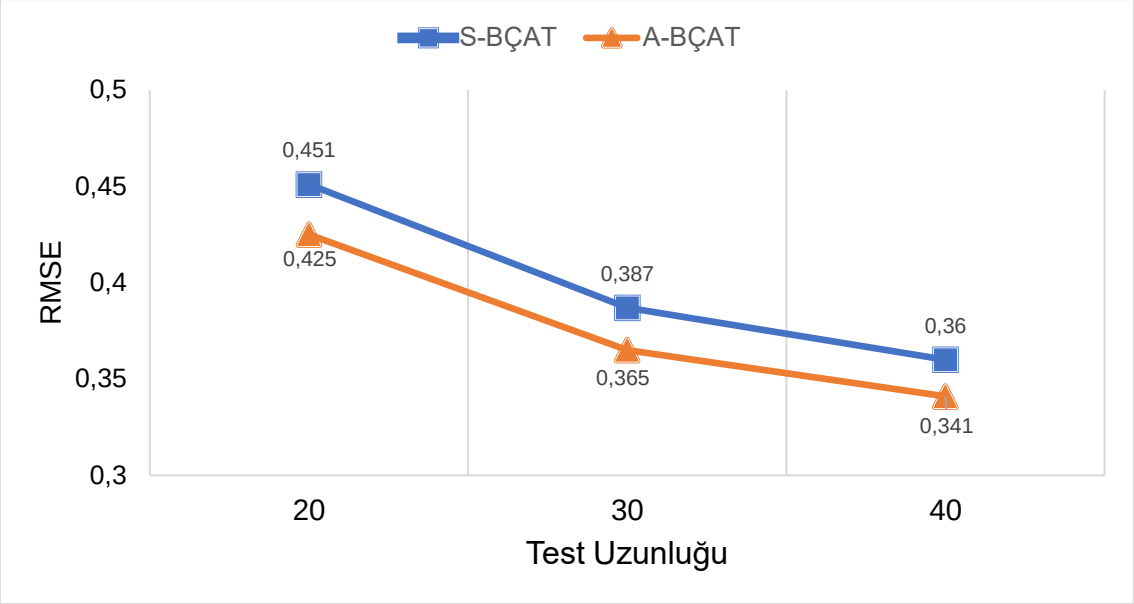
30 test uzunluğu için ortalama RMSE değeri S-BÇAT yaklaşımı için 0.387 iken, A-BÇAT yaklaşımı için 0.365'tir. Benzer şekilde 30 test uzunluğu için ortalama MAB değeri S-BÇAT için 0.293 iken A-BÇAT için 0.277'dir. Ayrıca ortalama Cohen d etki büyüklüğü değerinin de RMSE için 2.783, MAB için 2.516 olduğu görülmektedir. 20 tet uzunluğuna benzer şekilde 30 test uzunluğu için de A-BÇAT yaklaşımının S-BÇAT'a göre daha iyi yetenek kestirimi yaptığı belirtilebilir.

40 test uzunluğu için ortalama RMSE değeri S-BÇAT yaklaşımı için 0.360 iken, A-BÇAT yaklaşımı için 0.341'dir. Benzer şekilde 40 test uzunluğu için ortalama MAB değeri S-BÇAT için 0.272 iken A-BÇAT için 0.258'dir. Ayrıca ortalama Cohen d etki büyüklüğü değerinin de RMSE için 2.578, MAB için 2.411 olduğu görülmektedir. Diğer test uzunluklarına benzer şekilde, 40 test uzunluğu için A-BÇAT yaklaşımının S-BÇAT'a göre daha iyi yetenek kestirimi yaptığı söylenebilir.

Her üç test uzunluğu (20, 30, 40) düzeyinde de RMSE, MAB ve d değerlerine göre A-BÇAT yaklaşımının S-BÇAT yaklaşımına göre daha etkili yetenek kestirimi sunduğu görülmektedir. Tablo 11'e göre RMSE bulguları Şekil 13'de, MAB bulguları Şekil 14'te sunulmuştur.

Şekil 13

Test Uzunluğuna Göre RMSE Değerleri Grafiği

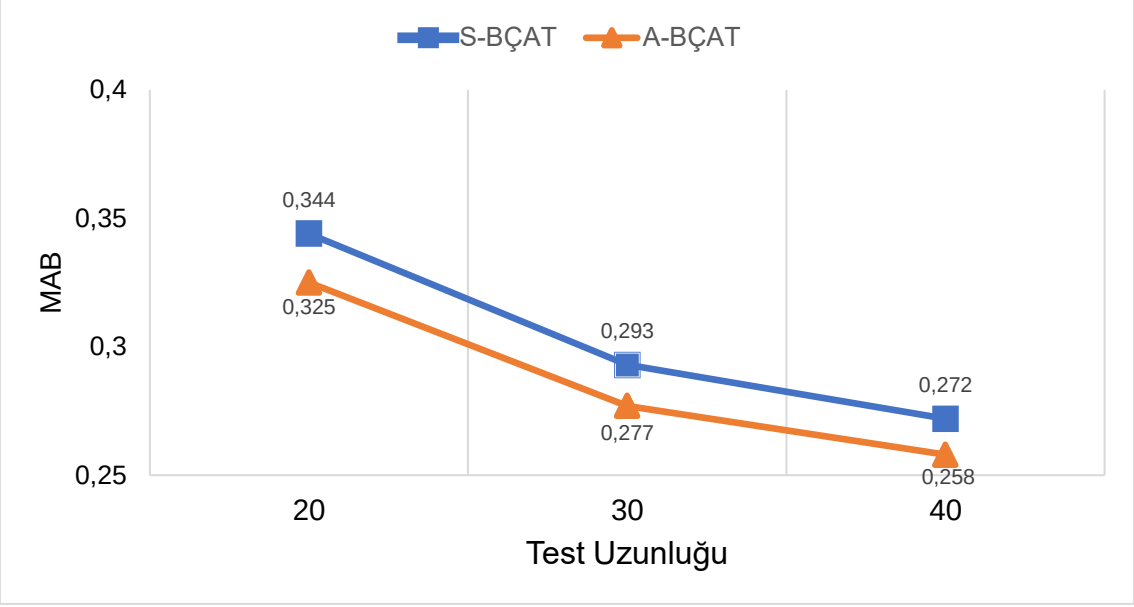


Şekil 13 incelendiğinde her üç test uzunluğu (20, 30, 40) düzeyinde de A-BÇAT'ın RMSE değerlerinin daha düşük olduğu, dolayısıyla S-BÇAT yaklaşımına göre daha iyi ölçme kesinliği sunduğu belirtilmişti. 20 test uzunluğu ile 30 test uzunluğu arasındaki RMSE değeri farkı S-BÇAT ve A-BÇAT için sırasıyla 0.064 (0.451-0.387) ve 0.060 (0.425-0.365) iken, 30 test uzunluğu ile 40 test uzunluğu RMSE değeri farkı sırasıyla 0.027 (0.387-0.360) ve 0.024 (0.365-0.341)'tür. Benzer şekilde Şekil 3 incelendiğinde, 20 test uzunluğu ile 30 test uzunluğu arasındaki MAB değeri farkı S-BÇAT ve A-BÇAT için sırasıyla 0.051 (0.344-0.293) ve 0.048 (0.325-0.277) iken, 30 test uzunluğu ile 40 test uzunluğu RMSE değeri farkı sırasıyla 0.021 (0.293-0.272) ve 0.019 (0.277-0.258)'dur. 20 ile 30 test uzunluğu değeri arasındaki MAB farkı fazla iken, 30 ve 40 test uzunluğu arasındaki MAB farkı azalmıştır. Hem RMSE farkı hem de MAB farkına göre bu durum, test uzunluğu arttırılsa da yetenek kestiriminin etkililiğinin test uzunluğu ile orantılı olarak artmayacağını göstermektedir. Azalan verimler kanunu, üretim faktörlerinin artması ile üretimin aynı oranda artış göstermeyeceğini, oransal olarak faydanın giderek azalacağını öne sürmektedir. Testlerde test uzunluğunun arttırılması belirli bir noktaya kadar ölçme kesinliğini iyileştirecektir. Belirli

noktadan daha fazla test uzunluğunun arttırılması ölçme kesinliğinde istenen iyileştirmeyi yapmayacağı gibi öğrencinin yorgunluk ve sıkılma-bunalma gibi fizyolojik ve psikolojik etkilerinden dolayı beklenen verim sağlanamayacaktır. Dolayısıyla test uzunluğunun iyi belirlenmesi gerekir. Bu araştırmada 20 test uzunluğunun 30 ve 40 test uzunluğuna göre düşük ölçme kesinliği gösterdiği Şekil 13 ve Şekil 14'ten anlaşılmaktadır. Dolayısıyla 30 veya 40 test uzunluğunun bu araştırma sonuçlarına göre daha iyi sonuçlar verdiği söylenebilir. 50, 60, 70 maddelik test uzunluklarının ise hem uyarlanabilir test yaklaşımlarının mantığı ile uyumlu olmadığı düşünülmektedir hem de azalan verimler kanununa göre ölçme kesinliğinin oransal olarak azalması beklenmemelidir (Yasuda, Mae, Hull & Taniguchi, 2021).

Şekil 14

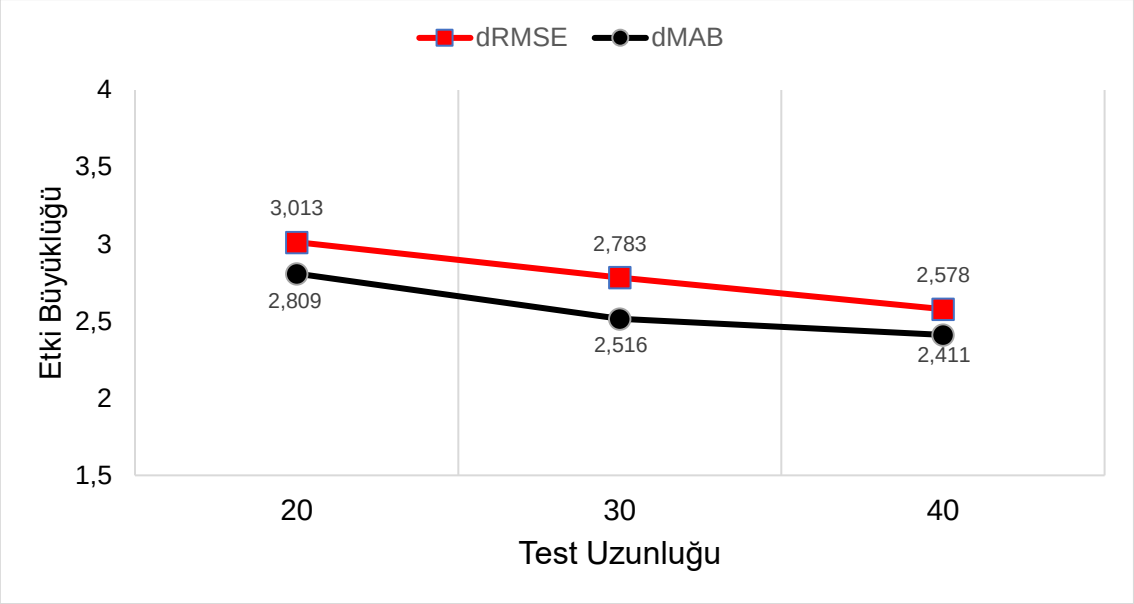
Test Uzunluğuna Göre MAB Değerleri Grafiği



Test uzunluklarına göre Tablo 11'de yer alan ortalama Cohen d etki büyüklükleri Şekil 15'te görselleştirilerek sunulmuştur.

Şekil 15

Test Uzunluğuna Göre Ortalama Etki Büyüklükleri

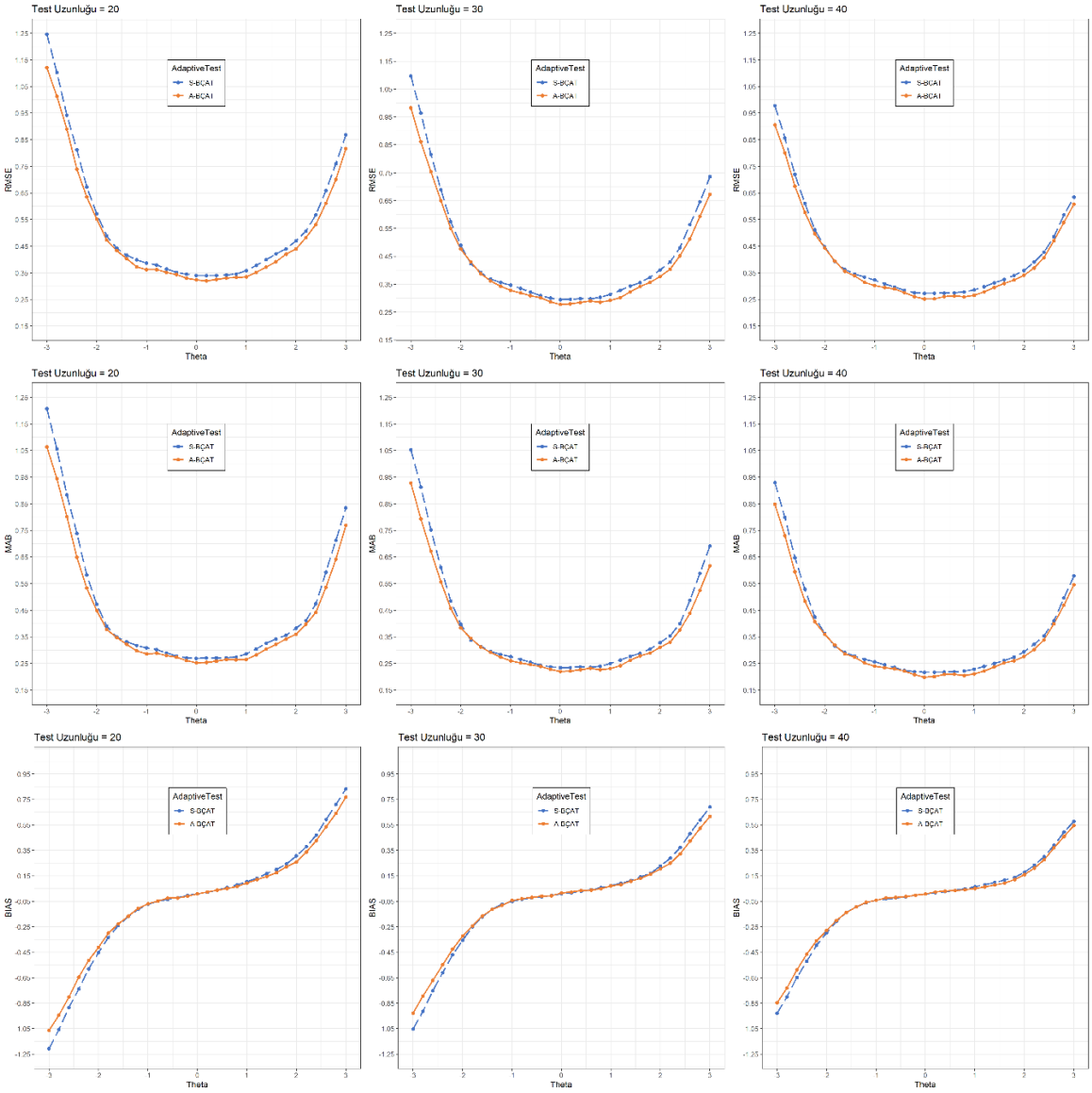


Şekil 15'te görüldüğü üzere S-BÇAT ile A-BÇAT arasındaki ortalama etki büyüklükleri 20 test uzunluğunda RMSE için 3.013, MAB için 2.809, 30 test uzunluğunda RMSE için 2.783, MAB için 2.516, 40 test uzunluğunda RMSE için 2.578, MAB için 2.411 olarak hesaplanmıştır. Bu durum, test uzunluğu azaldıkça A-BÇAT'ın S-BÇAT'a göre daha etkili kestirimler yaptığını göstermektedir. Diğer taraftan, her üç test uzunluğunda da A-BÇAT'ın S-BÇAT'a göre büyük etki gösterdiği belirtilmelidir. Literatürde, bu çalışmanın bulgularını destekleyecek şekilde, uyarlanabilir testlerde test uzunluğu arttıkça hataların azaldığı ve ölçme kesinliğinin iyileştiği bulgusunu bildiren pek çok çalışma bulunmaktadır (Balta & Uçar, 2022; Demir, 2022; Demir & Atar, 2021; Erdem Kara & Doğan, 2021; Gökçe & Glas, 2018; Gündeğer & Soysal, 2022; Gür & Gülleroğlu, 2020; Özberk & Gelbal, 2017; Ozdemir & Gelbal, 2022; Şahin, 2020).

Son olarak, test uzunluğuna göre A-BÇAT ve S-BÇAT arasındaki farkı daha ayrıntılı incelemek için yetenek düzeyine göre RMSE ve MAB grafikleri oluşturulmuştur. Grafikler Şekil 16'da sunulmuştur.

Şekil 16

Test Uzunluğuna Göre Yetenek Ölçeğinde Farklı Test Yaklaşımlarının RMSE, MAB ve BIAS Grafikleri



Şekil 16 incelendiğinde, tüm test uzunluğu koşullarında ve tüm yetenek ölçeğinde A-BÇAT'ın S-BÇAT'a göre iyi ölçme kesinliğine ulaştığı görülmektedir. Yetenek ölçeğinin orta noktalarında (-1:1 aralığında) S-BÇAT'ın ölçme kesinliği A-BÇAT'tan biraz düşük ölçme kesinliğine sahiptir. Yetenek ölçeğinin uç noktalarına doğru gidildikçe ise iki test yaklaşımının ölçme kesinliği arasındaki fark A-BÇAT lehine artmaktadır. Dolayısıyla özellikle uçyetenek düzeylerinde A-BÇAT'ın S-BÇAT'tan bariz bir şekilde daha iyi ölçme

kesinliğine ulaştığı söylenebilir. Diğer taraftan, üç test uzunluğuna göre incelendiğinde 20 test uzunluğunun uç yetenek düzeyindeki RMSE değerleri [1.15, 1.25] aralığında iken, 30 test uzunluğunda RMSE değerleri [1.00, 1.10] düzeyine düşmekte, 40 test uzunluğunda ise [0.90, 1.00] aralığına gerilemektedir. Bu durum test uzunluğu arttığında her iki test yaklaşımının uç yetenek düzeylerinde daha etkili kesitirim yaptığını göstermektedir. Ayrıca yetenek ölçeğinin orta noktalarında ise 20 test uzunluğu için RMSE değeri 0.35 düzeyinde iken, 30 test uzunluğu için 0.30 düzeyinde olmakta, 40 test uzunluğu düzeyi için ise 0.25 düzeyine gerilemektedir. Orta düzeyde de benzer şekilde test uzunluğu ile beraber ölçme kesinliğinin azaldığı söylenebilir.

İkinci Alt Araştırma Problemine İlişkin Bulgular

İkinci alt araştırma problemi "Yetenek dağılımına (normal dağılım, sağa çarpık, sola çarpık, uniform) göre S-BÇAT ve A-BÇAT yaklaşımlarının RMSE, MAB ve BIAS değerleri nasıl değişim göstermektedir?" şeklindeydi. Yetenek dağılımına göre Tablo 10'daki ilgili hücrelerin ortalamalarından elde edilen sonuçlar Tablo 12'de sunulmuştur.

Tablo 12

Yetenek Dağılımına Göre RMSE, MAB, BIAS ve d değerleri

Yetenek Dağılımı	RMSE			MAB			BIAS	
	S-BÇAT	A-BÇAT	d_{RMSE}	S-BÇAT	A-BÇAT	d_{BIAS}	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
Tekdüze	0,492	0,465	2,708	0,384	0,362	2,732	-0,048	-0,044

Tablo 12 incelendiğinde yetenek dağılımlarından normal dağılım için ortalama RMSE değeri S-BÇAT yaklaşımı için 0.345 iken, A-BÇAT yaklaşımı için 0.326 olarak hesaplanmıştır. Benzer şekilde normal dağılım için ortalama MAB değeri S-BÇAT için 0.272 iken A-BÇAT için 0.257'dir. Ayrıca ortalama Cohen d etki büyüklüğü değerinin de RMSE

için 2.401, MAB için 2.330 olduğu görülmektedir. Normal dağılım için A-BÇAT yaklaşımının S-BÇAT'a göre daha iyi yetenek kestirimi yaptığı görülmektedir.

Sağa çarpık dağılım için ortalama RMSE değeri S-BÇAT yaklaşımı için 0.379 iken, A-BÇAT yaklaşımı için 0.358 olarak hesaplanmıştır. Benzer şekilde sağa çarpık dağılım için ortalama MAB değeri S-BÇAT için 0.284 iken A-BÇAT için 0.269'dur. Ayrıca ortalama Cohen d etki büyüklüğü değerinin de RMSE için 3.054, MAB için 2.584 olduğu görülmektedir. Normal dağılıma benzer şekilde sağa çarpık dağılım için de A-BÇAT yaklaşımının S-BÇAT'a göre daha iyi yetenek kestirimi yaptığı belirtilebilir.

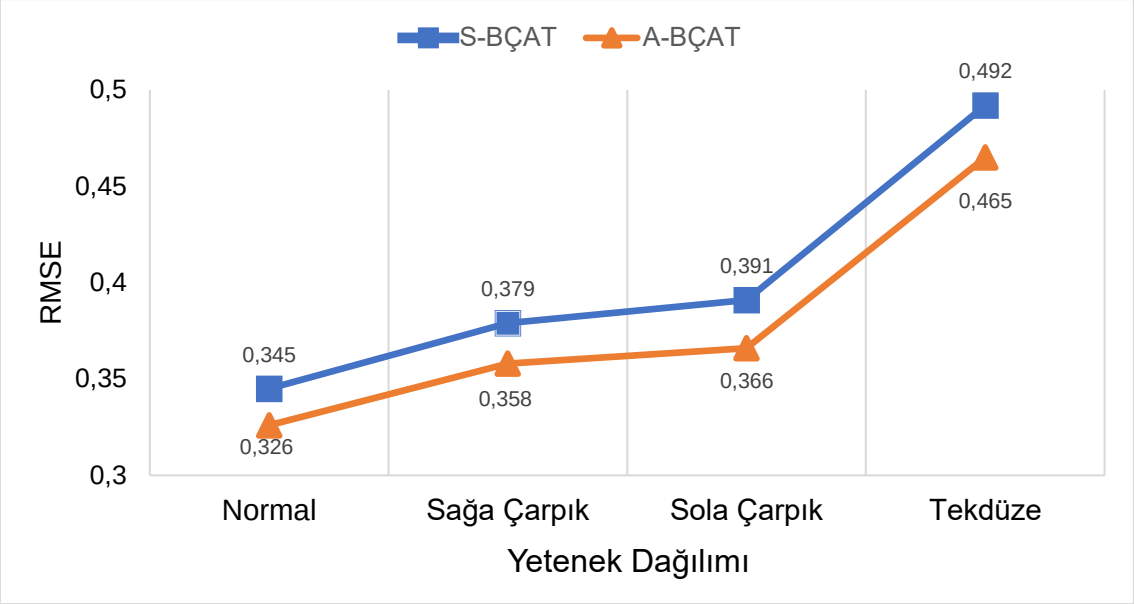
Sola çarpık dağılım için ortalama RMSE değeri S-BÇAT yaklaşımı için 0.391 iken, A-BÇAT yaklaşımı için 0.366'dır. Benzer şekilde sola çarpık dağılım için ortalama MAB değeri S-BÇAT için 0.282 iken A-BÇAT için 0.267'dir. Ayrıca ortalama Cohen d etki büyüklüğü değerinin de RMSE için 3.040, MAB için 2.711 olduğu görülmektedir. Diğer yetenek dağılımlarına benzer şekilde, sola çarpık yetenek dağılımı için de A-BÇAT yaklaşımının S-BÇAT'a göre daha iyi yetenek kestirimi yaptığı söylenebilir.

Tekdüze dağılım için ortalama RMSE değeri S-BÇAT yaklaşımı için 0.492 iken, A-BÇAT yaklaşımı için 0.465'dır. Benzer şekilde tekdüze dağılım için ortalama MAB değeri S-BÇAT için 0.384 iken A-BÇAT için 0.362'dir. Ayrıca ortalama Cohen d etki büyüklüğü değerinin de RMSE için 2.708, MAB için 2.732 olduğu görülmektedir. Diğer yetenek dağılımlarına benzer şekilde, tekdüze yetenek dağılımı için de A-BÇAT yaklaşımının S-BÇAT'a göre hem RMSE hem MAB hem de Cohen d etki büyüklüğü değerlerine göre daha iyi yetenek kestirimi yaptığı söylenebilir.

Dört farklı yetenek dağılımı (normal, sağa çarpık, sola çarpık ve tekdüze) düzeyinde de RMSE, MAB ve d değerlerine göre A-BÇAT yaklaşımının S-BÇAT yaklaşımına göre daha etkili yetenek kestirimi sunduğu görülmektedir. Tablo 12'ye göre farklı yetenek dağılımları bulgularından elde edilen RMSE bulguları Şekil 17'de, MAB bulguları Şekil 18'de sunulmuştur.

Şekil 17

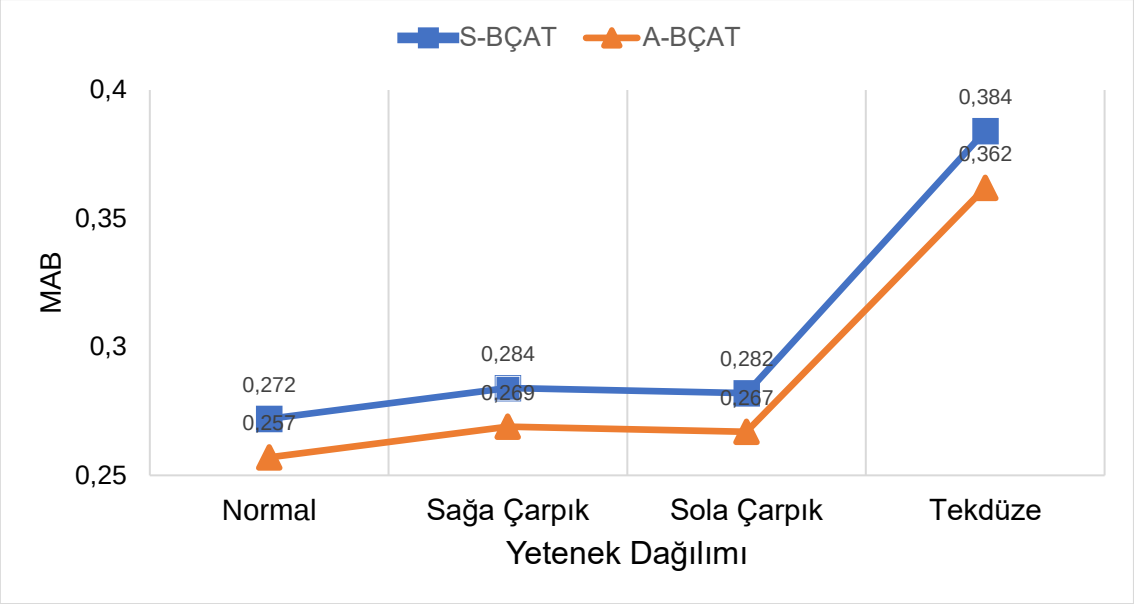
Yetenek Dağılımlarına Göre RMSE Değerleri



Şekil 17 incelendiğinde tüm yetenek dağılımlarında (normal, sağa çarpık, sola çarpık) A-BÇAT'ın RMSE değerlerinin daha düşük olduğu, dolayısıyla S-BÇAT yaklaşımına göre daha iyi ölçme kesinliği sunmaktadır. Hem S-BÇAT hem de A-BÇAT normal dağılımda en iyi ölçme kesinliğini sunmaktadır. Ölçme kesinliği açısından normal dağılımı sağa ve sola çarpık dağılım izlemektedir. Tekdüze dağılım ise ölçme kesinliği açısından en son sırada yer almaktadır. Bu durumun nedeni normal, sağa ve sola çarpık dağılımlarda yetenek ölçeğinin orta düzeylerindeki birey sayısı fazla iken, tekdüze dağılımlarda tüm yetenek ölçeğinde eşit bir dağılımın olmasıdır.

Şekil 18

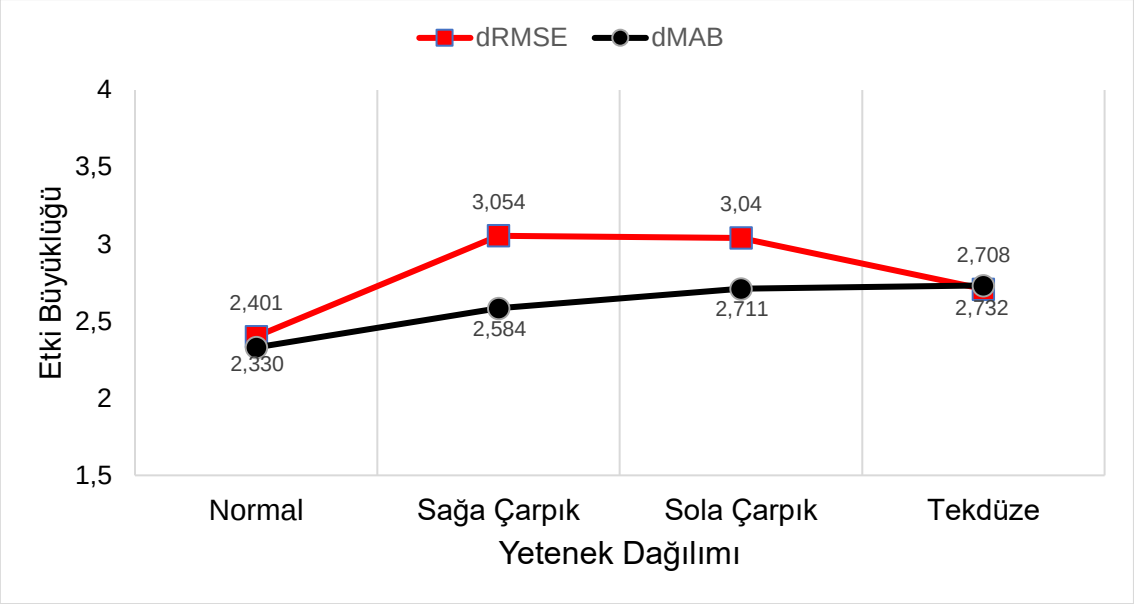
Yetenek Dağılımlarına Göre MAB Değerleri



Şekil 18 incelendiğinde MAB değerlerinde de RMSE değerlerine benzer şekilde her iki test yaklaşımı da normal dağılımda en iyi ölçme kesinliğini sunmaktadır. Normal dağılımı, sağa ve sola çarpık dağılım izlemektedir. Tekdüze dağılım ise ölçme kesinliği açısından son sırada yer almaktadır.

Şekil 19

Test Uzunluğuna Göre Ortalama Etki Büyüklükleri



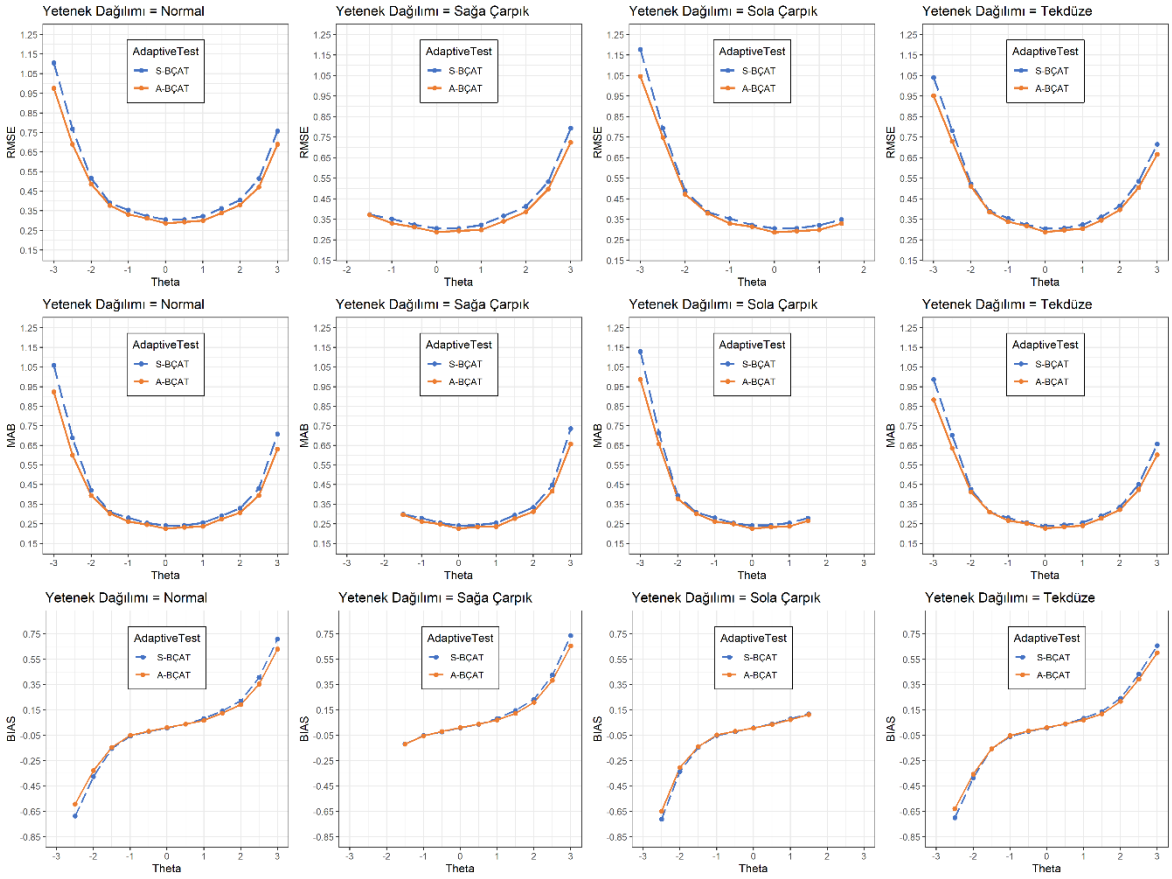
Şekil 19'da görüldüğü üzere S-BÇAT ile A-BÇAT arasındaki ortalama Cohen d etki büyüklükleri normal dağılımda RMSE için 2.401, MAB için 2.330, sağa çarpık dağılımda RMSE için 3.054, MAB için 2.584, sola çarpık dağılımda RMSE için 3.040, MAB için 2.711, tekdüze dağılımda RMSE için 2.702, MAB için 2.732 olarak hesaplanmıştır. Bu durumda A-BÇAT, S-BÇAT'a göre normal dağılımda daha düşük etki büyüklüğüne sahipken, normal dağılımı sırasıyla tekdüze, sola çarpık ve sağa çarpık dağılım izlemektedir. Diğer taraftan, her üç test uzunluğunda da A-BÇAT'ın S-BÇAT'a göre büyük etki gösterdiği vurgulanmalıdır (Cohen d > .80).

Tablo 12'deki BIAS değerleri incelendiğinde normal dağılım için BIAS değerinin düşük olduğu, çarpık dağılımlarda ise BIAS değerinin arttığı gözlemlenmektedir. Her iki BÇAT yaklaşımı için de sağa çarpık dağılımlarda pozitif, sola çarpık ve tekdüze dağılımlarda ise negatif yanlılık oluştuğu söylenebilir. Diğer taraftan, BIAS değerlerine göre A-BÇAT'ın S-BÇAT'tan çok az fark ile daha iyi yanlılık sonuçları sunduğu belirtilebilir.

Yetenek dağılımına göre A-BÇAT ve S-BÇAT arasındaki ölçme kesinliği farkını daha ayrıntılı incelemek için yetenek ölçeğinin düzeylerine göre RMSE ve MAB grafikleri oluşturulmuştur. Grafikler Şekil 20'de sunulmuştur.

Şekil 20

Yetenek Dağılımına Göre Yetenek Ölçeğinde Farklı Test Yaklaşımlarının RMSE, MAB ve BIAS Grafikleri



Şekil 20 incelendiğinde, tüm farklı yetenek dağılımı koşullarında ve tüm yetenek ölçeğinde A-BÇAT'ın S-BÇAT'a göre daha iyi ölçme kesinliğine ulaştığı görülmektedir. Yetenek ölçeğinin orta noktalarında (-1:1 aralığında) S-BÇAT'ın ölçme kesinliği A-BÇAT'tan oldukça yakın ölçme kesinliğine sahip iken, yetenek ölçeğinin uç noktalarına doğru gidildikçe ise iki test yaklaşımının ölçme kesinliği arasındaki fark A-BÇAT lehine artmaktadır. Dolayısıyla özellikle uçyetenek düzeylerinde A-BÇAT'ın S-BÇAT'tan bariz bir şekilde daha iyi ölçme kesinliğine ulaştığı ifade edilebilir. Normal dağılım ve tekdüze

dağılımın yetenek ölçeğinde RMSE ve MAB değerleri oldukça benzerdir. Bu iki dağılımın yetenek ölçeğinin uçnoktalarında ise tekdüze dağılımın daha düşük RMSE ve MAB değerlerine sahip olduğu görünmektedir. Bu durumun sebebi, tekdüze dağılımda uç yetenek düzeylerinde daha fazla katılımcının bulunmasıdır. Diğer taraftan, sağa çarpık dağılımda ise grafik yetenek ölçeğinin -1.5 düzeyinde, sağa çarpık dağılımda ise 1. Düzeyinde sonlanmaktadır. Bu durumun sebebi dağılımların şekli itibariyle bu noktaların daha ilerisinde katılımcı bulunmamaktadır (bkz. Şekil 12).

Üçüncü Alt Araştırma Problemine İlişkin Bulgular

Üçüncü alt araştırma problemi "Modül/test uzunluğu oranına (U-K-K, O-O-O, K-K-U) göre S-BÇAT ve A-BÇAT yaklaşımlarının RMSE, MAB ve BIAS değerleri nasıl değişim göstermektedir?" şeklindeydi. Modül/test oranına göre Tablo 11'deki ilgili hücrelerin ortalamalarından elde edilen sonuçlar Tablo 13'te sunulmuştur.

Tablo 13

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_{BIAS}	S-BÇAT	A-BÇAT
U-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-U	0,404	0,375	3,668	0,309	0,286	3,589	-0,006	-0,006

U: Uzun, O: Orta, K: Kısa.

Tablo 13 incelendiğinde U-K-K oranı için ortalama RMSE değeri S-BÇAT yaklaşımı için 0.402 iken, A-BÇAT yaklaşımı için 0.385'tir. Benzer şekilde U-K-K oranı için ortalama MAB değeri S-BÇAT için 0.305 iken A-BÇAT için 0.292'dir. Ayrıca ortalama Cohen d etki büyüklüğü değerinin de RMSE için 2.257, MAB için 1.902 olduğu görülmektedir. U-K-K oranı için A-BÇAT yaklaşımının S-BÇAT'a göre daha iyi düzeyde yetenek kestirimi yaptığı görülmektedir.

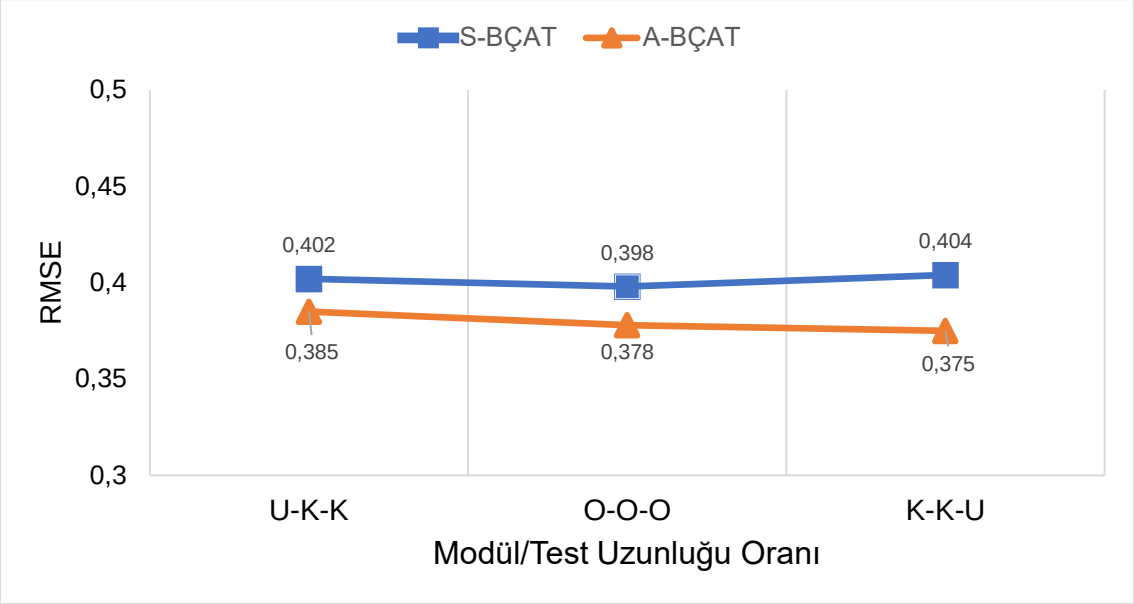
O-O-O oranı için ortalama RMSE değeri S-BÇAT yaklaşımı için 0.398 iken, A-BÇAT yaklaşımı için 0.378'dir. Benzer şekilde O-O-O oranı için ortalama MAB değeri S-BÇAT için 0.303 iken A-BÇAT için 0.288'dir. Ayrıca ortalama Cohen d etki büyüklüğü değerinin de RMSE için 2.478, MAB için 2.277 olduğu görülmektedir. U-K-K oranına benzer şekilde, O-O-O oranı için de A-BÇAT yaklaşımının S-BÇAT'a göre daha iyi yetenek kestirimi yaptığı söylenebilir.

K-K-U oranı için ortalama RMSE değeri S-BÇAT yaklaşımı için 0.404 iken, A-BÇAT yaklaşımı için 0.375'dir. Benzer şekilde K-K-U oranı için ortalama MAB değeri S-BÇAT için 0.309 iken A-BÇAT için 0.286'dır. Ayrıca ortalama Cohen d etki büyüklüğü değerinin de RMSE için 3.668, MAB için 3.589 olduğu görülmektedir. Diğer oranlara benzer şekilde, K-K-U oranı için de A-BÇAT yaklaşımının S-BÇAT'a göre daha iyi yetenek kestirimi yaptığı söylenebilir.

Her modül/test uzunluğu oranı (U-K-K, O-O-O, K-K-U) düzeyinde de RMSE, MAB ve d değerlerine göre A-BÇAT yaklaşımının S-BÇAT yaklaşımına göre daha etkili yetenek kestirimi sunduğu görülmektedir. Tablo 13'e göre RMSE bulguları Şekil 21'de, MAB bulguları Şekil 22'de görselleştirilerek sunulmuştur.

Şekil 21

Modül/Test Uzunluğu Oranına Göre RMSE Değerleri

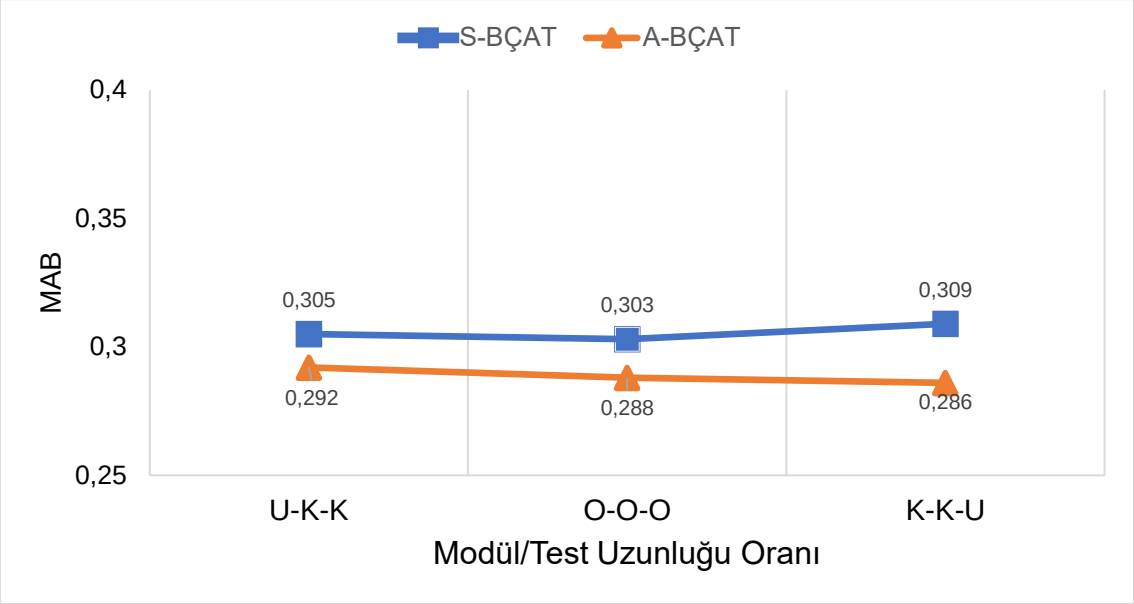


Not: U: Uzun, O: Orta, K: Kısa.

Şekil 21 incelendiğinde her modül/test uzunluğu oranı (U-K-K, O-O-O, K-K-U) düzeyinde de A-BÇAT'ın RMSE değerlerinin daha düşük olduğu, dolayısıyla S-BÇAT yaklaşımına göre daha iyi ölçme kesinliği sunduğu görülmektedir. S-BÇAT ve A-BÇAT yaklaşımlarının RMSE değeri farkı U-K-K oranına göre 0.017 (0.402-0.385), O-O-O oranına göre 0.020 (0.398-0.378) ve K-K-U oranına göre 0.029 (0.404-0.375)'dir. Bu durumda A-BÇAT yaklaşımının ölçme kesinliği RMSE değeri farklarına göre K-K-U oranı düzeyinde daha yüksektir. U-K-K ve O-O-O oranları düzeyinde ise daha benzerdir. Diğer taraftan, S-BÇAT yöntemi, en iyi ölçme kesinliğini RMSE değerlerine göre en iyi O-O-O oranında gösterirken, A-BÇAT yaklaşımı ise U-K-K oranında göstermektedir.

Şekil 22

Modül/Test Uzunluğu Oranına Göre MAB Değerleri



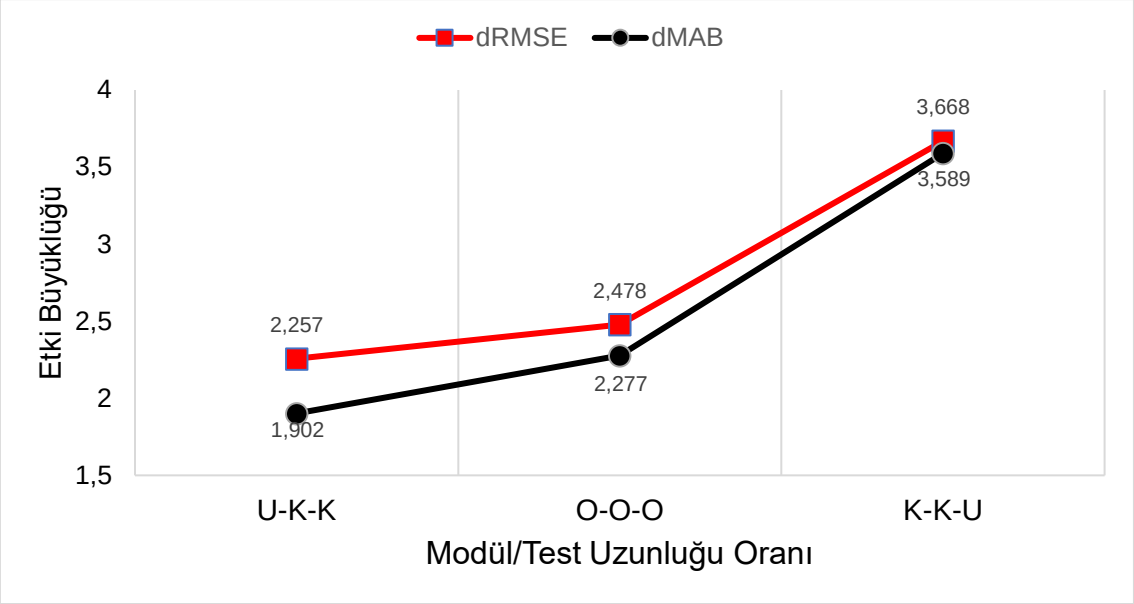
Not: U: Uzun, O: Orta, K: Kısa.

Şekil 22 incelendiğinde RMSE sonuçlarına benzer şekilde, tüm modül/test uzunluğu oranlarında A-BÇAT yaklaşımının S-BÇAT'a göre daha iyi ölçme kesinliği sunduğu görülmektedir. Diğer taraftan, U-K-K ve O-O-O oranlarında MAB farkları benzerlik gösterirken, K-K-U oranında MAB farkının arttığı gözlenmektedir. MAB değerlerine göre S-BÇAT yaklaşımı en iyi ölçme kesinliğini O-O-O oranında sunarken, A-BÇAT yaklaşımı K-K-U oranında sunmaktadır.

Modül/test uzunluğu oranlarına göre Tablo 13'te yer alan ortalama Cohen d etki büyüklükleri Şekil 23'de görselleştirilerek sunulmuştur.

Şekil 23

Modül/Test Uzunluğu Oranına Göre Ortalama Etki Büyüklükleri



Not: U: Uzun, O: Orta, K: Kısa.

Şekil 23'de görüldüğü üzere S-BÇAT ile A-BÇAT arasındaki ortalama etki büyüklükleri U-K-K oranında RMSE için 2.257, MAB için 1.902, O-O-O oranında RMSE için 2,478, MAB için 2.277, K-K-U oranında ise RMSE için 3.668, MAB için 3.589 olarak hesaplanmıştır. Bu durumda A-BÇAT ile S-BÇAT yaklaşımının arasındaki ölçme kesinliği farkı U-K-K oranında en az iken, O-O-O oranında da benzerdir. K-K-U oranında ise iki yaklaşım arasındaki ölçme kesinliği farkı en yüksek değerine ulaşmıştır. Bu durumda son modüldeki S-BÇAT yaklaşımında son modüldeki madde sayısı arttıkça ölçme kesinliğinin azaldığı yorumu yapılabilir. Diğer taraftan A-BÇAT yaklaşımında modüller sabit olmadığından A-BÇAT, modül/test oranı uzunluğundan daha az düzeyde etkilendiği görülmektedir. Dahası, A-BÇAT yaklaşımında son modülün uzunluğunu arttırmak bu yaklaşımın ölçme kesinliğini arttırmıştır. Ayrıca, her üç modül/test uzunluğu oranında da A-BÇAT'ın S-BÇAT'a göre büyük etki gösterdiği belirtilmelidir.

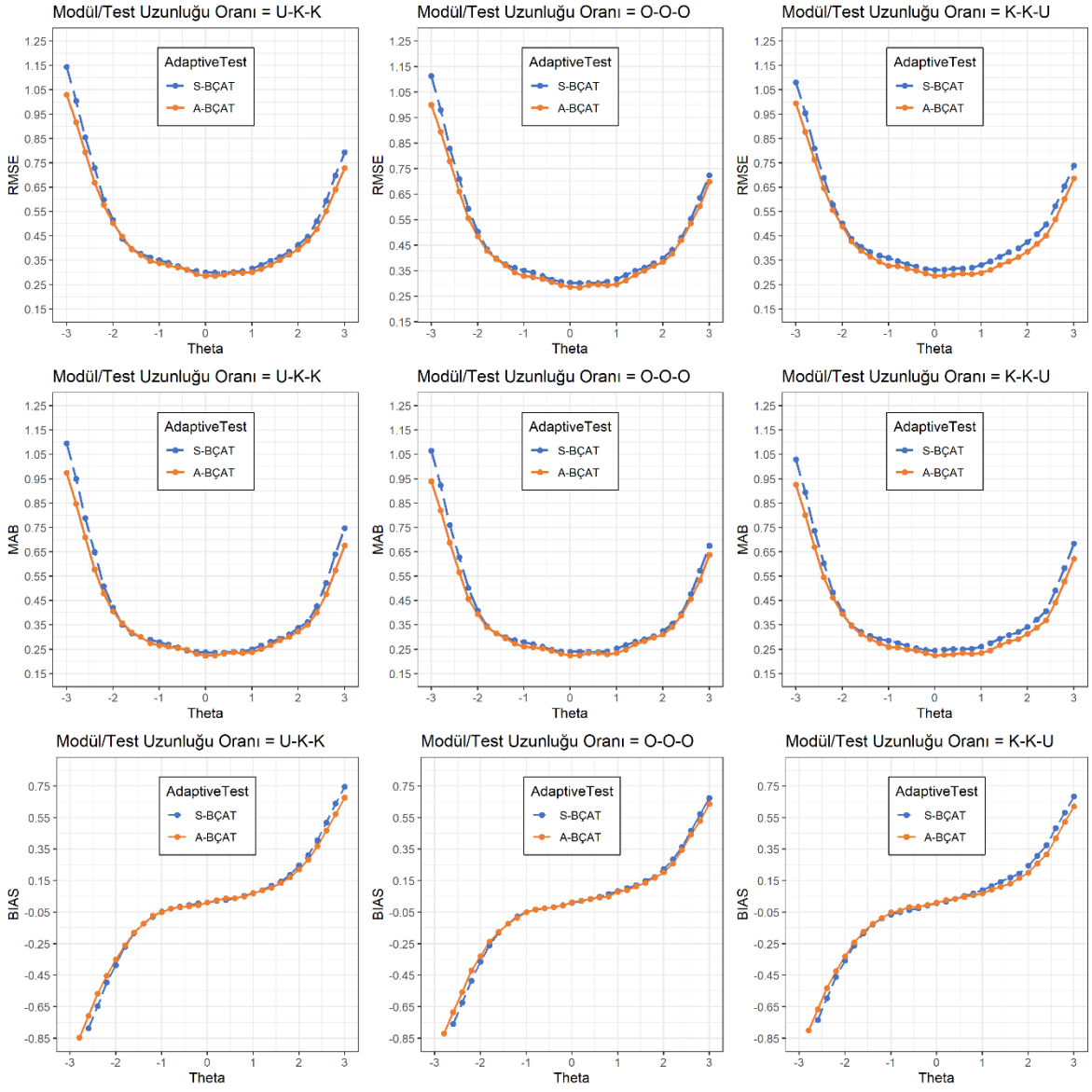
Literatürde bu araştırmanın bulguları ile benzerlik gösterecek şekilde yönlendirme modülünün uzunluğu arttığında S-BÇAT'ın daha iyi ölçme kesinliği sunduğunu belirten çalışmalar bulunmaktadır (Kim & Plake, 1993; Zheng, 2016; Boztunç, 2019; Cai, Anthony,

Albano & Roussos, 2021). Diğer taraftan, son modülün uzunluğu arttıkça S-BÇAT'ın ölçme kesinliği azalırken, A-BÇAT'ın ise ölçme kesinliğinin arttığı görülmüştür.

Son olarak, modül/test uzunluğu oranına göre A-BÇAT ve S-BÇAT arasındaki farkı daha ayrıntılı incelemek için yetenek düzeyine göre RMSE ve MAB grafikleri oluşturulmuştur. Grafikler Şekil 24'te sunulmuştur.

Şekil 24

Yetenek Dağılımına Göre Yetenek Ölçeğinde Farklı Test Yaklaşımlarının RMSE, MAB ve BIAS Grafikleri



Not: U: Uzun, O: Orta, K: Kısa.

Şekil 24 incelendiğinde, tüm farklı modül/test uzunluğu oranı koşullarında ve tüm yetenek ölçeğinde A-BÇAT'ın S-BÇAT'a göre daha iyi ölçme kesinliğine ulaştığı görülmektedir. Diğer yetenek ölçeği grafiklerine benzer şekilde yetenek ölçeğinin orta noktalarında (-1:1 aralığında) S-BÇAT, A-BÇAT'a oldukça yakın ölçme kesinliğine sahip iken, yetenek ölçeğinin uç noktalarına doğru gidildikçe ise iki test yaklaşımının ölçme kesinliği arasındaki fark A-BÇAT lehine artmaktadır. Diğer taraftan, Şekil 13'te görüldüğü üzere U-K-K ve O-O-O oranlarında S-BÇAT ve A-BÇAT'ın RMSE ve MAB grafikleri arasındaki ölçme kesinliği farkı oldukça az iken, K-K-U oranında ölçme kesinliği farkı artmıştır.

Araştırma Probleminin Madde Güvenliğine İlişkin Bulgular

Araştırma problemi "Belirlenen farklı benzetim koşulları altında S-BÇAT ve A-BÇAT yaklaşımlarının ölçme kesinliği ve madde güvenliği ne düzeyde değişim göstermektedir?" şeklindeydi. Bu bölümde araştırma probleminin madde güvenliği bölümüne yanıt vermek için tüm koşullara göre madde kullanım sıklığı betimsel istatistikleri Tablo 14'te, kullanılan madde sayıları Tablo 15'te sunulmuştur.

Tablo 14 incelendiğinde, S-BÇAT'ın ortalama minimum madde kullanım sıklığı ortalaması 0.054 iken, A-BÇAT'ın ortalama minimum madde kullanım sıklığı 0.001'dir. S-BÇAT'ın modül ve panelleri sabit olduğundan panel ve modüllerde yer alan maddelerin en düşük madde kullanım sıklığı %5. İken, A-BÇAT'ta modüller anında oluşturulduğundan ortalama minimum madde kullanım sıklığı %0.001'dir. Ortalama maksimum madde kullanım sıklığı S-BÇAT için 0.350 iken; A-BÇAT için 0.334'dür. Dolayısıyla A-BÇAT'ın maksimum madde kullanım sıklığının S-BÇAT'tan daha düşük olduğu görülmektedir.

Tablo 14

Madde Kullanım Sıklıkları

Koşul	Test Uzunluğu	Modül/Test Uzunluğu Oranı	Yetenek Dağılımı	S-BÇAT			A-BÇAT		
				Min	Maks	Ort	Min	Maks	Ort
1	20	U-K-K	Normal	0,046	0,356	0,190	0,001	0,369	0,123
2	20	U-K-K	Sağa Çarpık	0,063	0,345	0,167	0,001	0,371	0,125
3	20	U-K-K	Sola Çarpık	0,050	0,340	0,148	0,001	0,364	0,128
4	20	U-K-K	Tekdüze	0,061	0,353	0,192	0,001	0,371	0,115
5	20	O-O-O	Normal	0,055	0,355	0,167	0,001	0,312	0,114
6	20	O-O-O	Sağa Çarpık	0,063	0,345	0,147	0,001	0,316	0,116
7	20	O-O-O	Sola Çarpık	0,062	0,343	0,190	0,001	0,323	0,118
8	20	O-O-O	Tekdüze	0,060	0,353	0,167	0,001	0,319	0,110
9	20	K-K-U	Normal	0,054	0,343	0,148	0,001	0,302	0,114
10	20	K-K-U	Sağa Çarpık	0,049	0,344	0,190	0,001	0,308	0,114
11	20	K-K-U	Sola Çarpık	0,048	0,348	0,167	0,001	0,308	0,116
12	20	K-K-U	Tekdüze	0,049	0,346	0,148	0,001	0,300	0,108
13	30	U-K-K	Normal	0,049	0,345	0,192	0,001	0,355	0,136
14	30	U-K-K	Sağa Çarpık	0,054	0,368	0,167	0,001	0,356	0,136
15	30	U-K-K	Sola Çarpık	0,052	0,346	0,147	0,001	0,355	0,136
16	30	U-K-K	Tekdüze	0,055	0,350	0,190	0,001	0,359	0,129
17	30	O-O-O	Normal	0,049	0,355	0,167	0,001	0,329	0,131
18	30	O-O-O	Sağa Çarpık	0,049	0,338	0,148	0,001	0,323	0,130
19	30	O-O-O	Sola Çarpık	0,056	0,341	0,190	0,001	0,331	0,132
20	30	O-O-O	Tekdüze	0,043	0,358	0,167	0,001	0,321	0,123
21	30	K-K-U	Normal	0,041	0,355	0,148	0,001	0,318	0,126
22	30	K-K-U	Sağa Çarpık	0,054	0,363	0,192	0,001	0,323	0,127
23	30	K-K-U	Sola Çarpık	0,059	0,339	0,167	0,001	0,330	0,128
24	40	K-K-U	Tekdüze	0,055	0,368	0,147	0,001	0,294	0,119
25	40	U-K-K	Normal	0,057	0,336	0,190	0,001	0,351	0,150
26	40	U-K-K	Sağa Çarpık	0,056	0,367	0,167	0,001	0,354	0,150
27	40	U-K-K	Sola Çarpık	0,044	0,343	0,148	0,001	0,357	0,152
28	40	U-K-K	Tekdüze	0,062	0,355	0,190	0,001	0,374	0,142
29	40	O-O-O	Normal	0,057	0,345	0,167	0,001	0,333	0,14
30	40	O-O-O	Sağa Çarpık	0,052	0,357	0,148	0,001	0,335	0,142
31	40	O-O-O	Sola Çarpık	0,064	0,363	0,192	0,001	0,333	0,142
32	40	O-O-O	Tekdüze	0,057	0,351	0,167	0,001	0,345	0,135
33	40	K-K-U	Normal	0,054	0,355	0,147	0,001	0,331	0,138
34	40	K-K-U	Sağa Çarpık	0,059	0,339	0,190	0,001	0,333	0,139
35	40	K-K-U	Sola Çarpık	0,056	0,340	0,167	0,001	0,332	0,140
36	40	K-K-U	Tekdüze	0,053	0,351	0,148	0,001	0,313	0,132

Not: U: Uzun, O: Orta, K: Kısa, Min: Minimum, Maks: Maksimum, Ort: Ortalama.

Ortalama madde kullanım sıklıklarının ortalaması ise S-BÇAT için 0.168 iken, A-BÇAT için 0.129'dur. A-BÇAT'ın ortalama madde kullanım sıklığının S-BÇAT'a göre düşük olması A-BÇAT'ın madde havuzundan daha fazla maddeden faydalandığını ve dolayısıyla madde havuzunu daha efektif kullandığının bir göstergesi olarak yorumlanabilir.

Bu araştırmada S-BÇAT ile A-BÇAT'ın madde kullanım sıklıklarının sınırlandırılması için için S-BÇAT'ta 3 panel oluşturulmuştu ve A-BÇAT'ta ise uygunsuzluk (ineligibility) yöntemi ile madde kullanım sıklığı 0.33'e sabitlenmişti. Bu bulgulara göre A-BÇAT, S-BÇAT'a göre minimum, maksimum ve ortalama madde kullanım sıklığı değerlerine göre daha iyi sonuçlar vermektedir. Dolayısıyla A-BÇAT'ın madde güvenliği açısından S-BÇAT'a göre daha güvenli olduğu söylenebilir.

Tablo 15'te araştırmada ele alınan 72 koşula göre koşullarda kullanılan madde sayıları sunulmuştur. Madde havuzundan daha fazla maddenin kullanılması hem madde kullanım sıklığını düşürmekte hem de maddelerin açığa çıkma olasılığını azaltması açısından madde güvenliğini arttırmaktadır.

Tablo 15 incelendiğinde tüm koşullarda madde havuzundan A-BÇAT'ın S-BÇAT'tan fazla sayıda maddeyi kullandığı görülmektedir. Özellikle test uzunluğu arttıkça madde kullanım sayıları birbirine yaklaşsa da A-BÇAT'ın daha fazla madde kullandığı belirtilmelidir.

Tüm koşullara göre madde bazında madde kullanım sıklıklarını içeren grafikler araştırmanın son bölümünde Ek-4'te sunulmuştur. Ayrıca aşağıda yer alan ilgili değişkenlere göre madde bazında madde kullanım sıklıklarının incelenmesi için bu grafiklerden bazıları örnek olarak gösterilmektedir. Ek-4'te yer alan grafiklerin x ekseninde madde sayısı yer alırken, y ekseninde madde kullanım sıklıkları görülmektedir. Grafiklerin üzerinde yer alan kesikli çizgi ise kullanılan maddelerin ortalama madde kullanım sıklığını göstermektedir.

Araştırmanın devamında ise test uzunluğu, yetenek dağılımı ve modül/test uzunluğu değişkenlerine göre madde kullanım sıklığı ve madde kullanım sayıları incelenmiştir.

Tablo 15

Test Yaklaşımlarına Göre Kullanılan Madde Sayıları

Koşul	Test Uzunluğu	Modül/Test Uzunluğu Oranı	Yetenek Dağılımı	Toplam Madde Sayısı	Kullanılan Madde Sayısı		
					S-BÇAT	A-BÇAT	Fark
1	20	U-K-K	Normal	400	105	163	-58
2	20	U-K-K	Sağa Çarpık	400	105	160	-55
3	20	U-K-K	Sola Çarpık	400	105	156	-51
4	20	U-K-K	Tekdüze	400	105	174	-69
5	20	O-O-O	Normal	400	120	175	-55
6	20	O-O-O	Sağa Çarpık	400	120	173	-53
7	20	O-O-O	Sola Çarpık	400	120	170	-50
8	20	O-O-O	Tekdüze	400	120	181	-61
9	20	K-K-U	Normal	400	135	176	-41
10	20	K-K-U	Sağa Çarpık	400	135	176	-41
11	20	K-K-U	Sola Çarpık	400	135	173	-38
12	20	K-K-U	Tekdüze	400	135	185	-50
13	30	U-K-K	Normal	400	156	221	-65
14	30	U-K-K	Sağa Çarpık	400	156	221	-65
15	30	U-K-K	Sola Çarpık	400	156	221	-65
16	30	U-K-K	Tekdüze	400	156	232	-76
17	30	O-O-O	Normal	400	180	229	-49
18	30	O-O-O	Sağa Çarpık	400	180	230	-50
19	30	O-O-O	Sola Çarpık	400	180	227	-47
20	30	O-O-O	Tekdüze	400	180	244	-64
21	30	K-K-U	Normal	400	204	239	-35
22	30	K-K-U	Sağa Çarpık	400	204	236	-32
23	30	K-K-U	Sola Çarpık	400	204	235	-31
24	40	K-K-U	Tekdüze	400	204	252	-48
25	40	U-K-K	Normal	400	210	267	-57
26	40	U-K-K	Sağa Çarpık	400	210	267	-57
27	40	U-K-K	Sola Çarpık	400	210	264	-54
28	40	U-K-K	Tekdüze	400	210	281	-71
29	40	O-O-O	Normal	400	240	285	-45
30	40	O-O-O	Sağa Çarpık	400	240	282	-42
31	40	O-O-O	Sola Çarpık	400	240	282	-42
32	40	O-O-O	Tekdüze	400	240	297	-57
33	40	K-K-U	Normal	400	270	289	-19
34	40	K-K-U	Sağa Çarpık	400	270	288	-18
35	40	K-K-U	Sola Çarpık	400	270	286	-16
36	40	K-K-U	Tekdüze	400	270	304	-34

Not: U: Uzun, O: Orta, K: Kısa.

Araştırmanın devamında alt problemlere göre test uzunluğu, yetenek dağılımı ve modül/test uzunluğu açısından madde kullanım sıklığı ve kullanılan madde sayıları incelenmiştir.

Dördüncü Alt Araştırma Problemine İlişkin Bulgular

Dördüncü alt araştırma problemi "Farklı test uzunluklarında (20-30-40) S-BÇAT ve A-BÇAT yaklaşımlarının madde kullanım sıklığına ve kullanılan madde sayılarına göre madde güvenliği nasıl değişim göstermektedir?" şeklindeydi. Farklı test uzunluklarına göre Tablo 14 ve Tablo 15'teki ilgili hücrelerin ortalamalarından elde edilen sonuçlar Tablo 16'da sunulmuştur.

Tablo 16

Test Uzunluğuna Göre Kullanılan Madde Sayıları ve Ortalama Madde Kullanım Sıklıkları

Test Uzunluğu	Kullanılan Madde Sayıları		Ortalama Madde Kullanım Sıklıkları	
	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 16 incelendiğinde 20 test uzunluğu için ortalama kullanılan madde sayıları S-BÇAT için 120 iken, A-BÇAT için 172'dir. Bu bulguya benzer şekilde ortalama madde kullanım sıklığı S-BÇAT için 0.168 iken A-BÇAT için 0.117'dir. 20 test uzunluğunda A-BÇAT yaklaşımının hem kullanılan madde sayısı hem de ortalama madde kullanım sıklıklarına göre madde güvenliği açısından daha iyi sonuçlar sunduğu belirtilebilir.

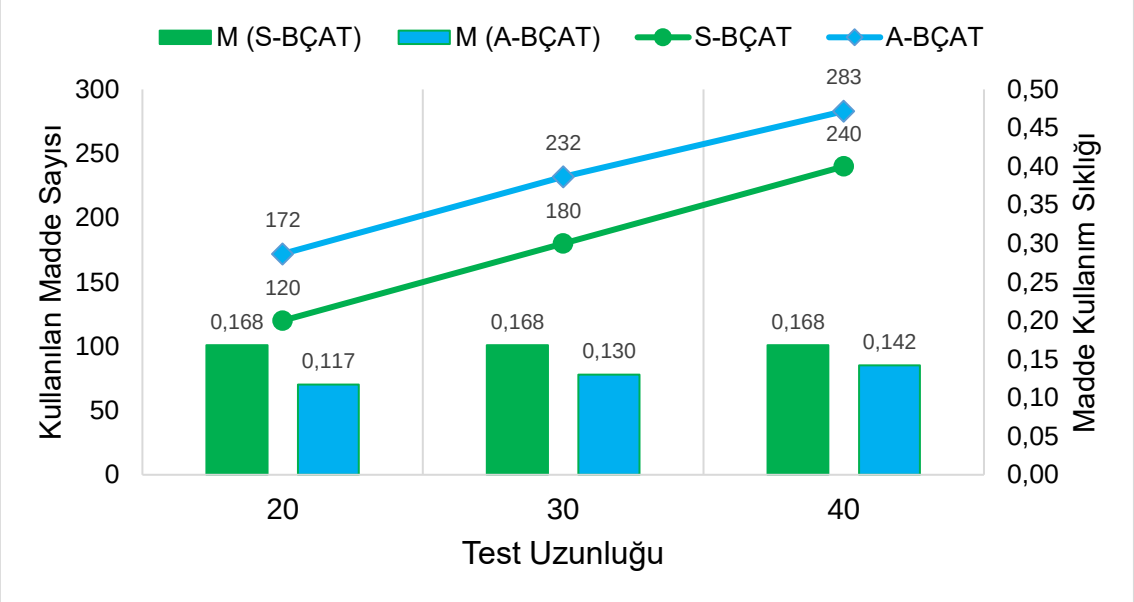
30 test uzunluğu için ortalama kullanılan madde sayıları S-BÇAT için 180 iken, A-BÇAT için 232'dur. Benzer şekilde ortalama madde kullanım sıklığı S-BÇAT için 0.168 iken A-BÇAT için 0.130'dur. 30 test uzunluğunda A-BÇAT yaklaşımının hem kullanılan madde sayısı hem de ortalama madde kullanım sıklıklarına göre madde güvenliği açısından daha iyi sonuçlar sunduğu söylenebilir.

40 test uzunluğunda ise ortalama kullanılan madde sayıları S-BÇAT için 240 iken, A-BÇAT için 283'tür. Benzer şekilde ortalama madde kullanım sıklığı S-BÇAT için 0.168 iken A-BÇAT için 0.142'dir. 40 test uzunluğunda A-BÇAT yaklaşımının hem kullanılan madde sayısı hem de ortalama madde kullanım sıklıklarına göre madde güvenliği açısından daha iyi sonuçlar sunduğu ifade edilebilir.

Tablo 16'da yer alan bulgular üzerinde karşılaştırma yapmak amacıyla Şekil 25'te grafik üzerinde görselleştirilmiştir.

Şekil 25

Test Uzunluğuna Göre Kullanılan Madde Sayıları ve Ortalama Madde Kullanım Sıklıkları



Not: M(S-BÇAT): S-BÇAT ortalama madde kullanım sıklığı, M(A-BÇAT): A-BÇAT ortalama madde kullanım sıklığı.

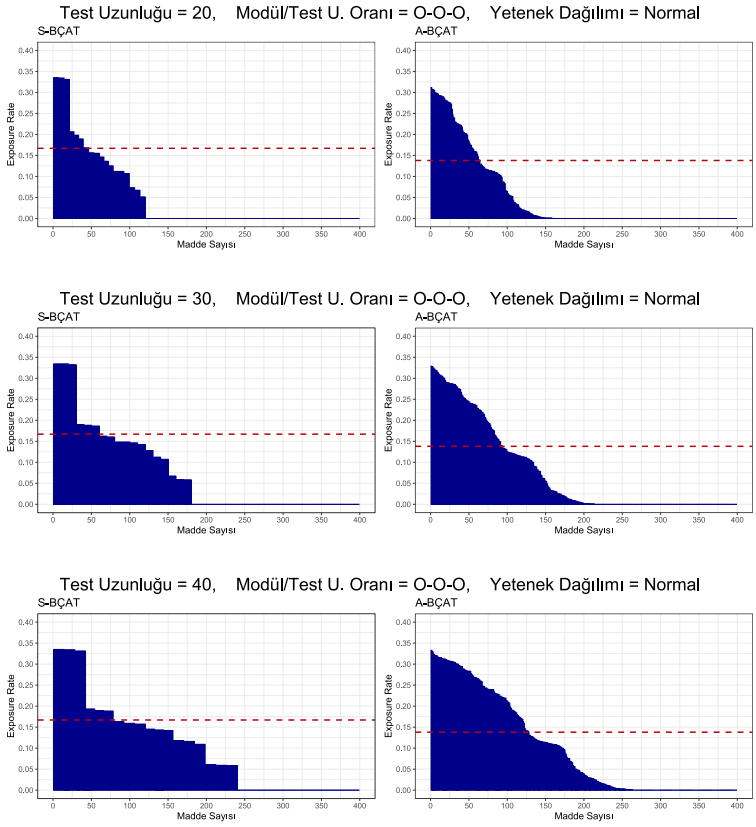
Şekil 25 incelendiğinde test uzunluğu arttıkça hem S-BÇAT hem de A-BÇAT yaklaşımlarında kullanılan madde sayısı artmaktadır. S-BÇAT ile A-BÇAT yaklaşımları arasındaki kullanılan madde sayısı farkı test uzunluğu arttıkça fark azalmaktadır. Bu bulguya benzer şekilde ortalama madde kullanım sıklığı S-BÇAT için tüm test uzunluklarında %16.8 iken, A-BÇAT'ta sırasıyla %11.7, %13.0 ve %14.2 olarak hesaplanmıştır. Bu durumda kısa test uzunluklarında A-BÇAT ile S-BÇAT arasındaki madde güvenliği A-BÇAT lehine daha iyiyken, test uzunluğu arttıkça her iki yaklaşımın madde güvenliği verimliliği birbirine yaklaşmıştır. Dolayısıyla A-BÇAT her üç test

uzunluğunda da S-BÇAT'a göre test ve madde güvenliği açısından verimlidir, bunun yanında kısa testlerde A-BÇAT'ın madde güvenliğine göre daha verimli olduğu söylenebilir.

Şekil 26'da test uzunluğu 20, 30 ve 40 olacak şekilde aynı yetenek dağılımı ve aynı modül/test uzunluğu oranına göre üç farklı koşulun madde bazında madde kullanım sıklığı grafikleri yer almaktadır.

Şekil 26

Test Uzunluğuna Göre Kullanılan Madde Bazında Madde Kullanım Sıklıkları Grafiği



Grafiklerde yer alan mavi sütunlar madde kullanım sıklığını göstermektedir. Maddeler, maksimum madde kullanım sıklığından minimum madde kullanım sıklığına doğru sıralanarak sütun grafikleri oluşturulmuştur. Grafiğin üzerindeki kırmızı kesikli çizgiler ilgili koşulun ortalama madde kullanım sıklığını göstermektedir.

Şekil 26 incelendiğinde tüm test uzunlukları koşullarında A-BÇAT yaklaşımının S-BÇAT yaklaşımına göre daha fazla sayıda madde kullandığı ve daha düşük ortalama madde kullanım sıklığına sahip olduğu görülmektedir. Diğer taraftan, S-BÇAT yaklaşımında

özellikle yönlendirme modülünde yer alan maddelerden dolayı bazı maddelerin madde kullanım sıklığı 0.35 düzeyinde olduğundan bu maddelerin açığa çıkmasının daha olası olduğu düşünülmektedir. A-BÇAT yaklaşımında ise bazı maddelerin madde kullanım sıklığı 0.35 düzeyinde iken ardından gelen maddelerin madde kullanım sıklıkları düşüşe geçmektedir. Ayrıca A-BÇAT yaklaşımında daha fazla maddenin kullanılmasının hem madde güvenliği hem de madde havuzundan daha efektif kullanılması adına faydalı olduğu söylenebilir.

Beşinci Araştırma Problemine İlişkin Bulgular

Beşinci alt araştırma problemi "Farklı yetenek dağılımlarında (normal dağılım, sağa çarpık, sola çarpık, uniform) S-BÇAT ve A-BÇAT yaklaşımlarının madde kullanım sıklığına ve kullanılan madde sayılarına göre madde güvenliği nasıl değişim göstermektedir?" şeklindeydi. Yetenek dağılımına göre Tablo 14 ve Tablo 15'teki ilgili hücrelerin ortalamalarından elde edilen sonuçlar Tablo 17'de sunulmuştur.

Tablo 17

Test Uzunluğuna Göre Kullanılan Madde Sayıları ve Ortalama Madde kullanım sıklıkları

Yetenek Dağılımı	Kullanılan Madde Sayıları		Ortalama Madde Kullanım Sıklıkları	
	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
Tekdüze	180	239	0,168	0,124

Tablo 17 incelendiğinde normal dağılımda ortalama kullanılan madde sayıları S-BÇAT için 180 iken, A-BÇAT için 227'dir. Bu bulguya benzer şekilde ortalama madde kullanım sıklığı S-BÇAT için 0.168 iken A-BÇAT için 0.130'dur. Normal dağılımda A-BÇAT yaklaşımının hem kullanılan madde sayısı hem de ortalama madde kullanım sıklıklarına göre madde güvenliği açısından daha iyi sonuçlar sunduğu ifade edilebilir.

Sağa çarpık dağılımda ortalama kullanılan madde sayıları S-BÇAT için 180 iken, A-BÇAT için 226'dır. Benzer şekilde ortalama madde kullanım sıklığı S-BÇAT için 0.168 iken A-BÇAT için 0.131'dir. Sağa çarpık dağılımda A-BÇAT yaklaşımının hem kullanılan madde sayısı hem de ortalama madde kullanım sıklıklarına göre madde güvenliği açısından daha iyi sonuçlar sunduğu belirtilebilir.

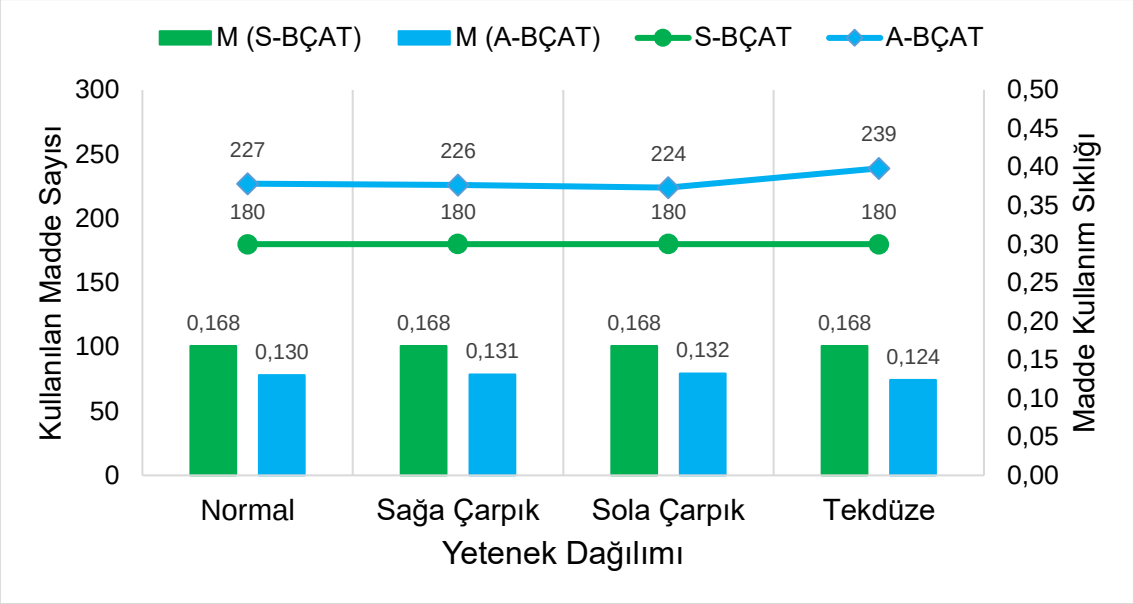
Sola çarpık dağılımda ortalama kullanılan madde sayıları S-BÇAT için 180 iken, A-BÇAT için 224'tür. Benzer şekilde ortalama madde kullanım sıklığı S-BÇAT için 0.168 iken A-BÇAT için 0.132'dir. Sola çarpık dağılımda A-BÇAT yaklaşımının hem kullanılan madde sayısı hem de ortalama madde kullanım sıklıklarına göre madde güvenliği açısından daha iyi sonuçlar sunduğu söylenebilir.

Tekdüze dağılımda ise ortalama kullanılan madde sayıları S-BÇAT için 240 iken, A-BÇAT için 239'dur. Benzer şekilde ortalama madde kullanım sıklığı S-BÇAT için 0.168 iken A-BÇAT için 0.124'dür. Tekdüze dağılımda da A-BÇAT yaklaşımının S-BÇAT'tan hem kullanılan madde sayısı hem de ortalama madde kullanım sıklıklarına göre madde güvenliği açısından daha iyi sonuçlar sunmaktadır.

Tablo 17'de yer alan bulgular, farklı yetenek dağılımlarını karşılaştırmak amacıyla Şekil 27'deki grafik üzerinde görselleştirilmiştir.

Şekil 27

Yetenek Dağılımına Göre Kullanılan Madde Sayıları ve Ortalama Madde Kullanım Sıklıkları

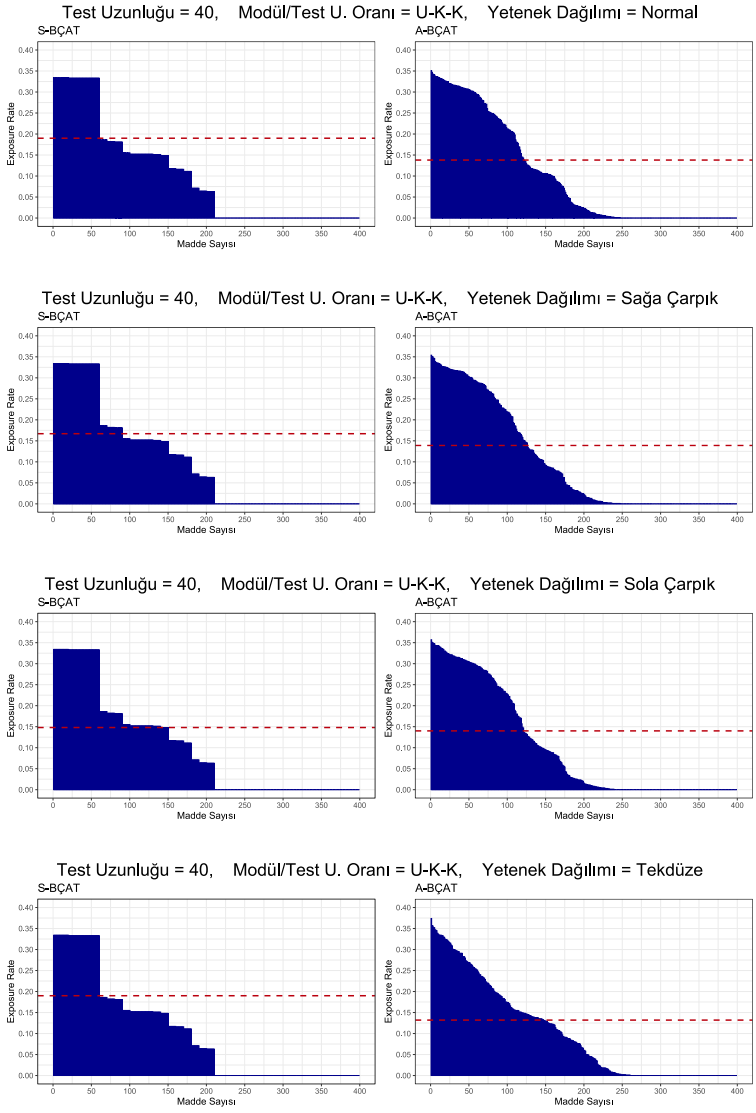


Şekil 27 incelendiğinde normal, sağa çarpık, sola çarpık ve tekdüze dağılımlarda A-BÇAT yaklaşımının S-BÇAT yaklaşımına göre kullanılan madde sayısının daha fazla olduğu görülmektedir. Bu bulguyu destekleyecek şekilde, grafiğin alt kısmında yer alan sütun grafiklerinde görüldüğü üzere tüm farklı yetenek dağılımlarında A-BÇAT yaklaşımının ortalama kullanım sıklığı S-BÇAT'tan daha düşüktür. Bunun yanında, ilk üç yetenek dağılımı için S-BÇAT ile A-BÇAT yaklaşımları arasındaki kullanılan madde sayısı farkı benzer iken, tekdüze dağılımda A-BÇAT yaklaşımının S-BÇAT yaklaşımından daha fazla kullanılan madde sayısına ve daha düşük madde kullanım sıklığına sahip olduğu görülmektedir. Dolayısıyla tekdüze dağılımlarda madde güvenliği açısından A-BÇAT yaklaşımının S-BÇAT'a göre daha iyi sonuçlar ürettiği belirtilebilir.

Şekil 28'de test uzunluğu 40 ve modül/test uzunluğu oranı U-K-K olacak şekilde dört farklı yetenek dağılımına (normal, sağa çarpık, sola çarpık ve tekdüze) göre dört farklı koşulun madde bazında madde kullanım sıklığı grafikleri yer almaktadır.

Şekil 28

Yetenek Dağılımına Göre Kullanılan Madde Bazında Madde Kullanım Sıklıkları Grafiği



Şekil 28 incelendiğinde tüm farklı yetenek dağılımları koşullarında A-BÇAT yaklaşımının S-BÇAT yaklaşımına göre daha fazla sayıda madde kullandığı ve daha düşük ortalama madde kullanım sıklığına (kırmızı kesikli çizgiler) sahip olduğu görülmektedir. Diğer taraftan, S-BÇAT yaklaşımında özellikle yönlendirme modülünde yer alan maddelerden dolayı bazı maddelerin madde kullanım sıklığı 0.35 düzeyindedir. A-BÇAT'ta ise bazı maddelerin madde kullanım sıklığı 0.35 düzeyinde iken ardından gelen maddelerin madde kullanım sıklıkları düşüşe geçmektedir. Dolayısıyla hem düşük madde kullanım

sıklığı hem de yüksek kullanılan madde sayılarına göre tüm yetenek dağılımlarında A-BÇAT yaklaşımının S-BÇAT'a göre daha iyi madde güvenliği sonuçları sunduğu söylenebilir.

Altıncı Araştırma Problemine İlişkin Bulgular

Altıncı alt araştırma problemi "Farklı modül/test uzunluğu oranlarında (U-K-K, O-O-O, K-K-U) S-BÇAT ve A-BÇAT yaklaşımlarının madde kullanım sıklığına ve kullanılan madde sayılarına göre madde güvenliği nasıl değişim göstermektedir?" şeklindeydi. Yetenek dağılımına göre Tablo 14 ve Tablo 15'deki ilgili hücrelerin ortalamalarından elde edilen sonuçlar Tablo 18'de sunulmuştur.

Tablo 18

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

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

U: Uzun, O: Orta, K: Kısa.

Tablo 17 incelendiğinde U-K-K modül/test uzunluğu oranında ortalama kullanılan madde sayıları S-BÇAT için 157 iken, A-BÇAT için 219'dur. Bu bulguya benzer şekilde ortalama madde kullanım sıklığı S-BÇAT için 0.174 iken A-BÇAT için 0.135'tir. U-K-K modül/test uzunluğu oranında A-BÇAT yaklaşımının hem kullanılan madde sayısı hem de ortalama madde kullanım sıklıklarına göre madde güvenliği açısından daha iyi sonuçlar sunduğu belirtilebilir.

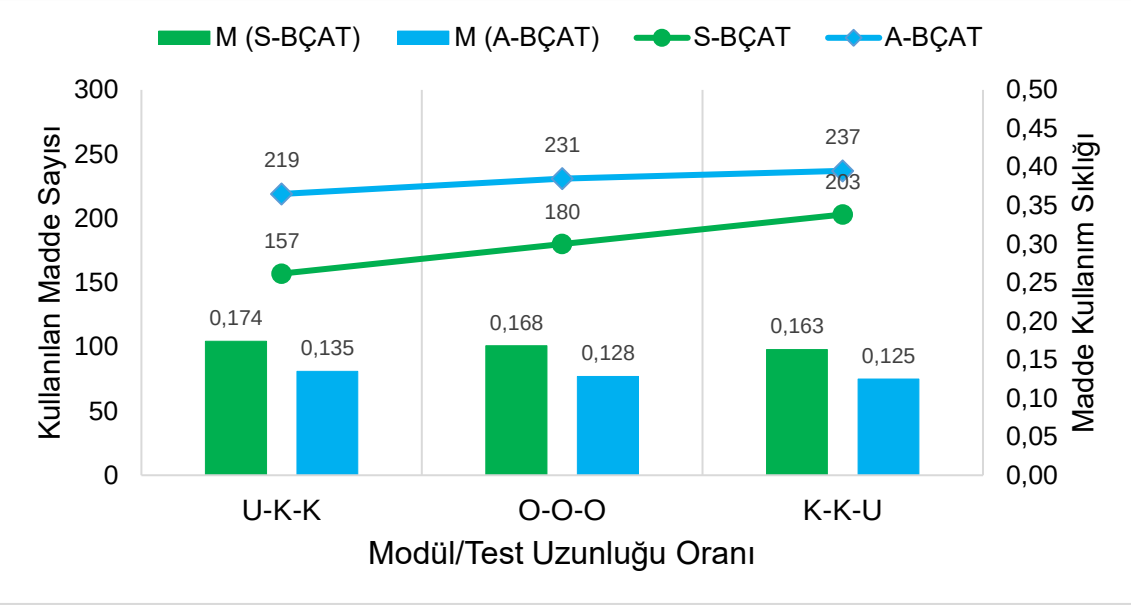
O-O-O modül/test uzunluğu oranında ortalama kullanılan madde sayıları S-BÇAT için 180 iken, A-BÇAT için 231'dir. Benzer şekilde ortalama madde kullanım sıklığı S-BÇAT için 0.168 iken A-BÇAT için 0.128'dir. O-O-O modül/test uzunluğu oranında A-BÇAT yaklaşımının hem kullanılan madde sayısı hem de ortalama madde kullanım sıklıklarına göre madde güvenliği açısından daha iyi sonuçlar sunduğu söylenebilir.

K-K-U modül/test uzunluğu oranında ise ortalama kullanılan madde sayıları S-BÇAT için 203 iken, A-BÇAT için 237'dir. Benzer şekilde ortalama madde kullanım sıklığı S-BÇAT için 0.163 iken A-BÇAT için 0.125'dir. K-K-U modül/test uzunluğu oranında da A-BÇAT yaklaşımının hem kullanılan madde sayısı hem de ortalama madde kullanım sıklıklarına göre madde güvenliği açısından daha iyi sonuçlar sunduğu ifade edilebilir.

Tablo 18'de yer alan bulgular üzerinde modül/test uzunlukları arasında karşılaştırma yapmak amacıyla Şekil 29'da grafik üzerinde görselleştirilmiştir.

Şekil 29

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



Not: U: Uzun, O: Orta, K: Kısa.

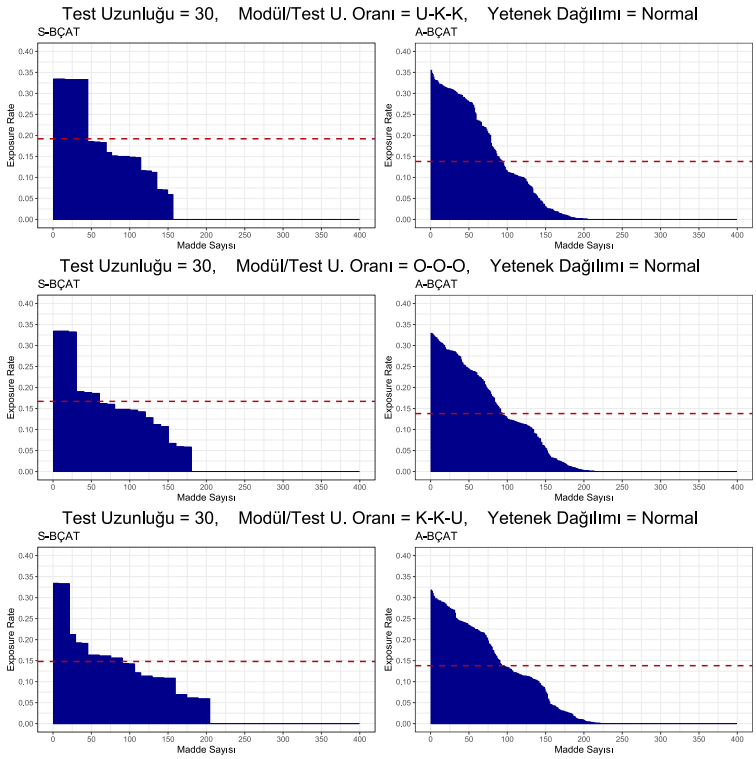
Şekil 29 incelendiğinde tüm modül/test uzunluğu oranlarında (U-K-K, O-O-O ve K-K-U) A-BÇAT yaklaşımının S-BÇAT yaklaşımına göre kullanılan madde sayısının daha fazla olduğu görülmektedir. Bu bulguyu destekleyecek şekilde, grafiğin alt kısmında yer alan sütun grafiklerinde görüldüğü üzere tüm farklı modül/test uzunluğu oranlarında A-BÇAT yaklaşımının ortalama madde kullanım sıklığı S-BÇAT'tan daha düşüktür. Bunun yanında, U-K-K oranında yönlendirme modülünde yer alan madde sayısı daha fazla olduğundan S-

BÇAT ile A-BÇAT'ın kullanılan madde sayısı arasındaki fark daha fazla iken, K-K-U oranına doğru gidildikçe kullanılan madde sayısı arasındaki fark azalmaktadır. K-K-U oranında iki yaklaşım arasındaki kullanılan madde sayısı farkı en aza inmiştir.

Şekil 30'da test uzunluğu 30 ve yetenek dağılımı normal dağılım olacak şekilde üç farklı modül/test uzunluğu oranına (U-K-K, O-O-O ve K-K-U) göre üç farklı koşulun madde bazında madde kullanım sıklığı grafikleri sunulmuştur.

Şekil 30

Modül/Test Uzunluğu Oranına Göre Kullanılan Madde Bazında Madde Kullanım Sıklıkları Grafiği



Not: U: Uzun, O: Orta, K: Kısa.

Şekil 30 incelendiğinde tüm farklı modül/test oranı uzunlukları koşullarında A-BÇAT yaklaşımının S-BÇAT yaklaşımına göre daha fazla sayıda madde kullandığı ve daha düşük ortalama madde kullanım Sıklığına (kırmızı kesikli çizgiler) sahip olduğu görülmektedir. Diğer taraftan, S-BÇAT yaklaşımında özellikle yönlendirme modülünde yer alan maddelerden dolayı bazı maddelerin madde kullanım sıklığı 0.35 düzeyindedir. A-BÇAT'ta ise bazı maddelerin madde kullanım sıklığı 0.35 düzeyinde iken ardından gelen maddelerin

madde kullanım sıklıkları düşüşe geçmektedir. Dolayısıyla hem düşük madde kullanım sıklığı hem de yüksek kullanılan madde sayılarına göre tüm yetenek dağılımlarında A-BÇAT yaklaşımının S-BÇAT'a göre daha iyi madde güvenliği sonuçları sunduğu söylenebilir. Diğer taraftan, U-K-K oranında yönlendirme modülünün uzunluğu daha fazla olduğundan S-BÇAT yaklaşımında yönlendirme modülünde yer alan maddelerin madde kullanım sıklığı daha yüksektir. Bu durum madde güvenliği açısından sorunlar oluşturabilir. K-K-U oranında ise yönlendirme modülü uzunluğu daha kısa olduğundan madde kullanım sıklıkları daha düşüktür. Fakat K-K-U oranının dezavantajı ise diğer iki modül/test uzunluğu oranına göre düşük ölçme kesinliğine sahip olmasıdır. Dolayısıyla modül/test uzunluğu oranına karar verirken hem madde güvenliği hem de ölçme kesinliği dikkate alınarak en uygun noktanın belirlenmesi gereklidir.

Bölüm 5

Sonuç, Tartışma ve Öneriler

Bu başlık altında araştırmadan elde edilen bulguların doğrultusunda sonuç ve tartışma bölümleri yer almaktadır.

Sonuç

1) Araştırma kapsamında S-BÇAT ve A-BÇAT yaklaşımları farklı koşullar altında karşılaştırılmıştır. İncelenen tüm koşullarda A-BÇAT yaklaşımının S-BÇAT'tan daha yüksek ölçme kesinliği sunduğu sonucuna varılmıştır. Ayrıca Cohen d katsayısı ile etki büyüklükleri incelenmiştir. Tüm koşullar altında A-BÇAT yaklaşımının S-BÇAT'a göre ölçme kesinliği açısından büyük etki gösterdiği sonucuna ulaşılmıştır. Bir diğer deyişle, A-BÇAT yaklaşımı S-BÇAT'a göre yetenek düzeyi açısından daha etkili kestirimler yapmaktadır.
2) Araştırmada iki farklı BÇAT yaklaşımının karşılaştırılmasında üç farklı test uzunluğu kullanılmıştır. Tüm test uzunluğu düzeylerinde A-BÇAT yaklaşımının S-BÇAT'a göre daha yüksek ölçme kesinliği sunduğu görülmüştür. Ayrıca tüm test uzunluklarında A-BÇAT yaklaşımının S-BÇAT'a göre büyük etki gösterdiği sonucuna ulaşılmıştır. Diğer taraftan, kısa test uzunluklarında A-BÇAT ile S-BÇAT arasındaki ölçme kesinliği farkı uzun test uzunluklarına göre daha fazladır. Bir diğer deyişle, özellikle kısa testlerde A-BÇAT'ın daha etkili yetenek kestirimi sunduğu sonucuna ulaşılmıştır.
3) Araştırmada iki BÇAT yaklaşımı dört farklı yetenek dağılımı altında karşılaştırılmıştır. Tüm yetenek dağılımlarında A-BÇAT yaklaşımının S-BÇAT'a göre daha yüksek ölçme kesinliği sunduğu sonucuna ulaşılmıştır. Ayrıca tüm yetenek dağılımlarında A-BÇAT yaklaşımının S-BÇAT'a göre büyük etki gösterdiği görülmektedir. Yetenek dağılımlarına göre hem ölçme kesinliği farkları hem de etki büyüklükleri incelendiğinde normal dağılımdaki farkın sağa çarpık, sola çarpık ve

tekdüze dağılımlara göre daha az olduğu görülmektedir. Bu durum S-BÇAT'ın normal dağılımda daha iyi yetenek kestirimi yaptığını belirtirken, sağa çarpık, sola çarpık ve tekdüze dağılımlarda S-BÇAT daha kötü ölçme kesinliği sunmaktadır. Sonuç olarak, A-BÇAT yaklaşımı değişen yetenek dağılımlarından daha az etkilenirken, S-BÇAT yaklaşımı daha fazla etkilenmektedir.

4) Araştırmada iki BÇAT yaklaşımı üç farklı modül/test uzunluğu oranı altında karşılaştırılmıştır. Tüm modül/test uzunluğu oranlarında A-BÇAT yaklaşımının S-BÇAT yaklaşımına göre daha yüksek ölçme kesinliği sunduğu sonucuna ulaşılmıştır. Ayrıca tüm modül/test uzunluğu oranlarında A-BÇAT'ın S-BÇAT'a göre ölçme kesinliği açısından büyük etki gösterdiği sonucuna ulaşılmıştır. Yönlendirme modülünün uzun olduğu U-K-K oranında A-BÇAT ile S-BÇAT arasındaki ölçme kesinliği farkı ve etki büyüklüğü değerleri diğer iki orana göre daha az olduğu görülmektedir. U-K-K oranını sırasıyla O-O-O ve K-K-U oranları izlemektedir. Özellikle K-K-U oranında S-BÇAT tasarımında son modüldeki madde sayısının fazla olmasından dolayı S-BÇAT'ın ölçme kesinliği azalmıştır. Fakat bu durumla beraber A-BÇAT ise K-K-U oranında en etkili yetenek kestirimini gerçekleştirmiştir. Sonuçlar, son modüldeki madde sayısı arttıkça A-BÇAT'ın ölçme kesinliği artarken, S-BÇAT'ın ölçme kesinliğinin azaldığımı göstermektedir.

5) Araştırma kapsamında S-BÇAT ve A-BÇAT yaklaşımlarının karşılaştırıldığı 36 farklı koşulun tümünde A-BÇAT yaklaşımının S-BÇAT'tan daha yüksek kullanılan madde sayısına ve daha düşük madde kullanım sıklığına sahip olduğu sonucuna ulaşılmıştır. Kullanılan madde sayısının daha yüksek olması A-BÇAT yaklaşımının S-BÇAT'a göre madde havuzundan daha iyi düzeyde faydalandığını göstermektedir. Benzer şekilde A-BÇAT yaklaşımının S-BÇAT'a göre madde kullanım sıklıklarının daha düşük olması maddelerin açığa çıkma olasılığını azaltmasından dolayı madde güvenliğinin daha iyi olduğunun bir göstergesi olarak yorumlanabilir. Ayrıca tüm koşullar altında A-BÇAT'ın minimum ve maksimum

madde kullanım sıklığının S-BÇAT'a göre daha düşük olması madde güvenliği açısından A-BÇAT'ın daha verimli olduğunun bir göstergesi olarak kabul edilebilir.

6) A-BÇAT yaklaşımının S-BÇAT yaklaşımına göre tüm test uzunluklarına göre daha yüksek kullanılan madde sayıları ve daha düşük madde kullanım sıklıklarına sahip olduğu sonucuna ulaşılmıştır. Bu durum tüm test uzunluklarında A-BÇAT'ın S-BÇAT'a göre madde güvenliği açısından daha verimli sonuçlar ürettiğini göstermektedir.
7) Farklı yetenek dağılımlarında A-BÇAT'ın S-BÇAT'a göre daha yüksek kullanılan madde sayısı ve daha düşük madde kullanım sıklığına sahip olduğu sonucuna ulaşılmıştır. Dolayısıyla madde güvenliği açısından A-BÇAT'ın S-BÇAT'a göre daha iyi sonuçlar verdiği ifade edilebilir. Ayrıca normal, sağa çarpık ve sola çarpık dağılımlarda A-BÇAT ile S-BÇAT yaklaşımları arasındaki kullanılan madde sayısı ve ortalama madde kullanım sıklığı farkı benzerlik gösterirken, tekdüze dağılımlarda iki yaklaşımın arasındaki farkın A-BÇAT lehine arttığı görülmüştür. Bu durum tekdüze dağılımlarda yer alan uç yetenek düzeylerindeki katılımcı sayısının artmasından kaynaklanmaktadır. Sonuç olarak, A-BÇAT yaklaşımının S-BÇAT'a göre tekdüze dağılımlarda madde güvenliği açısından daha etkili olduğu belirtilebilir.
8) Tüm modül/test uzunluğu oranlarında A-BÇAT'ın S-BÇAT'a göre daha yüksek kullanılan madde sayısı ve daha düşük madde kullanım sıklığına sahip olduğu görülmektedir. Bu durum madde güvenliği açısından farklı modül/test uzunluğu oranlarında A-BÇAT'ın S-BÇAT'a göre daha iyi sonuçlar verdiğini göstermektedir. Ayrıca S-BÇAT yaklaşımında U-K-K oranında yönlendirme modülü daha uzun olduğundan kullanılan madde sayısı daha düşük, madde kullanım sıklığı ise daha yüksektir. K-K-U oranında ise üçüncü aşamadaki modüller daha uzun olduğundan kullanılan madde sayısı daha yüksek, madde kullanım sıklığı ise daha düşüktür. Bu durumda U-K-K oranının K-K-U oranına göre madde güvenliği açısından daha verimli olduğu söylenebilir.

Tartışma

A-BÇAT yaklaşımının test uygulama süreçlerinde olduğu gibi çok sayıda ve karmaşık test özellikleri ve kısıtlamaların yer aldığı durumlarda başarılı çözümler sunan bir test yaklaşımı olduğu belirtilmektedir (Choi, van der Linden, 2018; van der Linden, 2021). Bu çalışmanın sonuçları, A-BÇAT'ın S-BÇAT'tan daha iyi ölçme kesinliği sunduğunu göstermektedir. Zheng ve Chang (2015), BBT, A-BÇAT ve S-BÇAT'ı karşılaştırdıkları simülasyon çalışmalarında BBT ve A-BÇAT'ın oldukça benzer ölçme kesinliği sunduğunu belirtirken, S-BÇAT'ın daha düşük ölçme kesinliği sonuçları ürettiğini belirtmektedirler. Tay (2015) A-BÇAT ve S-BÇAT'ın sınıflama doğruluğu ve sınıflama tutarlılığını karşılaştırdığı doktora tezinde tüm koşullarda A-BÇAT'ın S-BÇAT'a göre daha iyi sonuçlar sunduğunu belirtmektedir. van der Linden ve Diao (2016), BBT, A-BÇAT, S-BÇAT ve DT'yi karşılaştırdıkları çalışmalarında BBT ve A-BÇAT'ın ölçme kesinliği açısından en iyi sonuçlar sunduğunu, bu yaklaşımları S-BÇAT'ın izlediğini, DT'nin ise son sırada yer aldığını bildirmektedir. Han ve Guo (2016), BBT ile A-BÇAT'ın oldukça yakın ölçme kesinliği sunduğunu belirtirken, S-BÇAT'ın ise bu iki yaklaşımdan daha az ölçme kesinliği sunduğunu belirtmektedir. Van der Linden (2021) ise DT, BBT, A-BÇAT ve S-BÇAT yaklaşımlarını karşılaştırdıkları çalışmasında A-BÇAT'ın S-BÇAT'tan ölçme kesinliği açısından daha başarılı olduğunu ifade etmektedir. Literatürde yer alan ve A-BÇAT ile S-BÇAT'ı karşılaştıran çalışmalar ile bu çalışmanın sonuçların oldukça benzer olduğu görülmektedir (Han & Guo, 2016; Tay, 2015; van der Linden & Diao, 2016; van der Linden, 2021; Zheng & Chang, 2015;).

Bu çalışmada tüm test uzunlukları koşullarında A-BÇAT'ın S-BÇAT'tan daha iyi ölçme kesinliği sunduğu sonucuna ulaşılmıştır. Diğer taraftan 20 test uzunluğunda A-BÇAT yaklaşımı S-BÇAT'tan daha etkili iken, test uzunluğu arttıkça iki yaklaşımın ölçme kesinlikleri birbirine yaklaşmıştır. Diğer taraftan Yasuda, Mae, Hull ve Taniguchi (2021), BBT'nin test uzunluğu üzerine yaptığı çalışmada test uzunluğu arttıkça ölçme kesinliği ve yanlılığın ciddi düzeyde azalmadığı sonucuna ulaşmışlardır. Dolayısıyla bu çalışmada da

test uzunluğu arttıkça her iki yaklaşımın ölçme kesinliklerinin test uzunluğuyla orantılı şekilde azalmadığı görülmektedir. Bu durumu van der Linden (2021) azalan verimler kanunu ile açıklamaktadır. Ayrıca uzun test uzunluğuna sahip testlerin bireyselleştirilmiş test yaklaşımının mantığı ile çeliştiği ifade edilmektedir (van der Linden, 2021).

Araştırmada yer alan dört farklı yetenek dağılımında da A-BÇAT yaklaşımının S-BÇAT'a göre daha iyi ölçme kesinliği sunduğu sonucuna ulaşılmıştır. Her iki yaklaşımda da ölçme kesinliği en yüksek normal dağılımda iken, normal dağılımı sağa çarpık, sola çarpık dağılım izlemektedir. Tekdüze dağılımda ise ölçme kesinliği her iki yaklaşım için de düşüktür. Diğer taraftan, eğitimde hem ölçülen değişkenin yapısı hem de evrenin dağılımı düşünüldüğünde normal dağılmayan örneklemlerle sıklıkla karşılaştılmaktadır (MEB, 2021). Bu noktada A-BÇAT'ın özellikle sağa ve sola çarpık dağılımlarda S-BÇAT'a göre ölçme kesinliğinin oldukça iyi olduğu sonucuna ulaşılmıştır.

Modül-test uzunluğu oranına göre tüm oranlarda A-BÇAT'ın S-BÇAT'a göre daha yüksek ölçme kesinliği sunduğu görülmüştür. Yönlendirme modülünün uzunluğu fazla olduğunda S-BÇAT'ın daha iyi ölçme kesinliği sunduğunu belirten çalışmalar bulunmaktadır (Boztunç, 2019; Cai, Anthony, Albano & Roussos, 2021; Kim & Plake, 1993; Zheng, 2016). Bu çalışmada da literatürde yer alan çalışmalara benzer şekilde yönlendirme modülünün uzunluğu arttıkça S-BÇAT daha iyi ölçme kesinliği sonuçları vermektedir. Yine literatürdeki sonuçlara benzer şekilde, ikinci ve üçüncü aşamanın modül uzunlukları arttıkça S-BÇAT'ın ölçme kesinliği bu tasarımlarda fazla sayıda kullanılan madde ihtiyacı olmasından dolayı azalmaktadır. A-BÇAT ise modül/test uzunluğu oranından daha az etkilenmekte olup, son aşamadaki madde sayısının artması durumunda diğer modül/test uzunluğu oranlarına göre daha iyi ölçme kesinliği sunmaktadır. Sonuç olarak, S-BÇAT'ta ilk modül uzunluğu arttığında ölçme kesinliği artarken, A-BÇAT'ta son modülün uzunluğu arttığında ölçme kesinliği artmaktadır.

Madde güvenliği açısından tüm koşullarda A-BÇAT'ın S-BÇAT'tan daha düşük madde kullanım sıklığı ve daha yüksek kullanılan madde sayısına sahip olduğu sonucuna

ulaşılmıştır. Diğer taraftan, A-BÇAT'ta madde kullanım sıklıkları belirli bir ivme ile azalmakta iken, S-BÇAT'ta yönlendirme modülünde yer alan maddelerin madde kullanım sıklıkları yüksektir, ikinci ve üçüncü aşamalarda madde kullanım sıklıkları azalmaktadır. Zheng ve Chang (2015), bu araştırmanın sonuçlarına benzer şekilde A-BÇAT'ın test örtüşme oranlarının daha düşük olduğunu belirtmekte ve S-BÇAT'ta katılımcıların benzer rotaları izlemesinin madde güvenliği sorunları oluşturabileceğini bildirmektedirler. S-BÇAT'ta panel ve modüller sabittir, dolayısıyla katılımcının izleyeceği yolda yer alan modüllerdeki maddeler test yöneticisi tarafından bilinmektedir. A-BÇAT'ın her katılımcıya kendi yetenek düzeyinde test birleştirmesi sebebiyle katılımcıya uygulanacak maddeler test yöneticisi dahil kimse tarafından daha önceden bilinmemektedir ve belirli değildir. Dolayısıyla A-BÇAT kişiye özel benzersiz testler oluşturmaktadır. Bu bağlamda A-BÇAT madde ve test güvenliği açısından S-BÇAT'a göre oldukça avantajlı bir konumda yer almaktadır.

Bu araştırmanın sonuçları, A-BÇAT yaklaşımının S-BÇAT yaklaşımından hem ölçme kesinliği hem de madde ve test güvenliği açılarından daha avantajlı olduğunu göstermektedir. Diğer taraftan S-BÇAT yaklaşımının A-BÇAT yaklaşımına göre iki önemli avantajı bulunmaktadır: (I) sabit panel ve modül yapısının sınav katılımcılarına kolay açıklanabilirliği ve (II) test uygulamasının kolay yönetilebilirliği (Yan, Von Davier ve Lewis, 2016). Bu iki avantajlı durum S-BÇAT'ın tercih edilmesinin en önemli gerekçelerindendir. Sınavın yapısının ve puanlanma yönteminin sınav katılımcıları ve diğer paydaşlar tarafından kolayca anlaşılabilir olması, sınavın adil bir şekilde yürütüldüğünün hissettirilmesi noktasında önemlidir. Dolayısıyla uyarlanabilir test yaklaşımı tercihinde ölçme kesinliğinin ve madde-test güvenliğinin yanında yapılacak sınavın risk durumunun, değerlendirme türünün (norm veya kriter dayanaklı), test uygulamasının katılımcılara açıklanabilirliğinin ve toplumun sosyolojik olarak ilgili sınava bakış açısının irdelenerek karar verilmesi gerektiği düşünülmektedir.

Öneriler

Bu başlık altında uygulayıcılara ve araştırmacılara yönelik öneriler yer almaktadır.

Uygulayıcılara Yönelik Öneriler

1) Daha etkili yetenek kestirimleri sunmasından dolayı kısa test uzunluklarında S-BÇAT yerine A-BÇAT yönteminin tercih edilmesi önerilmektedir.
2) A-BÇAT tasarımı tercih edilecekse, ölçme kesinliğini iyileştireceğinden dolayı son modül uzunluğunun arttırılması önerilmektedir.
3) BÇAT tasarımının son aşamasındaki madde sayısının fazla olması isteniyorsa S-BÇAT yerine A-BÇAT yaklaşımının tercih edilmesi önerilmektedir.
4) Yönlendirme modülünde yer alacak madde sayısının fazla olması isteniyorsa S-BÇAT yöntemi ile A-BÇAT yönteminin etkililik düzeyinin birbirine yaklaştığından S-BÇAT yönteminin de tercih edilebilir olduğu ifade edilebilir.
5) Uygulanacak sınavın kesme puanı yetenek ölçeğinin orta noktalarında ($\vartheta = 0$ ciarında) ise A-BÇAT ve S-BÇAT yöntemleri oldukça benzer ölçme kesinliği sunmaktadır. Fakat kesme puanı uç yetenek düzeylerinde ise A-BÇAT yönteminin tercih edilmesi önerilmektedir.
6) Test uygulamasında madde ve test güvenliği sorunu yaşanacağı düşünülüyorsa ise S-BÇAT yerine A-BÇAT yönteminin tercih edilmesi önerilmektedir.

Araştırmacılara Yönelik Öneriler

1) Bu araştırmada TIMSS parametre dağılımlarına göre simülasyon ortamında madde havuzu üretilmiştir. Gerçek veri havuzu ile benzer bir çalışma tasarlanabilir.
2) Bu çalışmada S-BÇAT panel tasarımı olarak “1-2-3” tasarımı kullanılmıştır. “1-3”, “1-4”, “1-2-4” gibi farklı tasarımlar ile benzer araştırmalar gerçekleştirilebilir.
3) Bu çalışmada yetenek kestirimi yöntemi olarak EAP yöntemi tercih edilmiştir. Yetenek kestirimi yöntemi olarak MLE, MLEF, MAP yöntemleri kullanılarak farklı araştırmalar tasarlanabilir.

4) Bu çalışmada yalnızca dört farklı kapsama göre kısıtlama eklenerek test birleştirme işlemleri gerçekleştirilmiştir. Gerçek madde havuzları kullanarak ortak köklü maddeler, düşman maddeler, sözcük sayısı, ortalama süre gibi farklı test özellikleri ve kısıtlamalar altında benzer çalışmalar gerçekleştirilebilir. Ayrıca kısıtlama sayısı arttırılarak BBT ve A-BÇAT yaklaşımlarının etkililiği karşılaştırılabilir.
5) S-BÇAT'ta yönlendirme metodu olarak MFI yöntemi kullanılmıştır. Yönlendirme metodu olarak AMI, doğru sayısı ve diğer yöntemler kullanılarak farklı araştırmalar tasarlabilir.
6) Bu araştırmada A-BÇAT ile S-BÇAT karşılaştırılmıştır. Farklı S-BÇAT tasarımları ile hibrit BBT yaklaşımının karşılaştırıldığı araştırmalar tasarlanabilir.
7) Bu çalışmada test uzunluğu olarak 20, 30 ve 40 test uzunluğu ele alınmıştır. Farklı test uzunlukları altında benzer çalışmalar gerçekleştirilebilir. BBT yaklaşımları için en uygun (optimal) test uzunluğunun belirlenmesi için farklı çalışmalar gerçekleştirilebilir.
8) Bu çalışmada U-K-K, O-O-O ve K-K-U modül/test uzunluğu oranları ele alınmıştır. Farklı modül/test uzunluğu oranları altında farklı çalışmalar yapılabilir.
9) Bu çalışmada madde havuzu büyüklüğü 400 olarak belirlenmiştir. Madde havuzu büyüklüğü değişimlenerek benzer çalışmalar tasarlanabilir.
10) Bu çalışmada ölçme kesinliği ve madde güvenliği araştırılmıştır. Gelecek çalışmalarda A-BÇAT ile S-BÇAT'ın sınıflama doğruluğu karşılaştırılarak ele alınabilir.

Kaynaklar

Alkan, V. (2021). Mesleki alan ilgi envanteri'nin bilgisayar ortamında bireye uyarlanmış formunun geliştirilmesi. Doktora Tezi. Ankara Üniversitesi.

Andrew, D. (2014). The effects of routing and scoring within a computer adaptive multi-stage framework. (Unpublished Doctoral Dissertation). The University of North Carolina.

Arikan, S., Özer, F., Şeker, V., & Ertaş, G. (2020). The importance of sample weights and plausible values in large-scale assessments. *Egitimde ve Psikolojide Olcme ve Degerlendirme Dergisi*, 522–539. https://doi.org/10.21031/epod.602765

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

Armstrong, R., & Edmonds, J. (2004, March). A study of multiple stage adaptive test designs. In annual meeting of National Council of Measurement in Education,(NCME), San Diego, CA.

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

Aybek, E. C., & Çıkrıkçı, R. N. (2018). Kendini değerlendirme envanteri'nin bilgisayar ortamında bireye uyarlanmış test olarak uygulanabilirliği. *Turkish Psychological Counseling and Guidance Journal*, *8*(50), 117-141.

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

Balta, E., & Uçar, A. (2022). Bilgisayar Ortamında Bireye Uyarlanmış Test Uygulamalarında Ölçme Kesinliğinin ve Test Uzunluğunun Farklı Koşullar Altında İncelenmesi. *E-International Journal of Educational Research*. https://doi.org/10.19160/e-ijer.1023098

Baykul, Y. (2015). Eğitimde ve Psikolojide Ölçme Klasik Test Teorisi ve Uygulaması. *Pegem Yayınları*, Ankara.

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.

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

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

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.

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.

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

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

Bulut, O., & Sünbül, Ö. (2017). Monte Carlo Simulation Studies in Item Response Theory with the R Programming Language R Programlama Dili ile Madde Tepki Kuramında Monte

Carlo Simülasyon Çalışmaları. *Journal of Measurement and Evaluation in Education and Psychology*, 8(3), 266-287.

Buyukozturk, S., Kilic Cakmak, E., Akgun, O. E., Karadeniz, S., & Demirel, F. (2013). Bilimsel araştırma yöntemleri. *Ankara: Pegem Yayıncılık*.

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.

Chen, D.-S., Batson, R. G., & Dang, Y. 2010. Applied Integer Programming. New York, NY:Wiley.

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.

Cor, K., Alves, C., & Gierl, M. J. (2008). Computer Software Review: Conducting Automated Test Assembly Using the Premium Solver Platform Version 7.0 With Microsoft Excel and the Large-Scale LP/QP Solver Engine Add-In. Applied Psychological Measurement, 32(8), 652-663, https://doi.org/10.1177/0146621608316603.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

Crotts, K., Sireci, S. G., & Zenisky, A. (2012). Evaluating the content validity of multistage-adaptive tests. Journal of Applied Testing Technology, 13(1).

Çoban, E. (2020). Bilgisayar temelli bireyselleştirilmiş test yaklaşımlarının Türkiye'deki merkezi dil sınavlarında uygulanabilirliğinin araştırılması. Doktora Tezi. Ankara Üniversitesi

Dallas, A., Wang, X., Furter, R., & Luecht, R. M. (2012). Item pool size, targeted item writing, and panel replication strategies for a 1-3-3 multistage test design. In Annual Meeting of the National Council on Measurement in Education (pp. 1-23).

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.

DeMars, C. (2010). Item response theory. Oxford University Press.

Demir, S. (2022). The effect of item pool and selection algorithms on computerized classification testing (CCT) performance. *Journal of Educational Technology and Online Learning, 5*(3), 573-584. https://doi.org/10.31681/jetol.1099580

Demir, S., & Atar, B. (2021). Investigation of classification accuracy, test length and measurement precision at computerized adaptive classification tests. *Egitimde ve Psikolojide Olcme ve Degerlendirme Dergisi*. https://doi.org/10.21031/epod.787865

Diao, Q., & van der Linden, W. J. (2011). Automated test assembly using lp_solve version 5.5 in R. Applied Psychological Measurement, 35(5), 398-409.

Doğan, C., & Uçar, A. (2021). Defining cut point for Kullback-Leibler divergence to detect answer copying. *International Journal of Assessment Tools in Education*, 156–166. https://doi.org/10.21449/ijate.864078

Doğruöz, E. (2018). Bireyselleştirilmiş çok aşamalı testlerin test birleştirme yöntemlerine göre incelenmesi. Doktora Tezi. Hacettepe Üniversitesi.

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

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.

Embretson S. E., & Reise S. P. (2000). Item Response Theory for Psychologists. Mahwah, NJ: Lawrence Earlbaum.

Erdem Kara, B., & Doğan, N. (2022). The effect of ratio of items indicating differential item functioning on computer adaptive and multi-stage tests. *International Journal of Assessment Tools in Education*, *9*(3), 682–696. https://doi.org/10.21449/ijate.1105769

Erdem Kara, B., & Doğan, N. (2022). The effect of ratio of items indicating differential item functioning on computer adaptive and multi-stage tests. *International Journal of Assessment Tools in Education*, *9*(3), 682–696. https://doi.org/10.21449/ijate.1105769

Erkuş, A. (2012). Psikolojide ölçme ve ölçek geliştirme. Ankara: Pegem Akademi Yayınları.

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

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.

Georgiadou, E., Triantafillou, E., & Economides, A. A. (2007). A Review of Item Exposure Control Strategies for Computerized Adaptive Testing Developed from 1983 to 2005. Journal of Technology, Learning, and Assessment, 5(8), n8.

Gökçe, S., & Glas, C. A. W. (2018). Can TIMSS mathematics assessments be implemented as a computerized adaptive test? *Egitimde ve Psikolojide Olcme ve Degerlendirme Dergisi*, 421–436. https://doi.org/10.21031/epod.487351

Gündeğer, C., & Soysal, S. (2022). The effect of strong and weak unidimensional item pools on computerized adaptive classification testing. *Journal of Teacher Education and Lifelong Learning*. https://doi.org/10.51535/tell.1202804

Gür, R., & Gülleroğlu, H. D. (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*. https://doi.org/10.15390/eb.2020.8256

Hambleton, R. K., & Swaminathan, H. (1985). Assumptions of item response theory. In *Item Response Theory* (pp. 15-31). Springer, Dordrecht.

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.

Harter, R., Hornik, K., Theussl, S., Szymanski C. ve Schwendinger F. (2021). Rsymphony: An R interface to the SYMPHONY solver for mixed-integer linear programs, https://cran.r-project.org/web/packages/Rsymphony/Rsymphony.pdf

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

Hembry, I. F. (2014). Operational characteristics of mixed-format multistage tests using the 3PL testlet response theory model. 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

Himelfarb, I. (2019). A primer on standardized testing: History, measurement, classical test theory, item response theory, and equating. Journal of Chiropractic Education, 33(2), 151-163.

Huang, Y. M., Lin, Y. T., & Cheng, S. C. (2009). An adaptive testing system for supporting versatile educational assessment. Computers & Education, 52(1), 53-67.

International Business Machine Corporation. 2010. Efficient modeling with the IBM ILOG CPLEX optimization studio [White paper]. Retrieved from ftp://public.dhe.ibm.com/common/ssi/ecm/en/wsw14059usen/WSW14059USEN.PDF.

Jodoin, M. G., Zenisky, A., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. Applied Measurement in Education, 19(3), 203-220.

Kalender, İ. (2012). Computerized adaptive testing for student selection to higher education. *Yükseköğretim Dergisi*, *2*(1), 13-19.

Keng, L. (2008). A comparison of the performance of testlet-based computer adaptive tests and multistage tests. The University of Texas at Austin.

Keng, L. and Dodd, B. G. (2009). A comparison of the performance of testlet- based computer adaptive tests and multistage tests. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Kezer, F., & Koç, N. (2014). Bilgisayar ortamında bireye uyarlanmış test stratejilerinin karşılaştırılması. *Eğitim Bilimleri Araştırmaları Dergisi*, 4(1), 145-174.

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.

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.

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

Konis, K. (2016). lpSolveAPI, version 5.5. 2.0 [Computer software].

LaRoche, S., Joncas, M., & Foy, P. (2020). Sample design in TIMSS 2019. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), Methods and Procedures: TIMSS 2019 Technical Report (pp. 3.1-3.33). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/methods/chapter-3.html

Laukaityte, I., & Wiberg, M. (2017). Using plausible values in secondary analysis in large-scale assessments. *Communications in Statistics: Theory and Methods*, *46*(22), 11341–11357. https://doi.org/10.1080/03610926.2016.1267764

Li, J., & van der Linden, W. J. (2018). A Comparison of Constraint Programming and Mixed-Integer Programming for Automated Test-Form Generation. Journal of educational measurement, 55(4), 435-456, https://doi.org/10.1111/jedm.12187.

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.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*(3), 229–249. https://doi.org/10.1111/j.1745-3984.1998.tb00537.x

Luecht, R., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. Applied Measurement in Education, 19(3), 189-202.

Luo, X. (2020). Automated Test Assembly with Mixed-Integer Programming: The Effects of Modeling Approaches and Solvers. Journal of Educational Measurement, 57(4), 547-565, https://doi.org/10.1111/jedm.12262.

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.

Melican, G. J., Breithaupt, K., Zhang, Y., van der Linden, W. J., & Glas, C. A. W. (2010). Elements of Adaptive Testing.

Mills, C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (2002). Computer-based testing: Building the foundation for future assessments. New Jersey: Laurence Erlbaum Associates.

Mooney, C. Z. (1997). Monte carlo simulation (No. 116). Sage.

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

Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). TIMSS 2019 Assessment Frameworks. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2019/frameworks/

Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *ETS Research Report Series*, *1992*(1), i–30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

National Center for Education Statistics (NCES). (2019). Program for International Student Assessment 2022 (PISA 2022) Main Study Recruitment and Field Test.

Optimization, G. Inc.(2017). Gurobi optimizer, version 7.5.

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

Ozdemir, B., & Gelbal, S. (2022). Measuring language ability of students with compensatory multidimensional CAT: A post-hoc simulation study. *Education and Information Technologies*, *27*(5), 6273–6294. https://doi.org/10.1007/s10639-021-10853-0

Ozturk, N. B., & Dogan, N. (2015). Investigating item exposure control methods in computerized adaptive testing. Yayımlanmamış Doktora Tezi. Hacettepe Üniversitesi.

Özberk, E. H., & Gelbal, S. (2017). Basit ve Karmaşık Test Desenlerinde Çok Boyutlu Madde Seçme Yöntemlerinin Karşılaştırılması. *Egitimde ve Psikolojide Olcme ve Degerlendirme Dergisi*, 34–34. https://doi.org/10.21031/epod.286956

Park, R. (2015). Investigating the impact of a mixed-format item pool on optimal test designs for multistage testing (Doctoral dissertation).

Park, R., Kim, J., Chung, H., & Dodd, B. G. (2014). Enhancing pool utilization in constructing the multistage test using mixed-format tests. Applied Psychological Measurement, 38(4), 268-280.

Patsula, L. N. (1999). A comparison of computerized adaptive testing and multistage testing. University of Massachusetts Amherst.

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.

Purves, A. C. 1987. The evolution of the IEA: A memoir. Comparative Education Review, 31(1), 10–28.

Raborn, A., & Sari, H. (2021). Mixed adaptive multistage testing: A new approach. *Egitimde ve Psikolojide Olcme ve Degerlendirme Dergisi*. https://doi.org/10.21031/epod.871014

Reckase, M. D. (2009). Multidimensional item response theory models. In Multidimensional item response theory (pp. 79-112). Springer, New York, NY.

Reckase, M.D. (1989). Adaptive testing: The evolution of a good idea. Educational Measurement Issues and Practice, 8, 11-15.

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.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher (Washington, D.C.: 1972)*, *39*(2), 142–151. https://doi.org/10.3102/0013189x10363170

Sari, H. I., Yahsi-Sari, 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.

Sari, H. İ., & Huggins-Manley, A. C. (2017). Examining content control in adaptive tests: Computerized adaptive testing vs. computerized adaptive multistage testing. *Educational Sciences: Theory & Practice*, *17*(5).

Schnipke, D. L., & Reese, L. M. (1999). A Comparison [of] Testlet-Based Test Designs for Computerized Adaptive Testing. Law School Admission Council Computerized Testing Report. LSAC Research Report Series.

Schwendinger, F. (2020). lpSolveAPI: R Interface to 'lp_solve', https://cran.r-project.org/web/packages/lpSolveAPI/lpSolveAPI.pdf

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

Stefan, T, S., Hornik, K., Buchta, C., Schwendinger, F. and Schuchardt, H. (2019). Rglpk: R/GNU Linear Programming Kit Interface, https://cran.r-project.org/web/packages/Rglpk/Rglpk.pdf

Stocking, M. L., & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In Computerized adaptive testing: Theory and practice (pp. 163-182). Springer, Dordrecht.

Sulak, S., & Kelecioğlu, H. (2019). Investigation of item selection methods according to test termination rules in CAT applications. *Egitimde ve Psikolojide Olcme ve Degerlendirme Dergisi*, 315–326. https://doi.org/10.21031/epod.530528

Sünbül, Ö., & Erkuş, A. (2013). Madde parametrelerininin değişmezliğinin çeşitli boyutluluk özelliği gösteren yapılarda madde tepki kuramınana göre incelenmesi. *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, *9*(2), 378-398.

Svetina, D., Liaw, Y. L., Rutkowski, L., & Rutkowski, D. (2019). Routing Strategies and Optimizing Design for Multistage Testing in International Large-Scale Assessments. Journal of Educational Measurement, 56(1), 192-213.

Şahin Kürşad, M. (2020). Madde parametre sapmasının bireye uyarlanmış testlerden elde edilen yetenek kestirimlerine etkisinin farklı koşullar altında incelenmesi. Yayımlanmamış Doktra Tezi. Ankara Üniversitesi

Şahin, M. D., & Gelbal, P. D. S. (2020). Development of a multidimensional computerized adaptive test based on the bifactor model. *International Journal of Assessment Tools in Education*, 323–342. https://doi.org/10.21449/ijate.707199

Şahin, M. G. (2020). Analyzing different module characteristics in computer adaptive multistage testing. *International Journal of Assessment Tools in Education*, *7*(2), 191–206. https://doi.org/10.21449/ijate.676947

Şenel, S. (2017). Bilgisayar ortamında bireye uyarlanmış testlerin görme engelli öğrencilere uygunluğunun incelenmesi. Yayımlanmamış Doktora Tezi. Ankara Üniversitesi.

Şenel, S. Y. (2021). Bilgisayar ortamında bireye uyarlanmış testlerin görme engelli öğrencilere uygunluğunun incelenmesi (Doctoral dissertation, Ankara Üniversitesi Eğitim Bilimleri Enstitüsü Ölçme Ve Değerlendirme Anabilim Dalı).

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

The Organisation for Economic Co-operation and Development (OECD). (2017). PISA 2015 Technical Report. Paris: OECD Publishing. https://www.oecd.org/pisa/sitedocument/PISA-2015-technicalreport-final.pdf

Theussl, S., Hornik, K., Buchta, C., Schwendinger, F., Schuchardt, H., & Theussl, M. S. (2019). Package 'Rglpk'. *GitHub, Inc., San Francisco, CA, USA, Tech. Rep. 0.6-4*.

Tian, C. (2018). Comparison of four stopping rules in computerized adaptive testing and examination of their application to on-the-fly multistage testing. Master Thesis, University of Illinois.

van Der Linden, W. J. (2005). A comparison of item-selection methods for adaptive tests with content constraints. Journal of Educational Measurement, 42(3), 283-302, https://doi.org/10.1111/j.1745-3984.2005.00015.x.

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 (Vol. 10, pp. 978-0). 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., & 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

von Davier, M. (2020). TIMSS 2019 scaling methodology: Item Response Theory, population models, and linking across modes. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), Methods and Procedures: TIMSS 2019 Technical Report (pp. 11.1-11.25). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/methods/chapter-11.html

Wainer, H. (1990). An Adaptive Algebra Test: A Testlet-Based, Hierarchically-Structured Test with Validity-Based Scoring. Technical Report No. 90-92.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). Testlet response theory and its applications. Cambridge University Press.

Wang, W., Chen, J., & Kingston, N. (2020). How Well Do Simulation Studies Inform Decisions About Multistage Testing?. Journal of Applied Measurement, 21(3), 271-281.

Wang, X., Fluegge, L., & Luecht, R. M. (2012). A large-scale comparative study of the accuracy and efficiency of ca-MST panel design configurations. In Annual Meeting of the National Council on Measurement in Education.

Weissman, A., Belov, D. I., & Armstrong, R. D. (2007). LSAC RESEARCH REPORT SERIES.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2–3), 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005

Xing, D., & Hambleton, R. K. (2004). Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. Educational and Psychological Measurement, 64(1), 5-21.

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). https://doi.org/10.1103/physrevphyseducres.17.010115

Yigiter, M. S. & Doğan, N. (2023). The Effect of Test Design on Misrouting in Computerized Multistage Testing. *International Journal of Turkish Education Sciences*.

Yiğiter, M. S. & Doğan, N. (2023). Bireyselleştirilmiş Çok Aşamalı Testlerde Test Tasarımının Yanlış Yönlendirmeye Etkisi. *Uluslararası Türk Eğitim Bilimleri Dergisi*.

Yin, L., & Foy, P. (2021). TIMSS 2023 Assessment Design. *TIMSS 2023 Assessment Frameworks*, 71.

Zenisky, A. L. (2004). Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment. University of Massachusetts Amherst.

Zenisky, A. L., & Hambleton, R. K. (2014). Multistage test designs: Moving research results into practice. In D. Yan, C. Lewis, & A. von Davier (Eds.), Computerized multistage testing theory and applications (pp. 21–38). CRC Press.

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2009). Multistage testing: Issues, designs, and research. In Elements of adaptive testing (pp. 355-372). Springer, New York, NY.

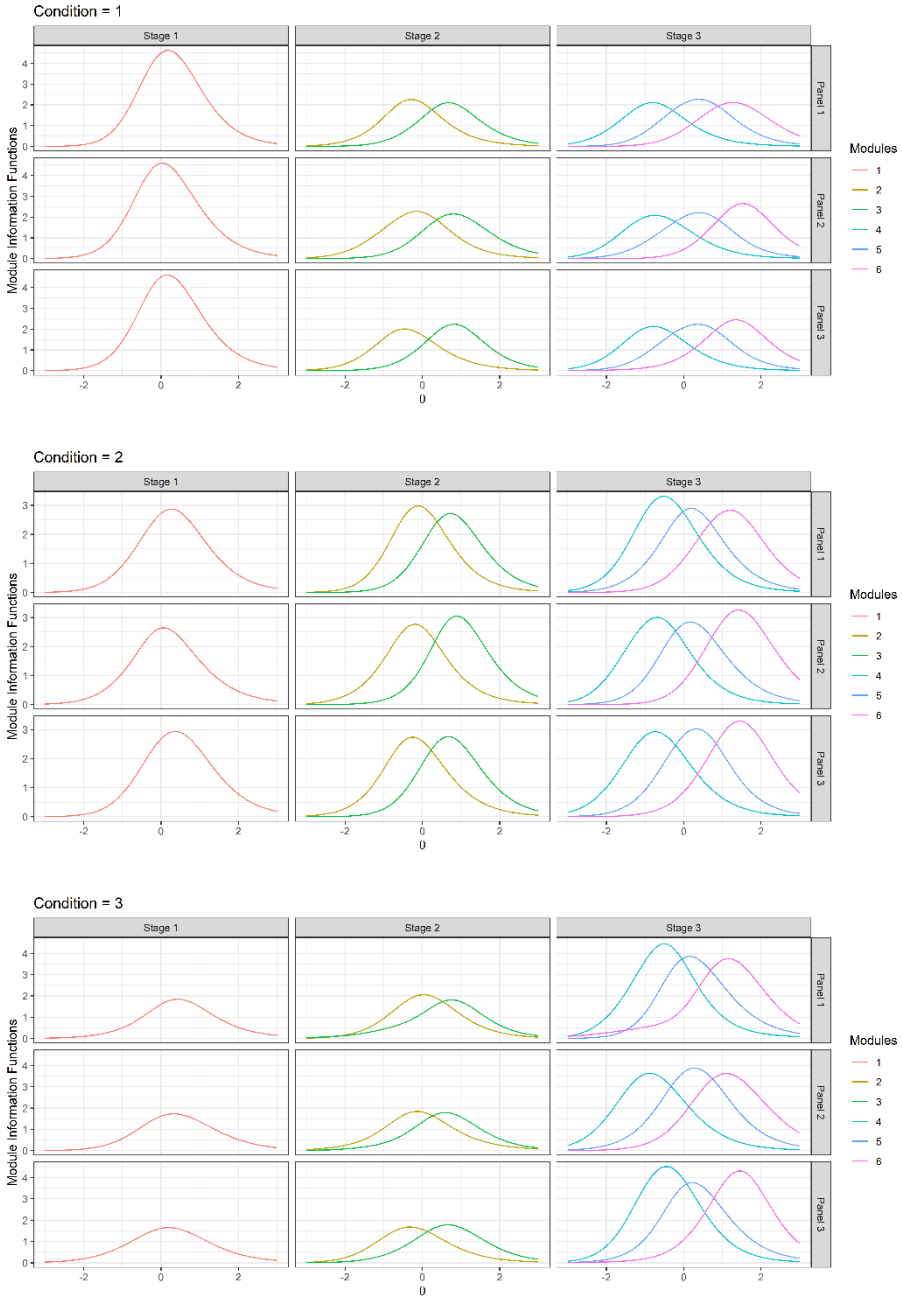
Zheng, W. (2016). Making test batteries adaptive by using multistage testing techniques (Doctoral dissertation, University of North Carolina, Greensboro, NC).

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

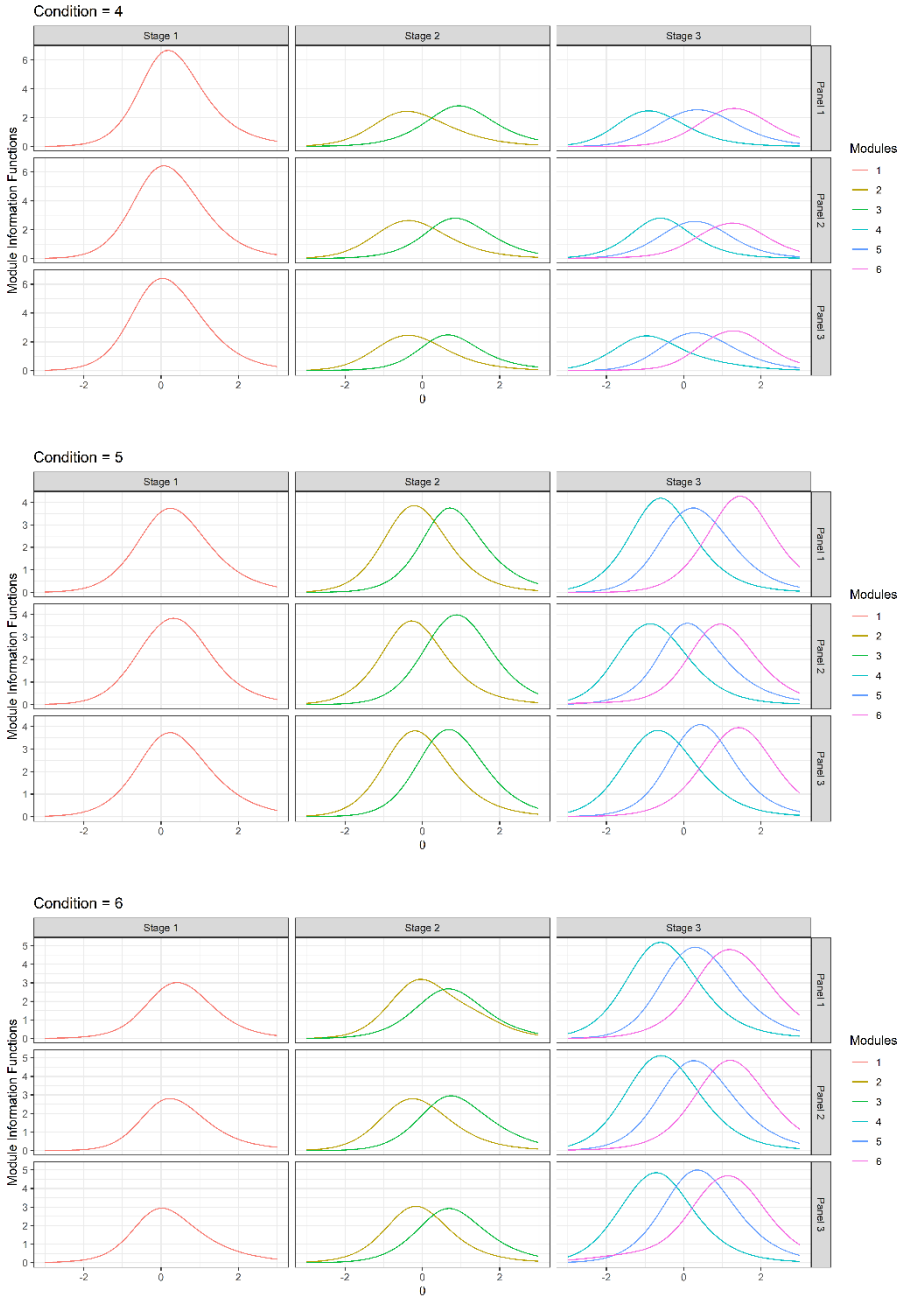
Zheng, Y., Nozawa, Y., Gao, X., & Chang, H. H. (2012). Multistage Adaptive Testing for a Large-Scale Classification Test: Design, Heuristic Assembly, and Comparison with Other Testing Modes. ACT Research Report Series, 2012 (6). ACT, Inc.

Zheng, Y., Wang, C., Culbertson, M. J., & Chang, H. H. (2016). Overview of test assembly methods in multistage testing. In D. L. Yan, A. A. von Davier, & C. Lewis (Eds.),Computerized multistage testing: Theory and practice(pp. 125–138). Boca Raton, FL: Chapman& Hall/CRC.

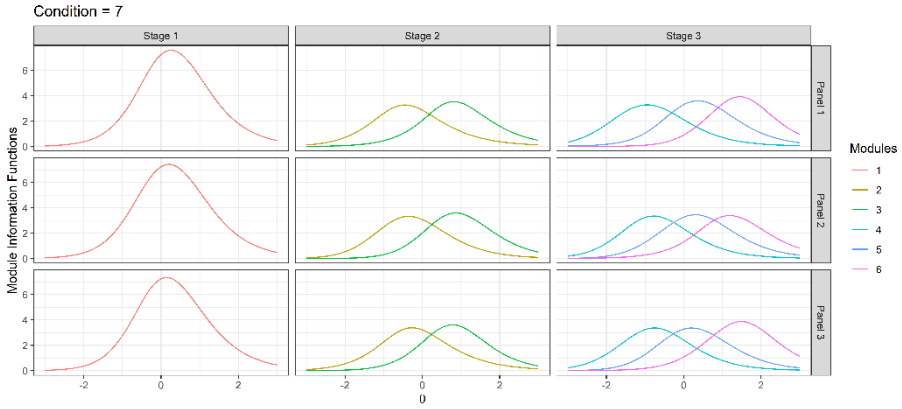
EK-A: S-BÇAT Tasarımlarının Panel ve Modüllere Göre Madde Bilgi Fonksiyonu Grafikleri

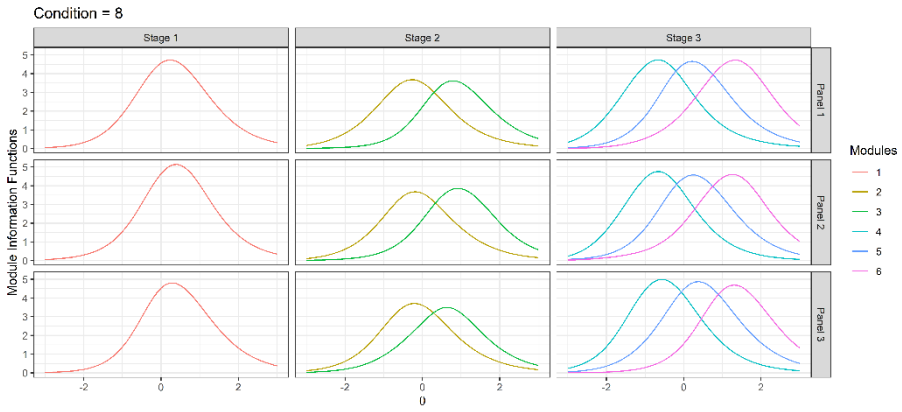


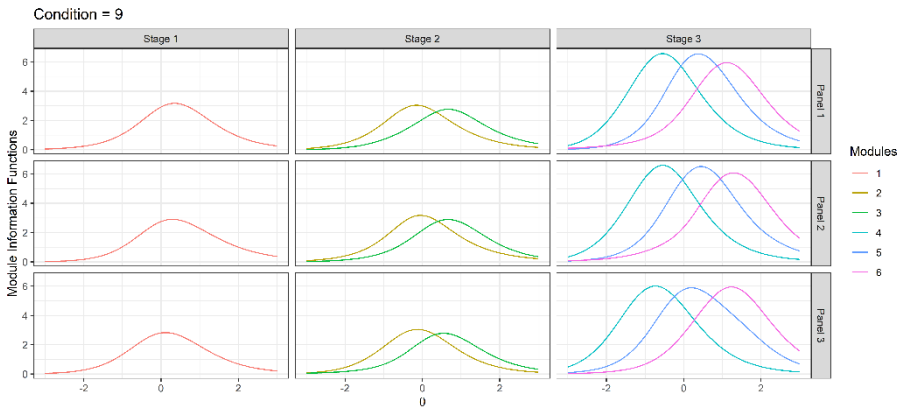
EK-A: S-BÇAT Tasarımlarının Panel ve Modüllere Göre Madde Bilgi Fonksiyonu Grafikleri (DEVAMI)



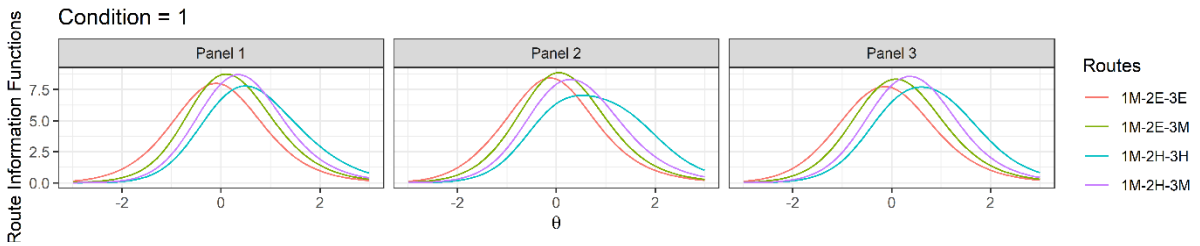
EK-A: S-BÇAT Tasarımlarının Panel ve Modüllere Göre Madde Bilgi Fonksiyonu Grafikleri (DEVAMI)

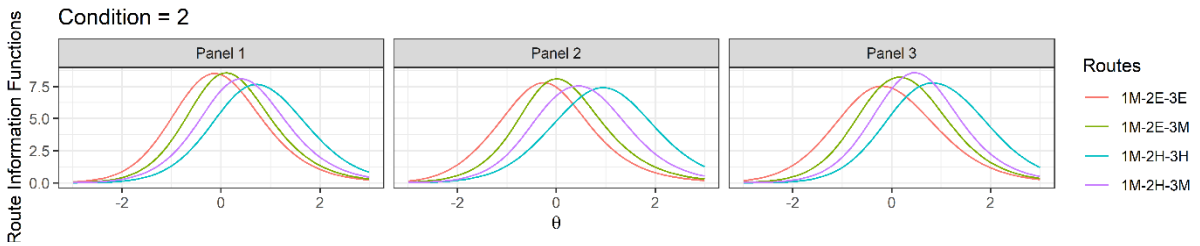


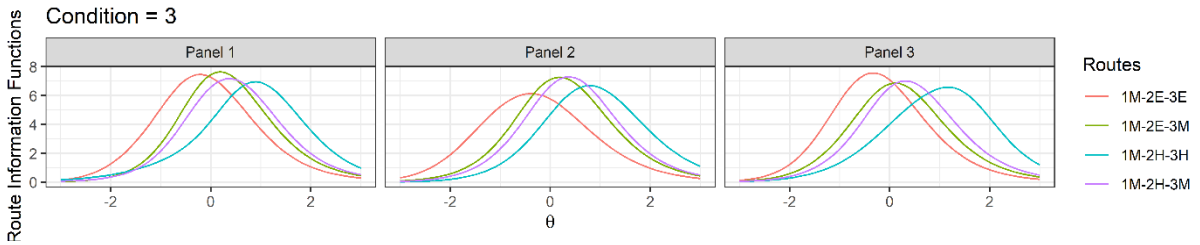


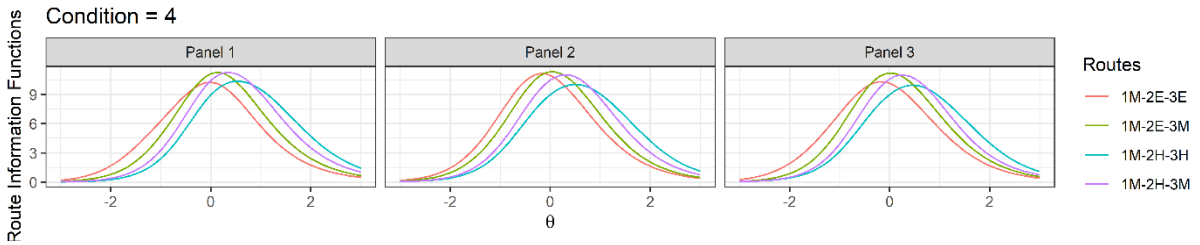


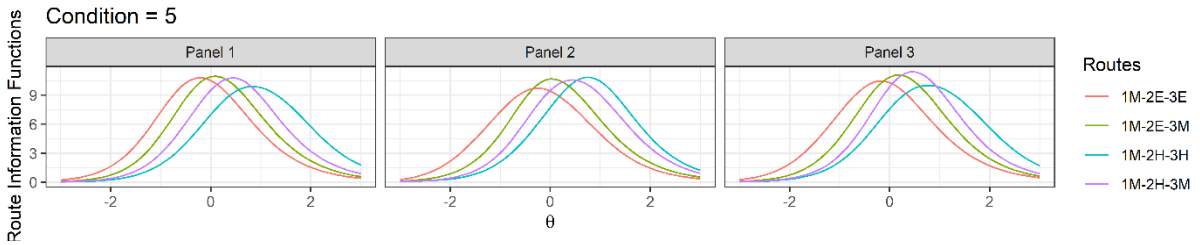
EK-B: S-BÇAT Tasarımlarının Panel ve Rotalara Göre Madde Bilgi Fonksiyonu Grafikleri

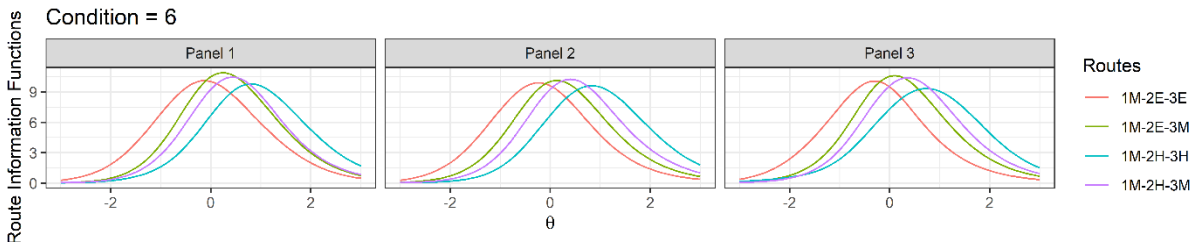




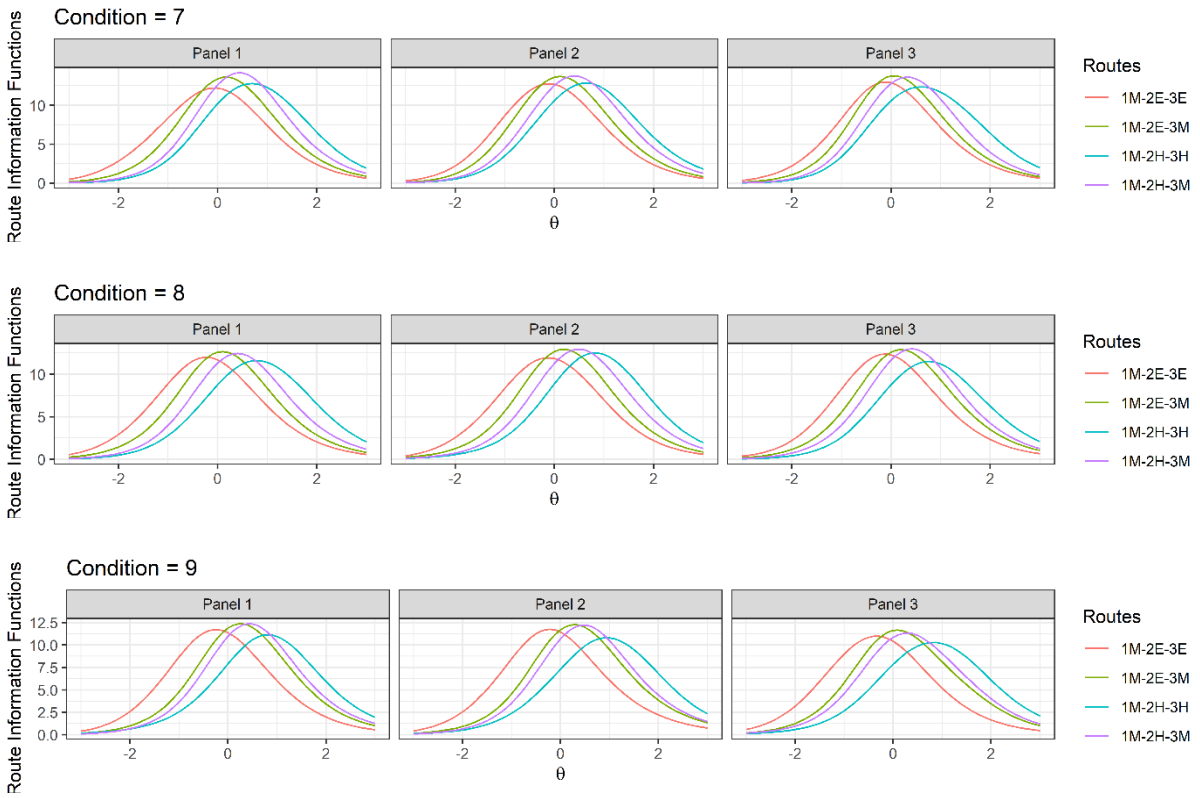




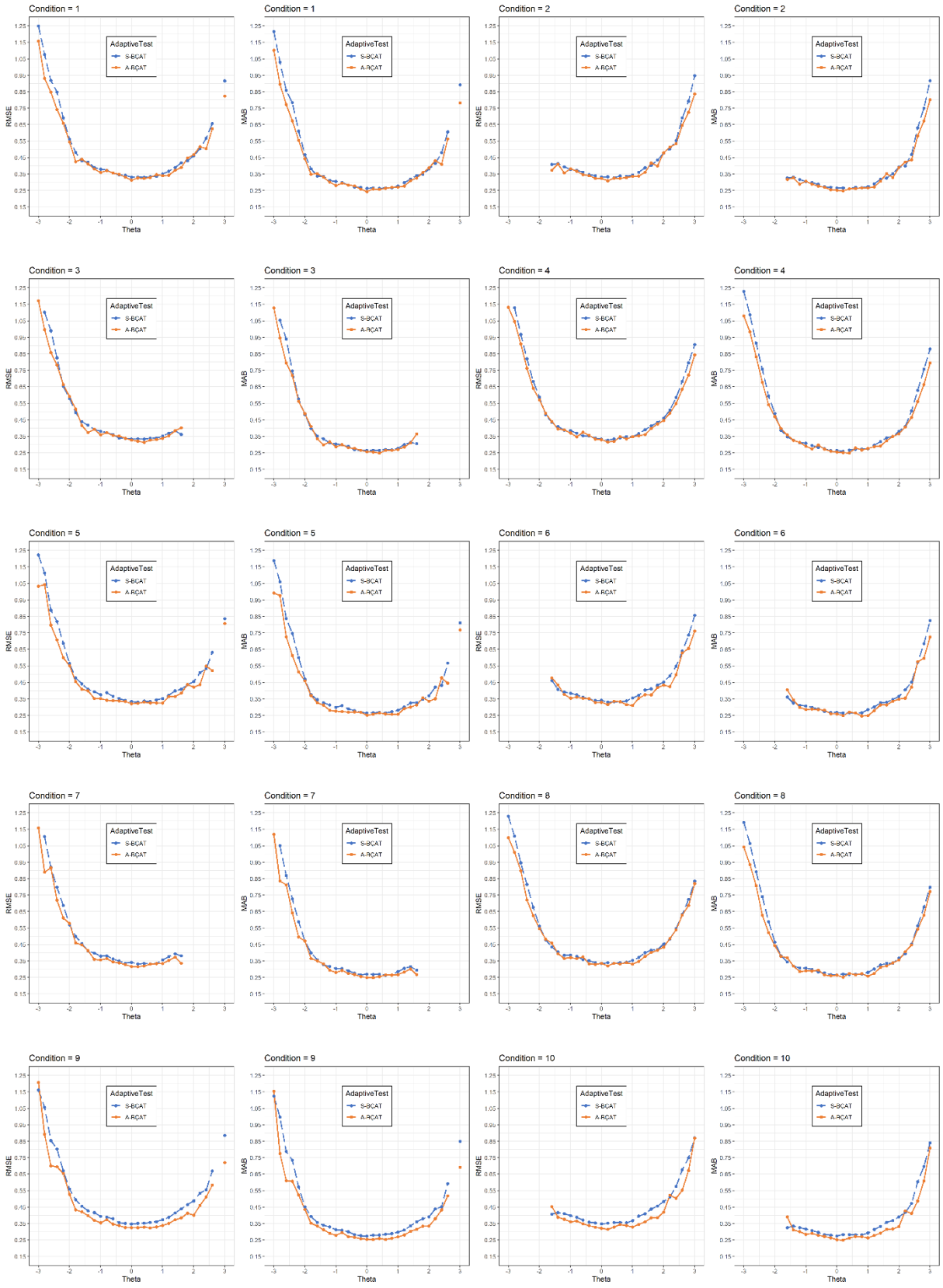




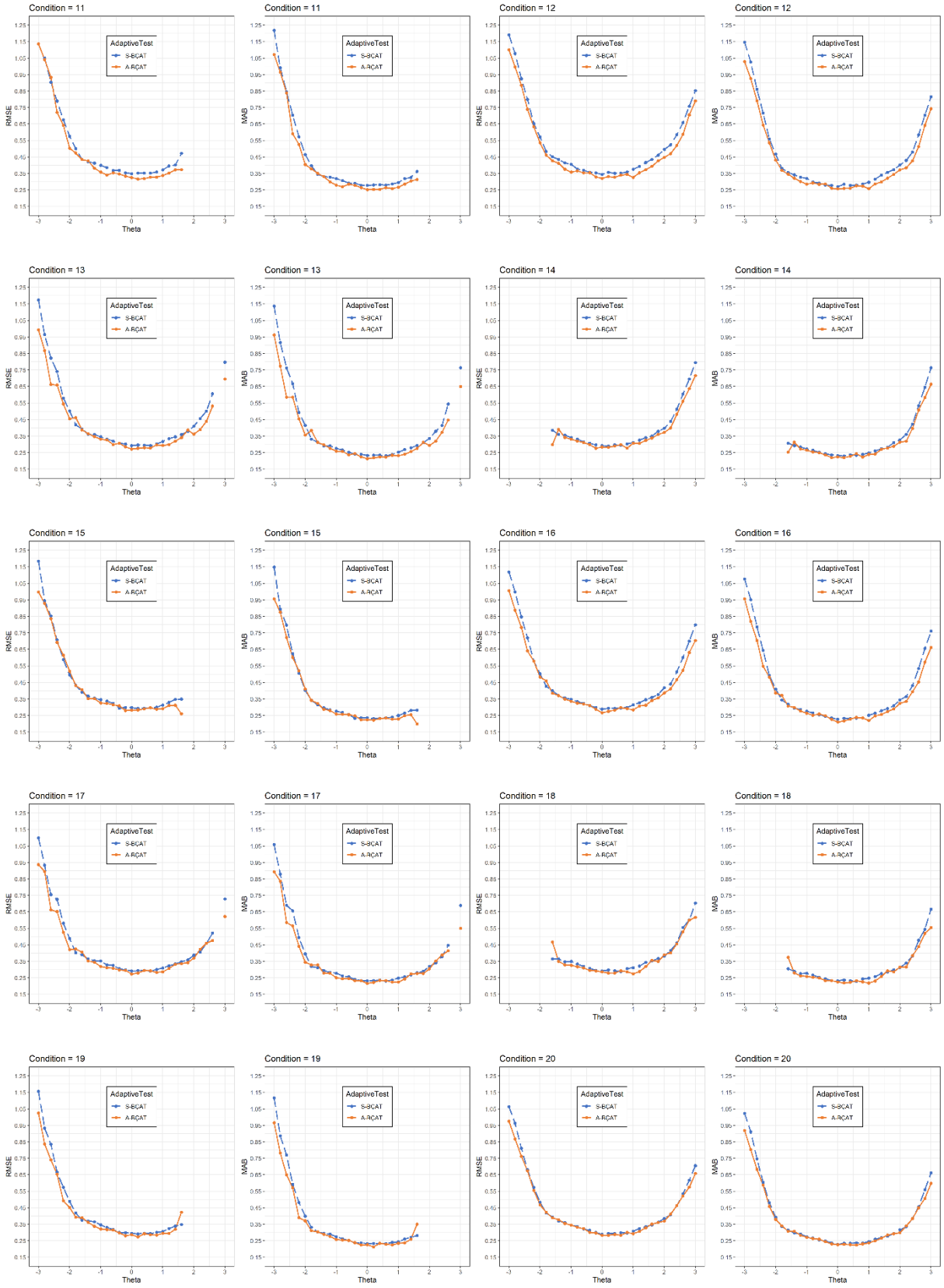
EK-B: S-BÇAT Tasarımlarının Panel ve Rotalara Göre Madde Bilgi Fonksiyonu Grafikleri

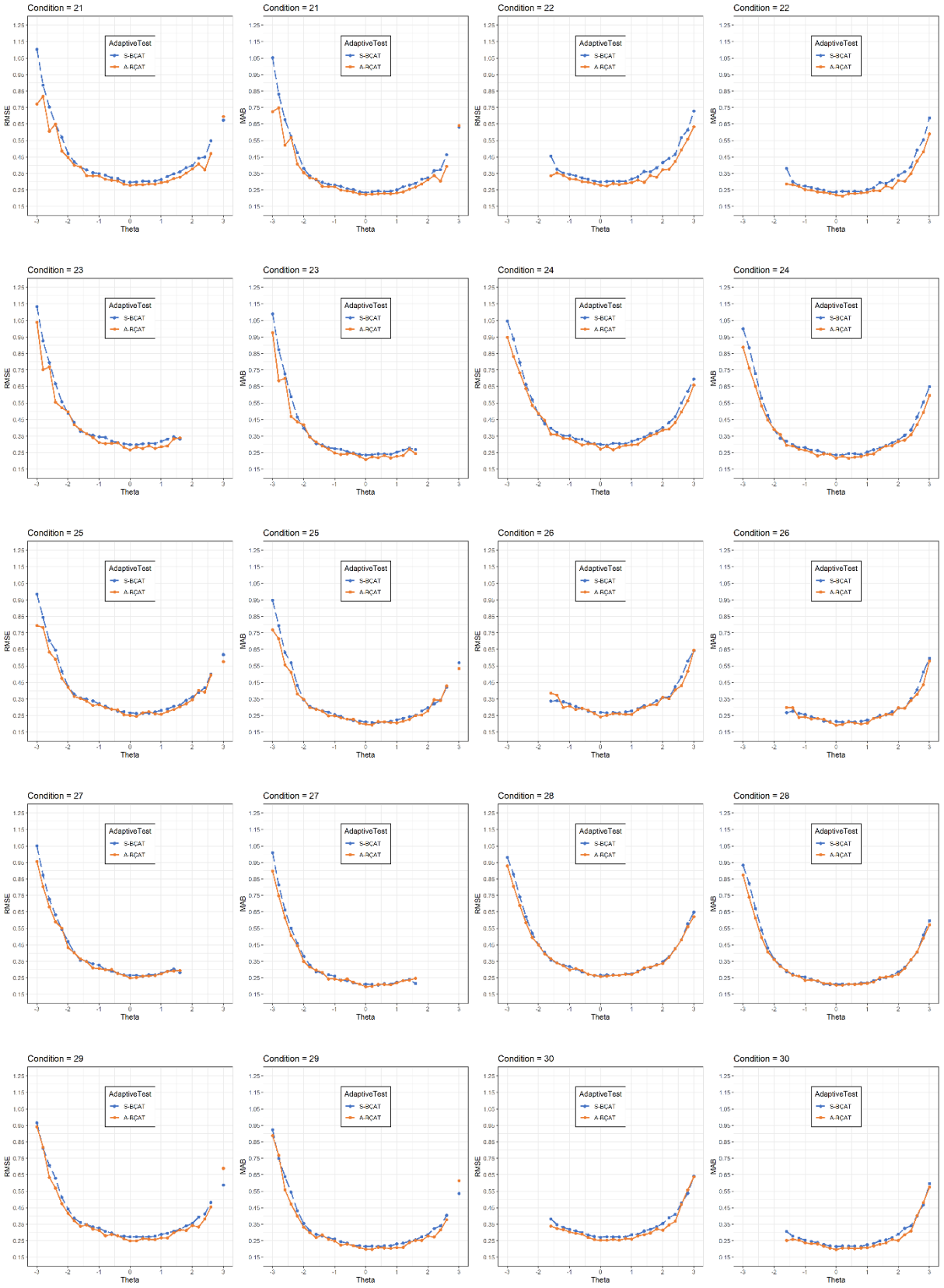


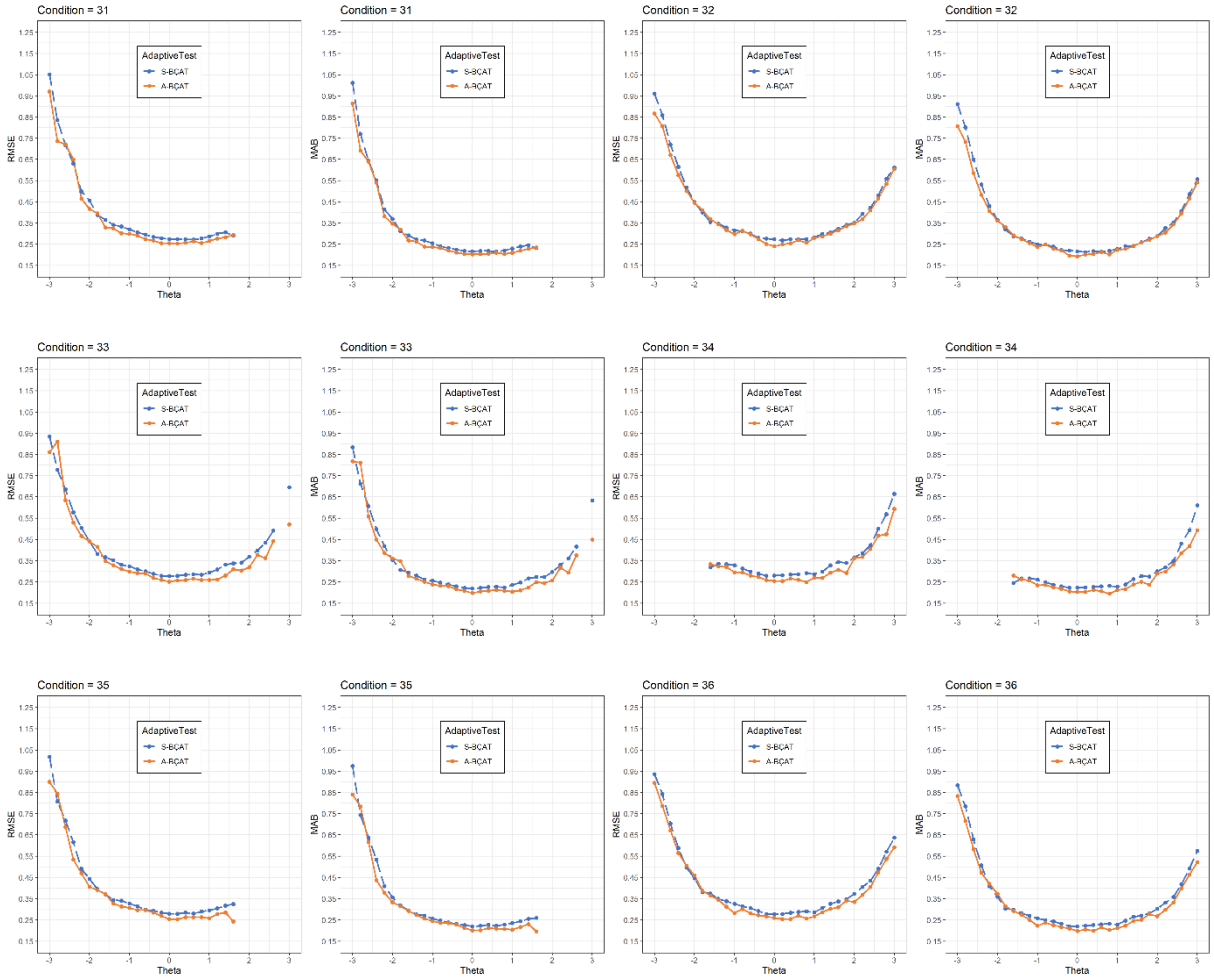
EK-C: S-BÇAT ile A-BÇAT'ın Ölçme Kesinliğinin Yetenek Düzeyine Göre Karşılaştırıldığı Grafikler (Tüm Koşullar)



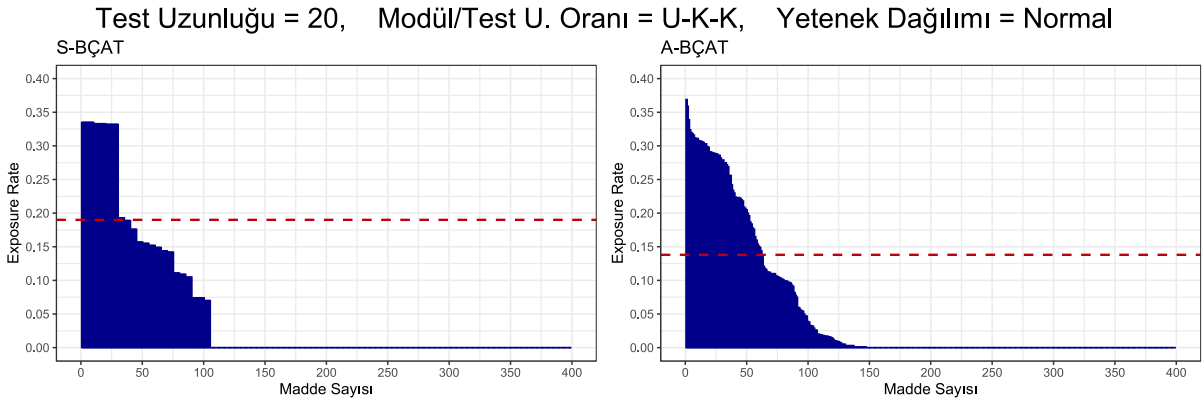
cxl

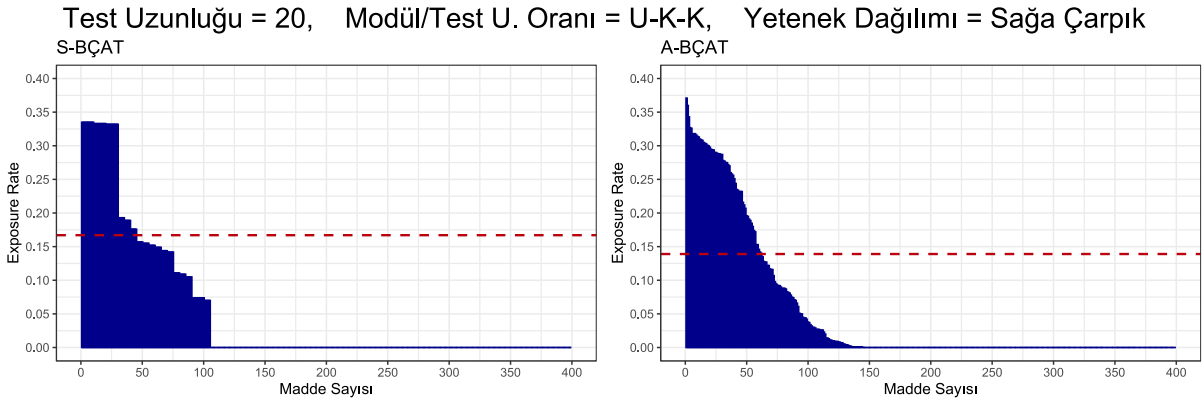

Condition = 11
Condition = 11
Condition = 12
Condition = 12
Condition = 13
Condition = 13
Condition = 14
Condition = 14
Condition = 15
Condition = 15
Condition = 16
Condition = 16
Condition = 17
Condition = 17
Condition = 18
Condition = 18
Condition = 19
Condition = 19
Condition = 20
Condition = 20
AdaptiveTest
S-BCAT
A-BCAT
RMSE
MAB
Theta

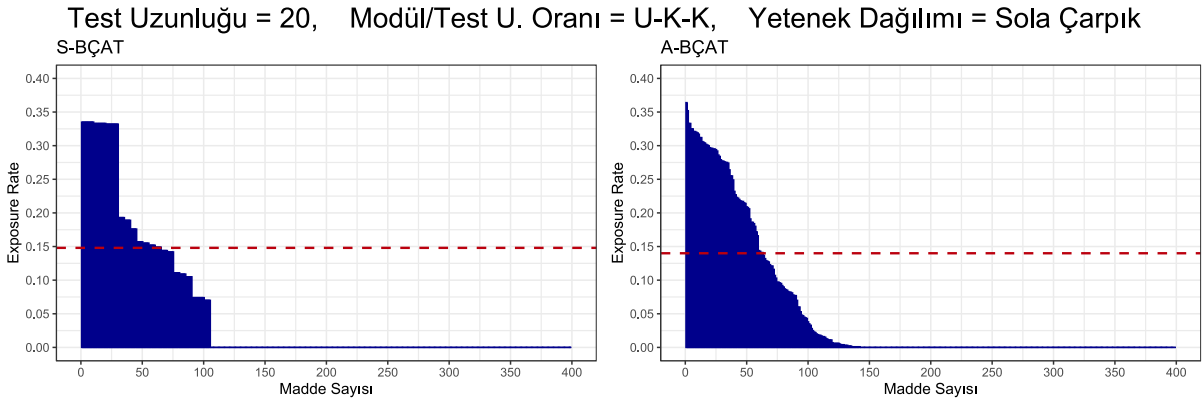

Condition = 21
Condition = 21
Condition = 22
Condition = 22
Condition = 23
Condition = 23
Condition = 24
Condition = 24
Condition = 25
Condition = 25
Condition = 26
Condition = 26
Condition = 27
Condition = 27
Condition = 28
Condition = 28
Condition = 29
Condition = 29
Condition = 30
Condition = 30
AdaptiveTest
S-BCAT
A-BCAT
RMSE
MAB
Theta


Condition = 31
Condition = 31
Condition = 32
Condition = 32
Condition = 33
Condition = 33
Condition = 34
Condition = 34
Condition = 35
Condition = 35
Condition = 36
Condition = 36
AdaptiveTest
S-BCAT
A-BCAT
RMSE
MAB
Theta

EK-D: S-BÇAT ile A-BÇAT'ın Madde Düzeyinde Madde Kullanım Sıklıklarının Karşılaştırılması (Tüm Koşullar)

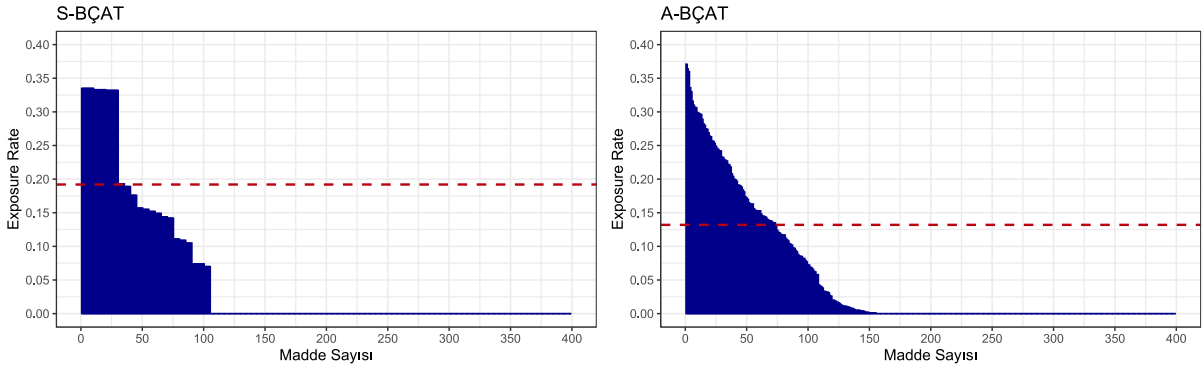




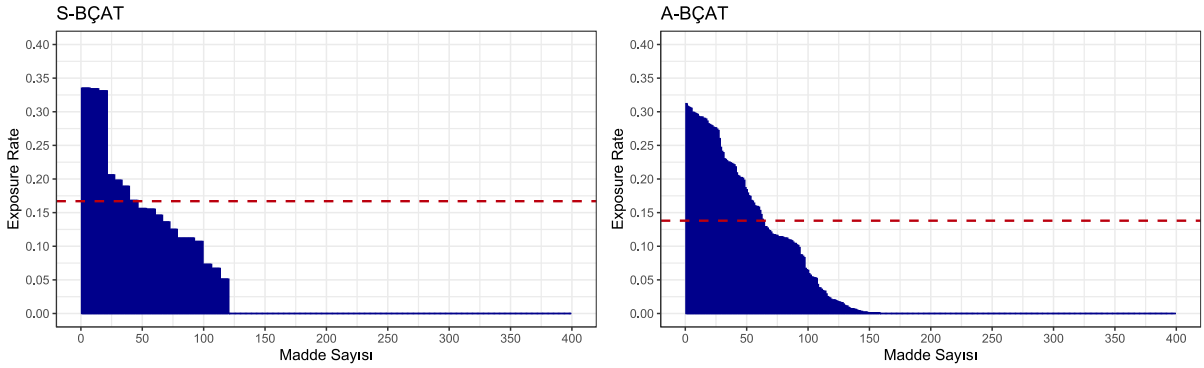


EK-D: S-BÇAT ile A-BÇAT'ın Madde Düzeyinde Madde Kullanım Sıklıklarının Karşılaştırılması (Tüm Koşullar) (DEVAMI)

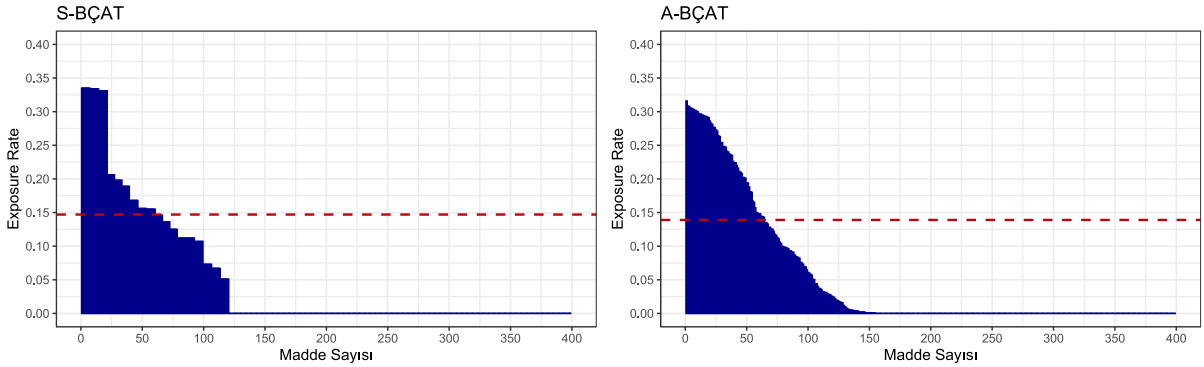
Test Uzunluğu = 20, Modül/Test U. Oranı = U-K-K, Yetenek Dağılımı = Tekdüze



Test Uzunluğu = 20, Modül/Test U. Oranı = O-O-O, Yetenek Dağılımı = Normal

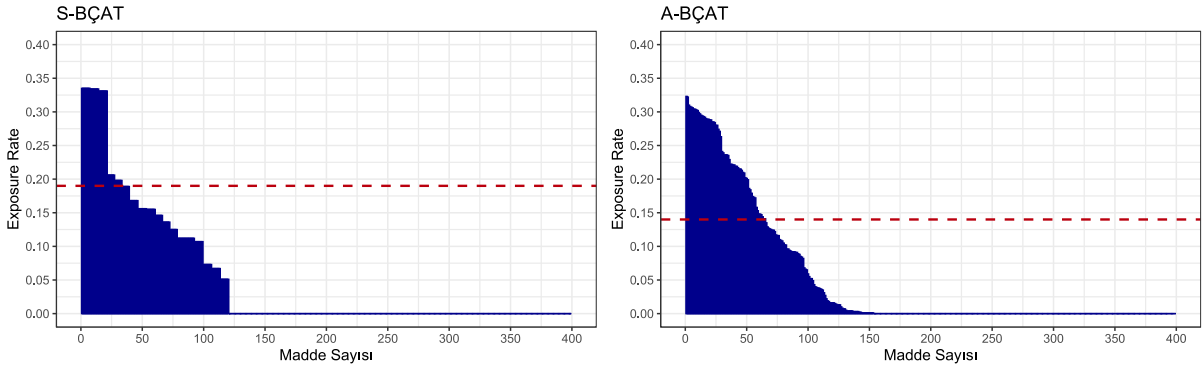


Test Uzunluğu = 20, Modül/Test U. Oranı = O-O-O, Yetenek Dağılımı = Sağa Çarpık

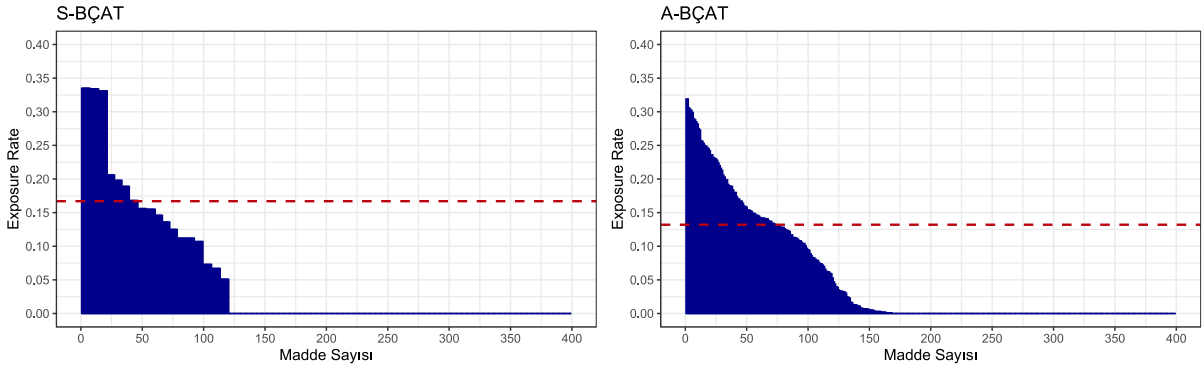


EK-D: S-BÇAT ile A-BÇAT'ın Madde Düzeyinde Madde Kullanım Sıklıklarının Karşılaştırılması (Tüm Koşullar) (DEVAMI)

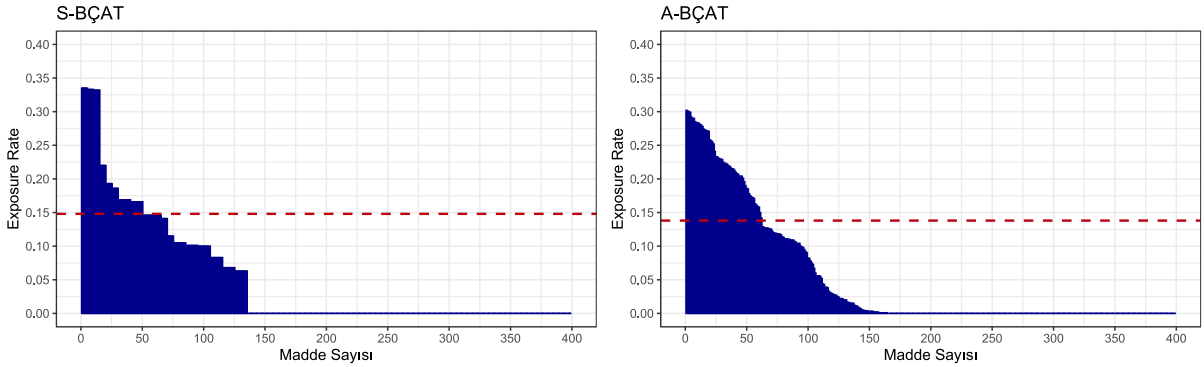
Test Uzunluğu = 20, Modül/Test U. Oranı = O-O-O, Yetenek Dağılımı = Sola Çarpık



Test Uzunluğu = 20, Modül/Test U. Oranı = O-O-O, Yetenek Dağılımı = Tekdüze

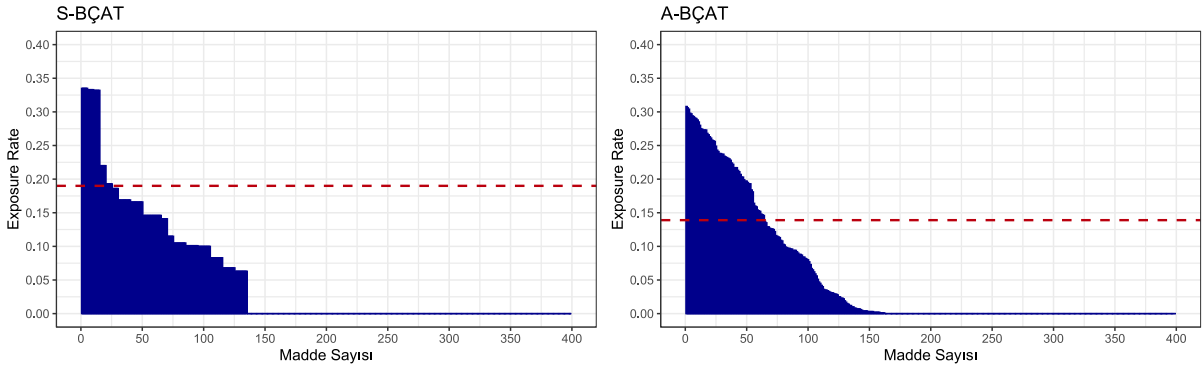


Test Uzunluğu = 20, Modül/Test U. Oranı = K-K-U, Yetenek Dağılımı = Normal

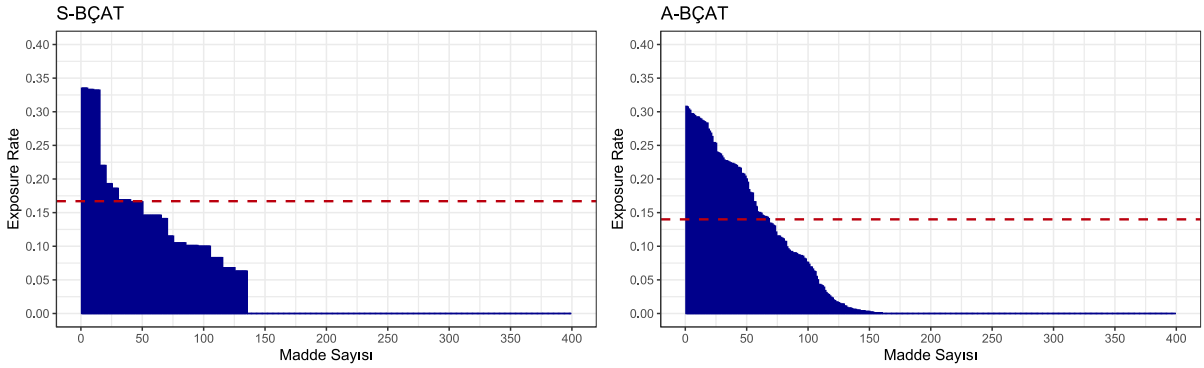


EK-D: S-BÇAT ile A-BÇAT'ın Madde Düzeyinde Madde Kullanım Sıklıklarının Karşılaştırılması (Tüm Koşullar) (DEVAMI)

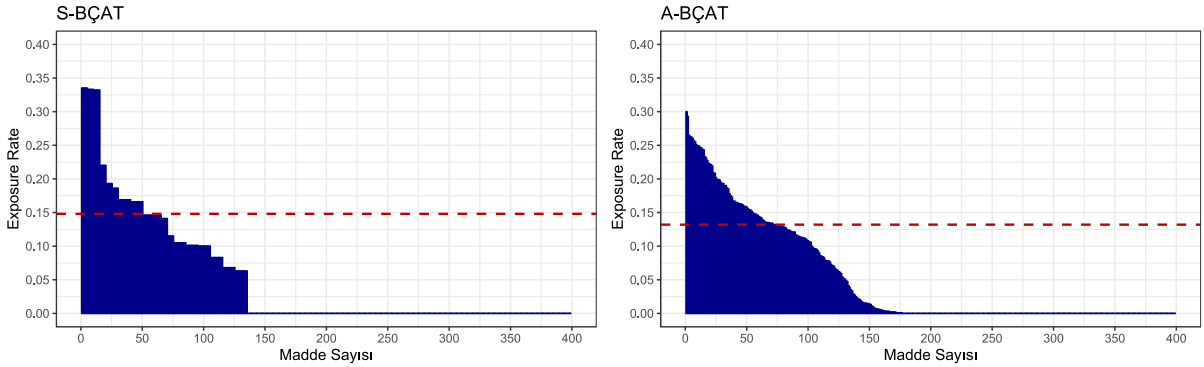
Test Uzunluğu = 20, Modül/Test U. Oranı = K-K-U, Yetenek Dağılımı = Sağa Çarpık



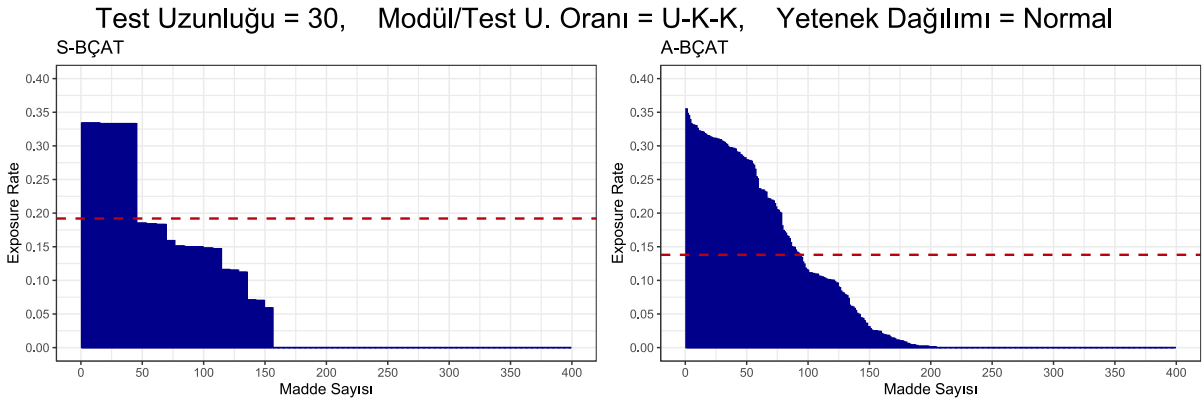
Test Uzunluğu = 20, Modül/Test U. Oranı = K-K-U, Yetenek Dağılımı = Sola Çarpık

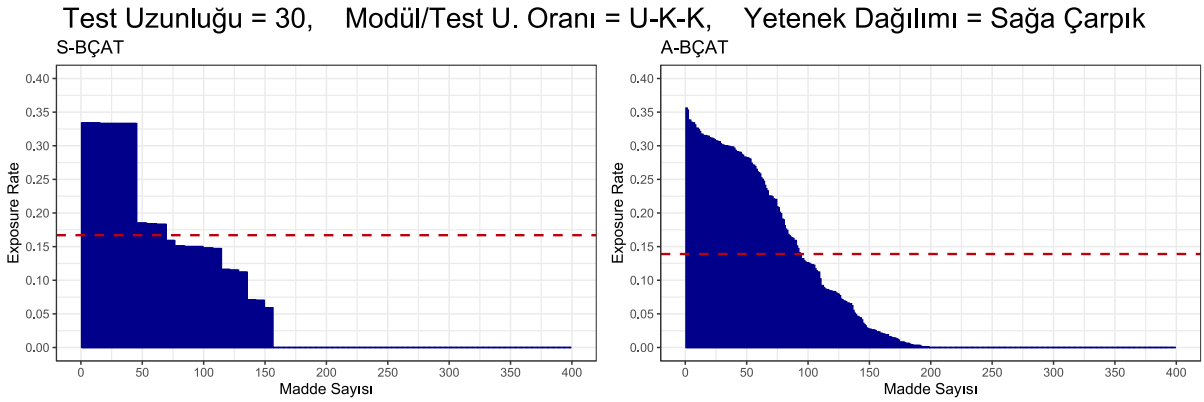


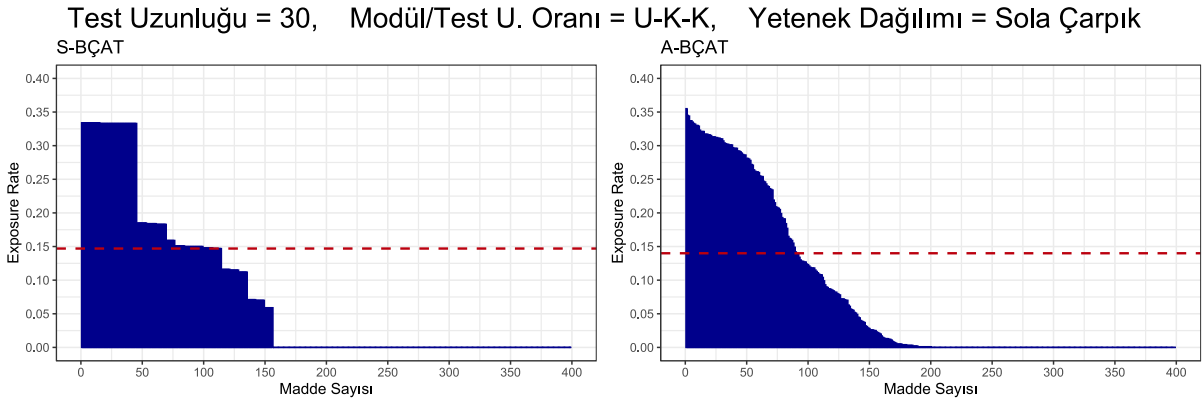
Test Uzunluğu = 20, Modül/Test U. Oranı = K-K-U, Yetenek Dağılımı = Tekdüze



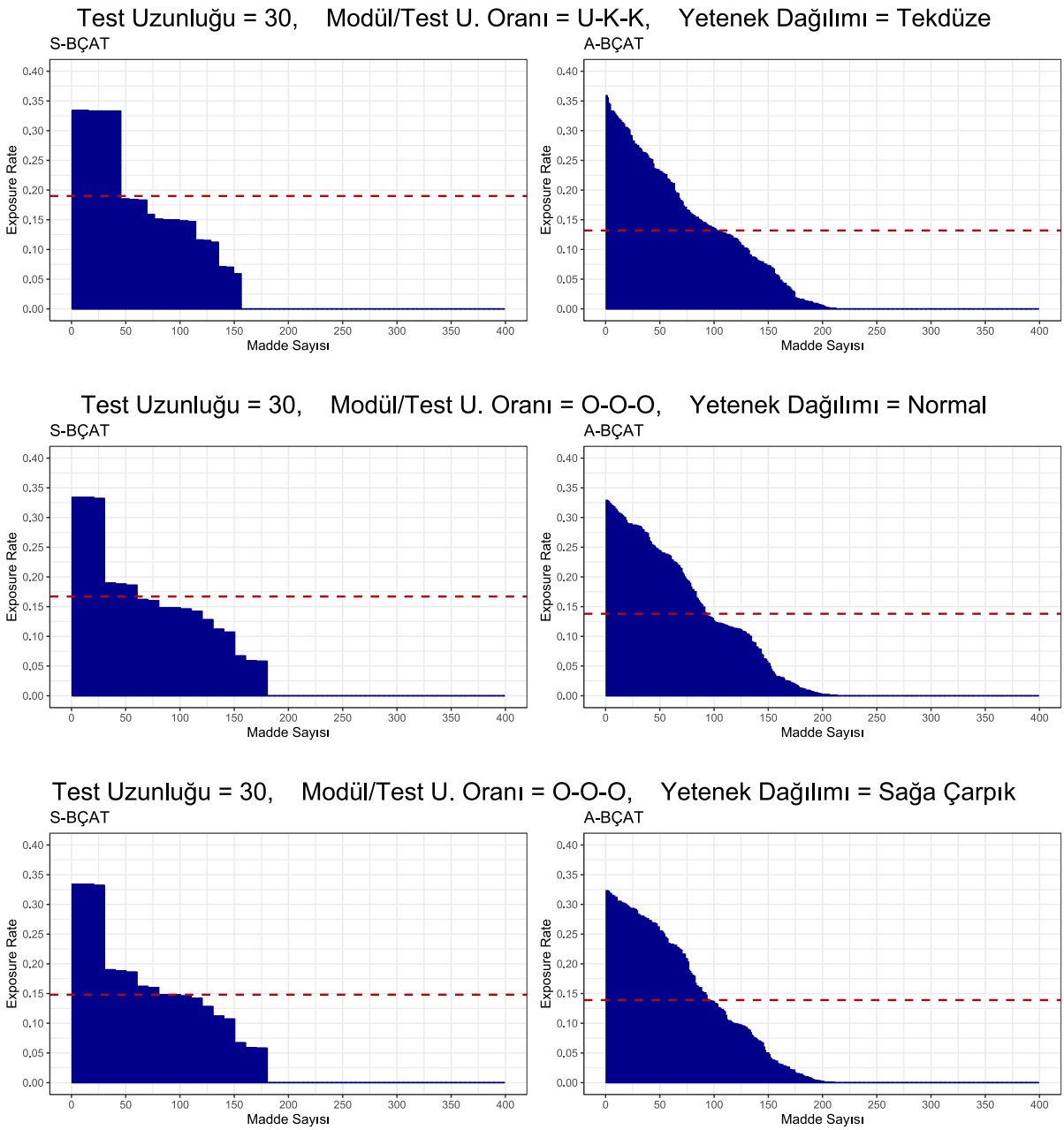
EK-D: S-BÇAT ile A-BÇAT'ın Madde Düzeyinde Madde Kullanım Sıklıklarının Karşılaştırılması (Tüm Koşullar) (DEVAMI)





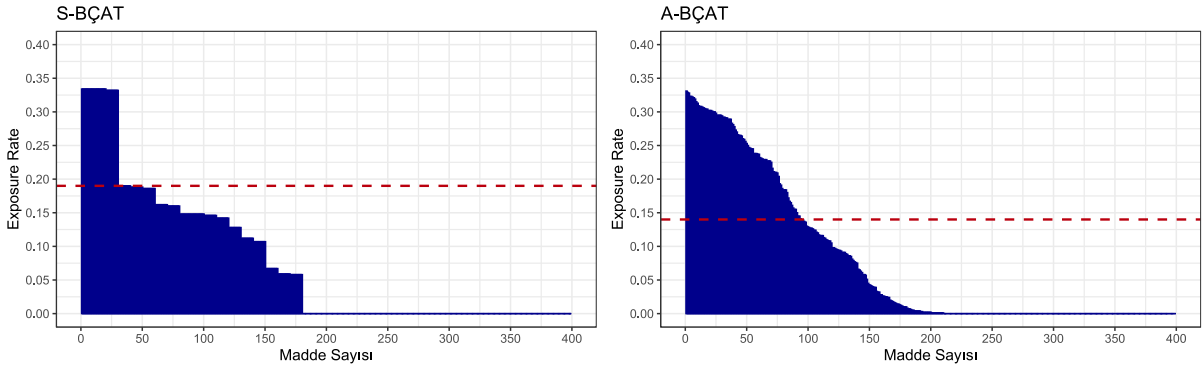


EK-D: S-BÇAT ile A-BÇAT'ın Madde Düzeyinde Madde Kullanım Sıklıklarının Karşılaştırılması (Tüm Koşullar) (DEVAMI)

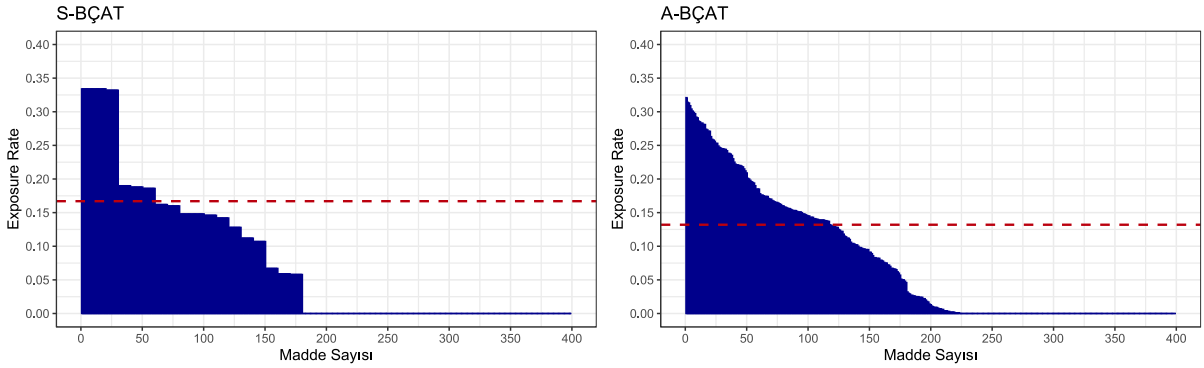


EK-D: S-BÇAT ile A-BÇAT'ın Madde Düzeyinde Madde Kullanım Sıklıklarının Karşılaştırılması (Tüm Koşullar) (DEVAMI)

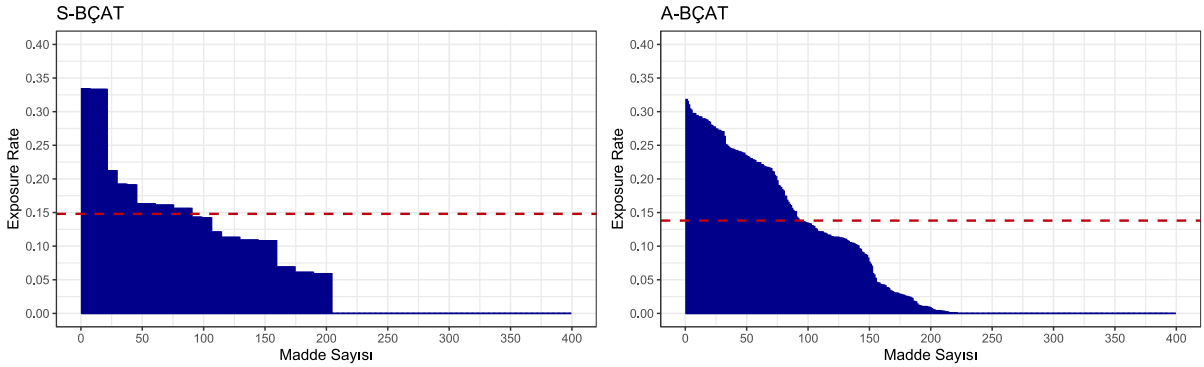
Test Uzunluğu = 30, Modül/Test U. Oranı = O-O-O, Yetenek Dağılımı = Sola Çarpık



Test Uzunluğu = 30, Modül/Test U. Oranı = O-O-O, Yetenek Dağılımı = Tekdüze

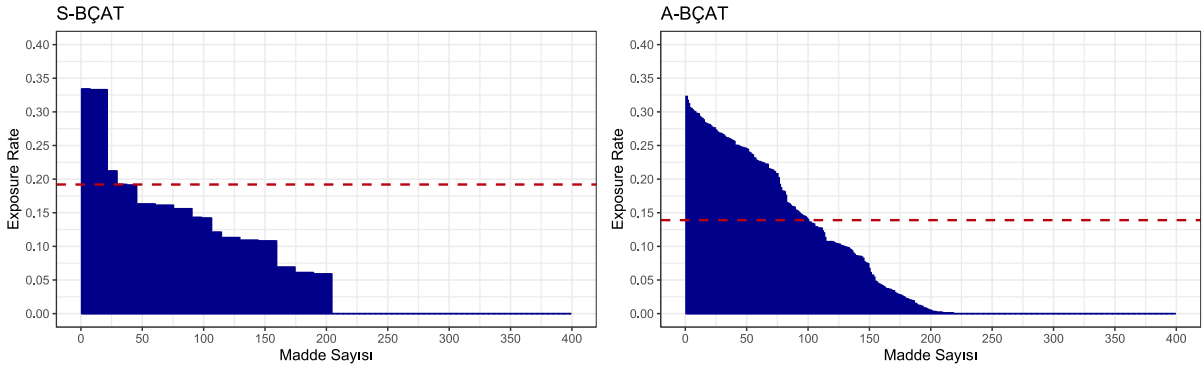


Test Uzunluğu = 30, Modül/Test U. Oranı = K-K-U, Yetenek Dağılımı = Normal

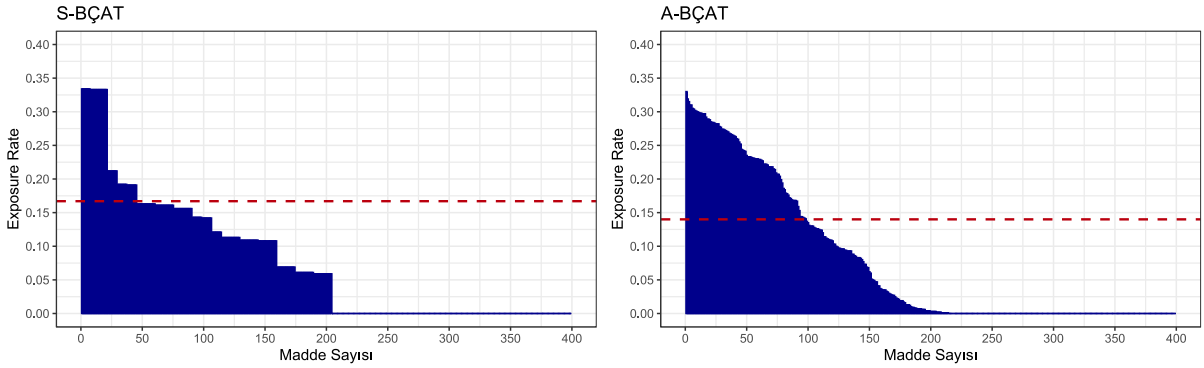


EK-D: S-BÇAT ile A-BÇAT'ın Madde Düzeyinde Madde Kullanım Sıklıklarının Karşılaştırılması (Tüm Koşullar) (DEVAMI)

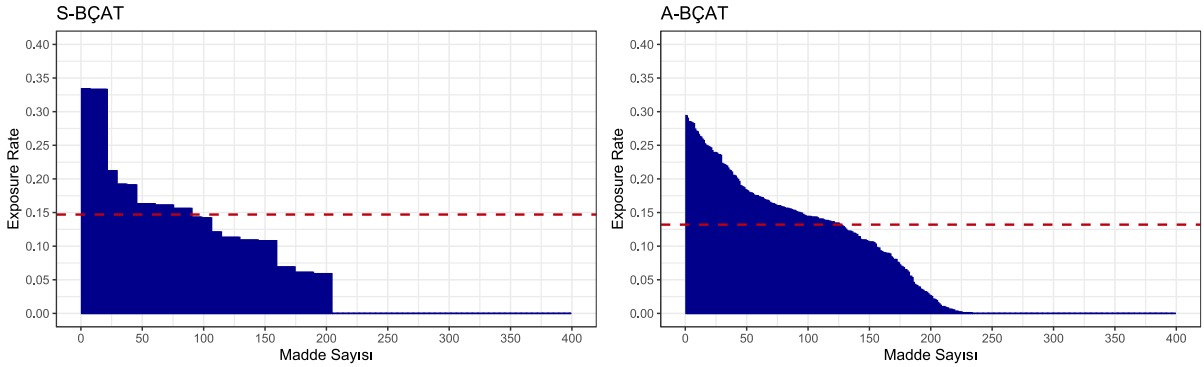
Test Uzunluğu = 30, Modül/Test U. Oranı = K-K-U, Yetenek Dağılımı = Sağa Çarpık



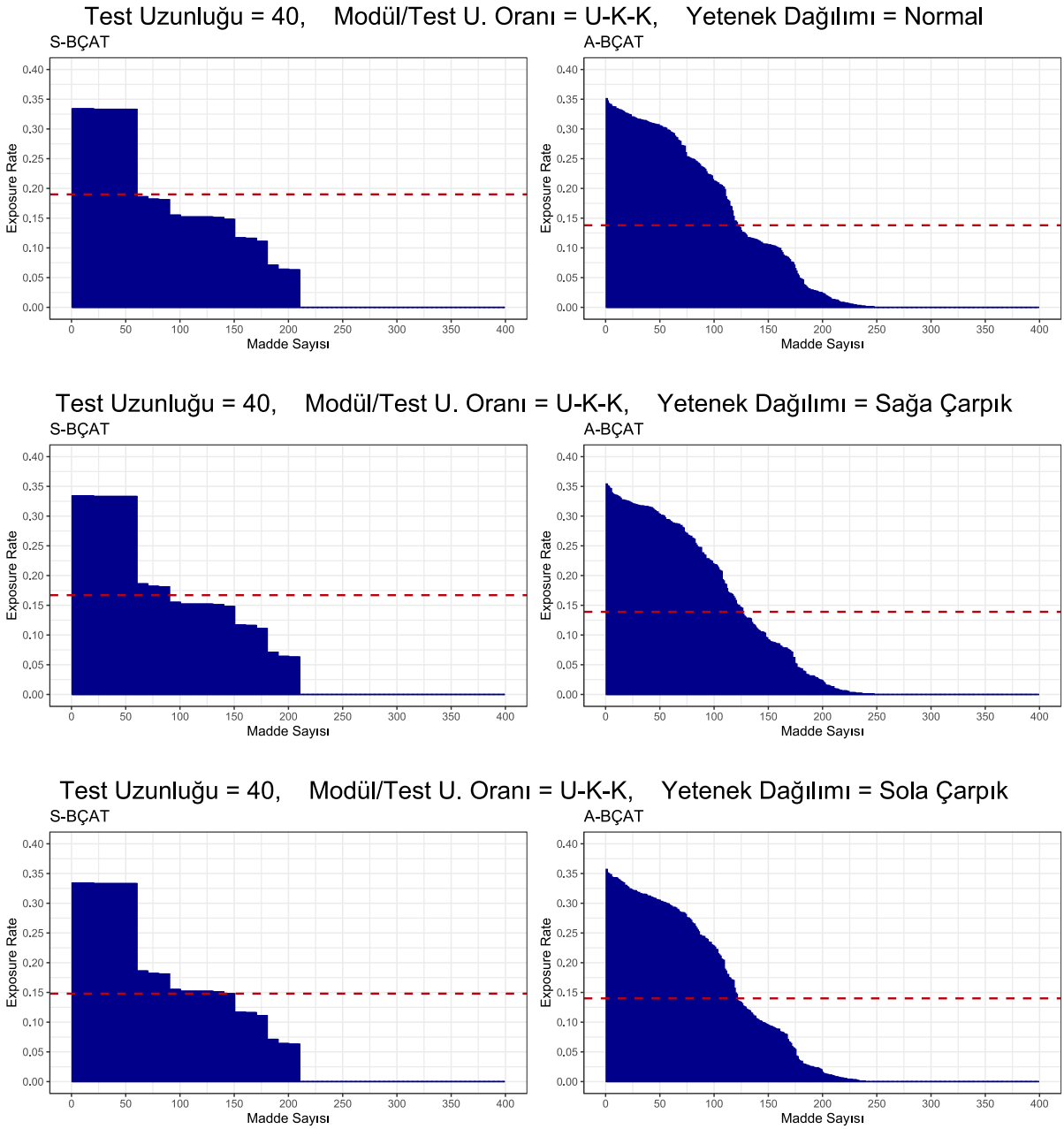
Test Uzunluğu = 30, Modül/Test U. Oranı = K-K-U, Yetenek Dağılımı = Sola Çarpık



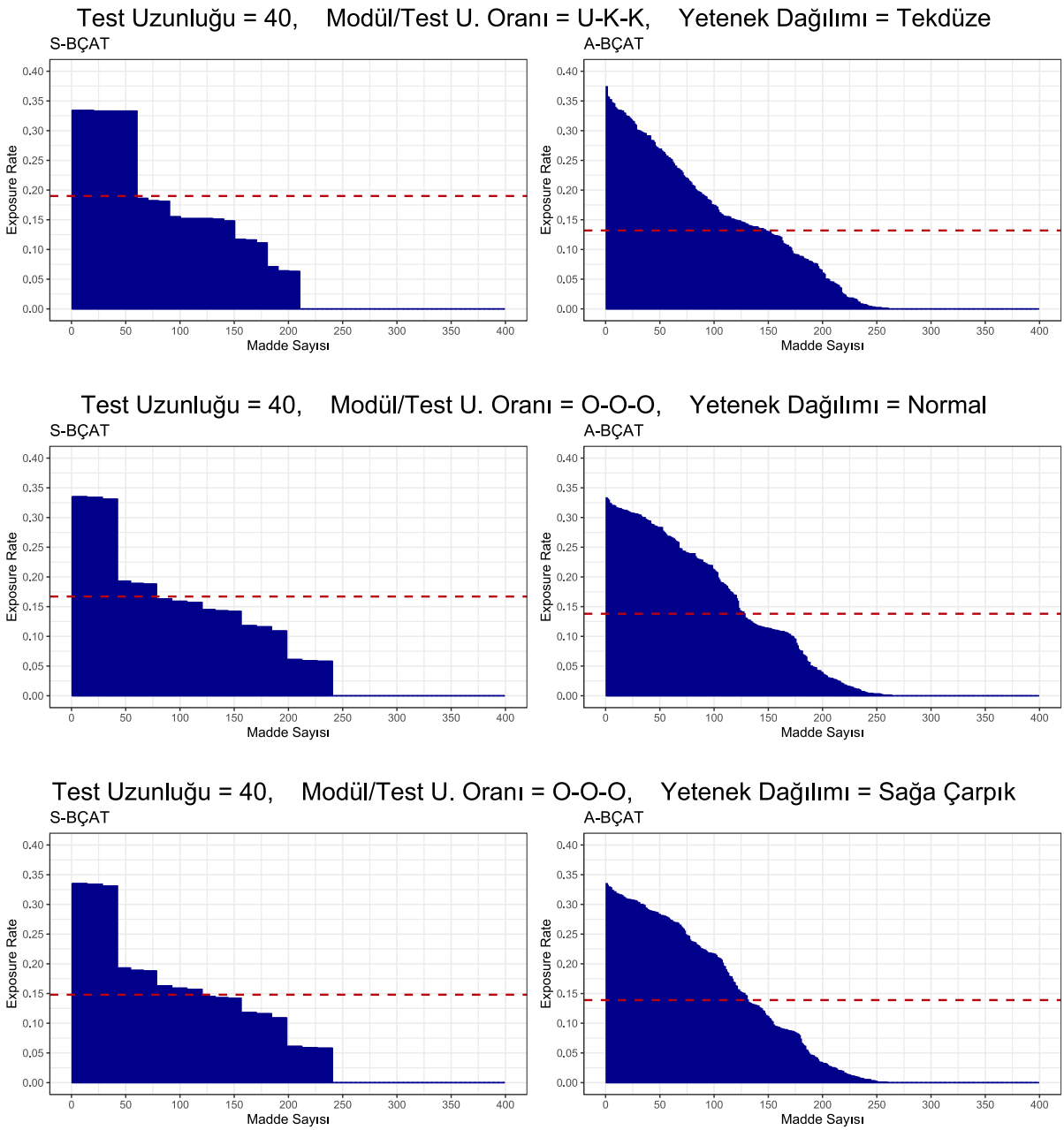
Test Uzunluğu = 40, Modül/Test U. Oranı = K-K-U, Yetenek Dağılımı = Tekdüze



EK-D: S-BÇAT ile A-BÇAT'ın Madde Düzeyinde Madde Kullanım Sıklıklarının Karşılaştırılması (Tüm Koşullar) (DEVAMI)

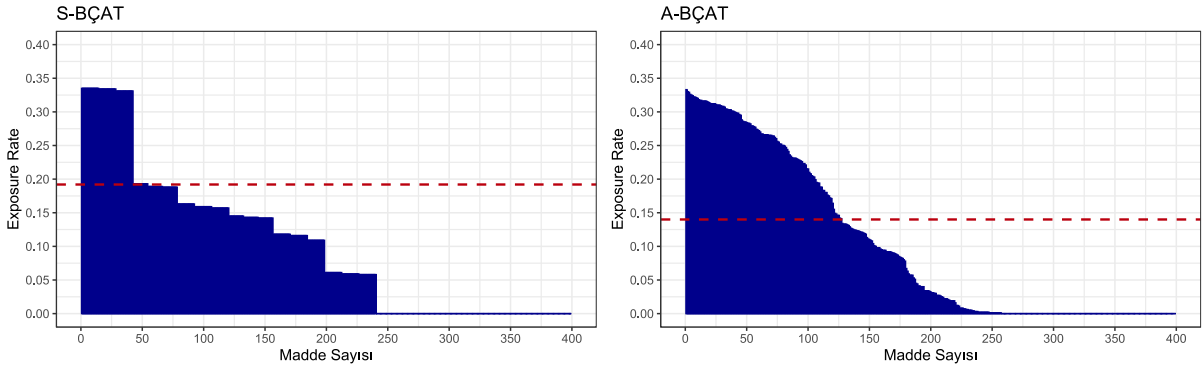


EK-D: S-BÇAT ile A-BÇAT'ın Madde Düzeyinde Madde Kullanım Sıklıklarının Karşılaştırılması (Tüm Koşullar) (DEVAMI)

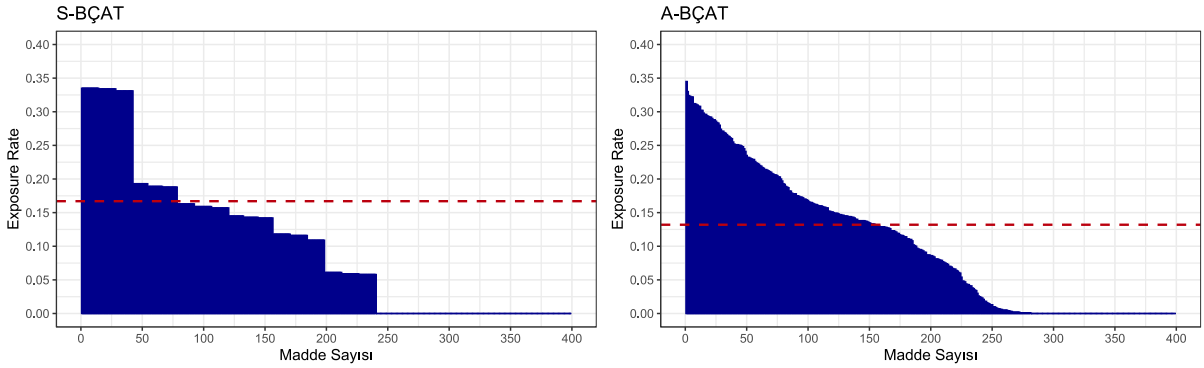


EK-D: S-BÇAT ile A-BÇAT'ın Madde Düzeyinde Madde Kullanım Sıklıklarının Karşılaştırılması (Tüm Koşullar) (DEVAMI)

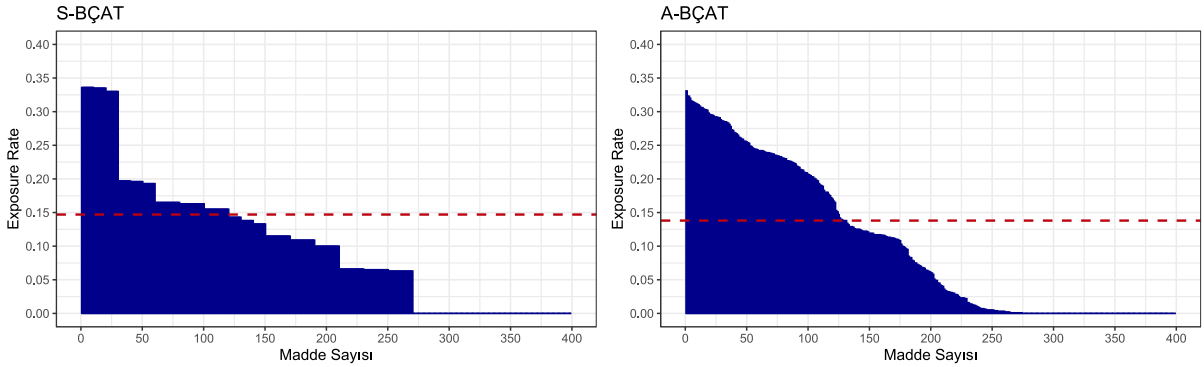
Test Uzunluğu = 40, Modül/Test U. Oranı = O-O-O, Yetenek Dağılımı = Sola Çarpık



Test Uzunluğu = 40, Modül/Test U. Oranı = O-O-O, Yetenek Dağılımı = Tekdüze

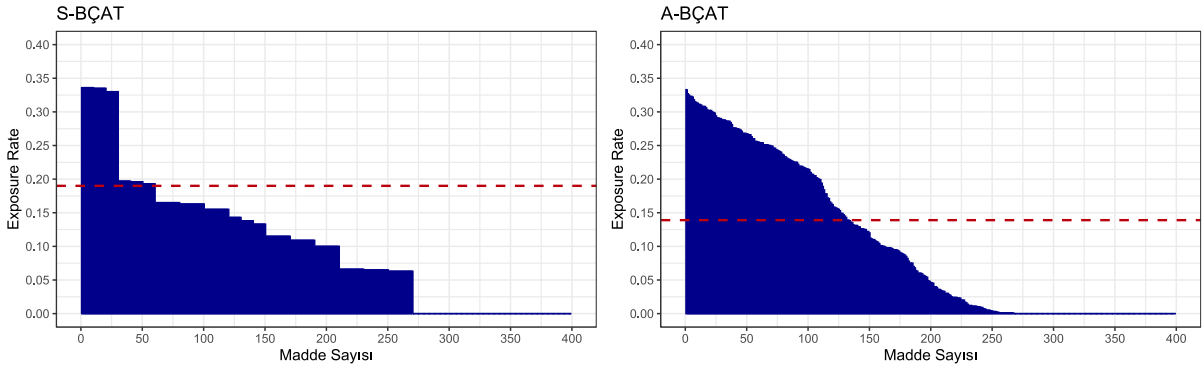


Test Uzunluğu = 40, Modül/Test U. Oranı = K-K-U, Yetenek Dağılımı = Normal

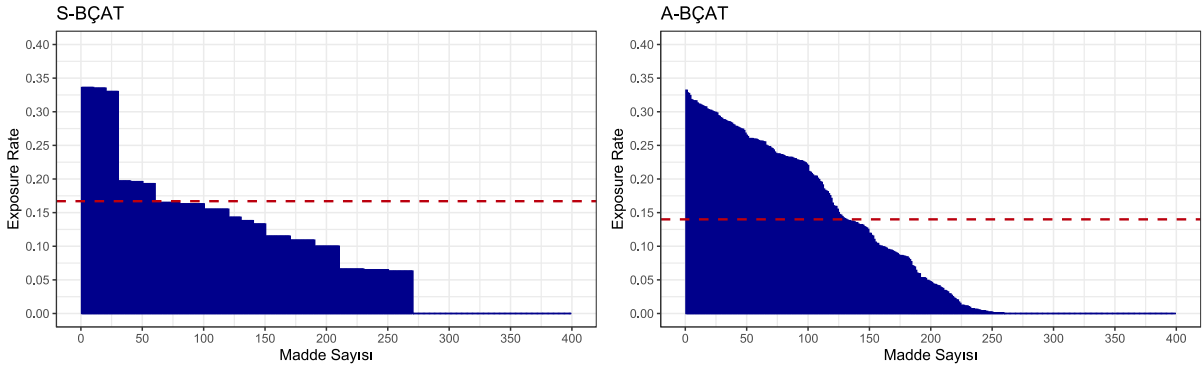


EK-D: S-BÇAT ile A-BÇAT'ın Madde Düzeyinde Madde Kullanım Sıklıklarının Karşılaştırılması (Tüm Koşullar) (DEVAMI)

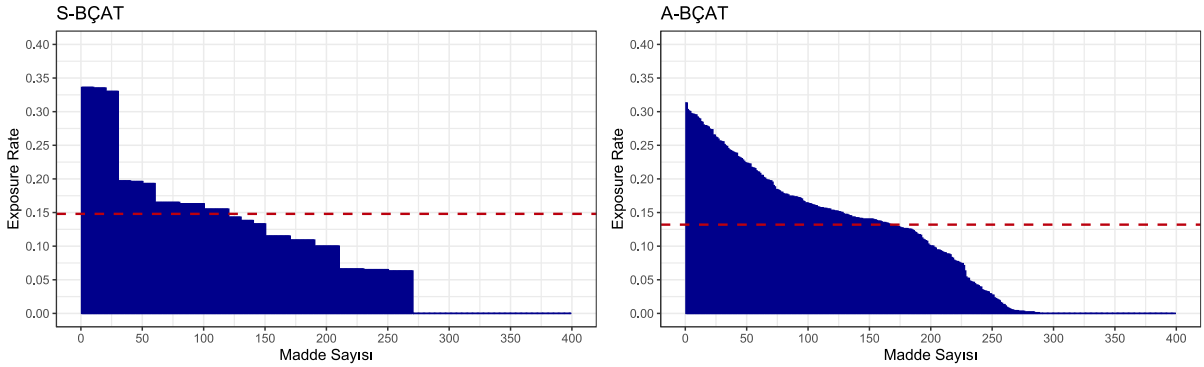
Test Uzunluğu = 40, Modül/Test U. Oranı = K-K-U, Yetenek Dağılımı = Sağa Çarpık



Test Uzunluğu = 40, Modül/Test U. Oranı = K-K-U, Yetenek Dağılımı = Sola Çarpık



Test Uzunluğu = 40, Modül/Test U. Oranı = K-K-U, Yetenek Dağılımı = Tekdüze



EK-E: Araştırma Etik Komisyon İzin Muafiyeti Formu/ Araştırma Etik Komisyonu Onay Bildirimi

Hacettepe Üniversitesi
Eğitim Bilimleri Enstitüsü
Tez Çalışması/Araştırma Etik Komisyon İzin Muafiyeti Formu

F46

03 / 05 / 2023

Hacettepe Üniversitesi
Eğitim Bilimleri Enstitüsü
Eğitim Bilimleri Ana Bilim Dalı Başkanlığına

Tez/Araştırma Başlığı	Sabit ve Anında Bireyselleştirilmiş Çok Aşamalı Testlerin Karşılaştırılması

Yukarıda başlığı/konusu verilen tez/araştırma çalışmam,

1. İnsan ve hayvan üzerinde deney niteliği taşımamaktadır.
2. Biyolojik materyal (kan, idrar vb. biyolojik sıvılar ve numuneler) kullanılmasını gerektirmemektedir.
3. Beden bütünlüğüne veya ruh sağlığına müdahale içermemektedir.
4. Anket, ölçek (test), mülakat, odak grup çalışması, gözlem, deney, görüşme gibi teknikler kullanılarak katılımcılardan veri toplanmasını gerektiren nitel ya da nicel yaklaşımlarla yürütülen araştırmalar niteliğinde değildir.
5. Diğer kişi ve kurumlardan temin edilen veri kullanımını (kitap, belge vs.) gerektirmektedir. Ancak bu kullanım, diğer kişi ve kurumların izin verdiği ölçüde Kişisel Bilgilerin Korunması Kanuna riayet edilerek gerçekleştirilecektir.

Çalışmada kullanacağım veriler:
( ) Kamusal erişime açık (buraya yazınız): .....
( ) Özel izin ve onaya tabi (buraya yazınız): .....
(X) Üretilmiş veri (buraya yazınız): Simülasyon ortamında üretilmiş veri
( ) Diğer (buraya yazınız): .....

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

Gereğini saygılarımla arz ederim.

Mahmut Sami YİĞİTER

Araştırmacı Bilgileri

Adı Soyadı	**Mahmut Sami YİĞİTER**
Öğrenci ise No	N19146394
Ana Bilim Dalı	Eğitim Bilimleri Ana Bilim Dalı
Programı	Eğitimde Ölçme ve Değerlendirme Programı
Statüsü	☐ Yüksek Lisans ☒ Doktora ☐ Bütünleşik Dr. ☐ Diğer

Danışman Görüşü ve Onayı*

Bu çalışmada kullanılan veriler bilgisayarda simülasyon ortamında üretilmiş olup herhangi bir uygulama yapılmamıştır.

Prof. Dr. Nuri DOĞAN

Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü. Beytepe Yerleşkesi, 06800, Çankaya / ANKARA
Telefon: 0(312) 297 85 72 Belgegeçer: 0(312) 297 85 66 e-Ağ: http://ebe.hacettepe.edu.tr/ e-Posta: ebe@hacettepe.edu.tr

EK-F: Etik Beyanı

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

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

beyan ederim.

03/05/2023

(İmza)

Mahmut Sami YİĞİTER

EK-G: Doktora Tez Çalışması Orijinallik Raporu

15/05/2023

HACETTEPE ÜNİVERSİTESİ
Eğitim Bilimleri Enstitüsü
Eğitim Bilimleri Ana Bilim Dalı Başkanlığına,

Tez Başlığı : Sabit ve Anında Bireyselleştirilmiş Çok Aşamalı Testlerin Karşılaştırılması

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

Rapor Tarihi	Sayfa Sayısı	Karakter Sayısı	Savunma Tarihi	Benzerlik Oranı	Gönderim Numarası
08/05/2023	117	164 391	15/08/2023	%2	2087009637

Uygulanan filtreler:
1. Kaynaklar hariç
2. Alıntılar dâhil
3. 5 kelimeden daha az örtüşme içeren metin kısımları hariç

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

Ad Soyadı: Mahmut Sami YİĞİTER

Öğrenci No.: N19146394

Ana Bilim Dalı: Eğitim Bilimleri Ana Bilim Dalı

İmza

Programı: Eğitimde Ölçme ve Değerlendirme Programı

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

DANIŞMAN ONAYI

UYGUNDUR.

Prof. Dr. Nuri DOĞAN

Danışman

EK-H: Thesis/Dissertation Originality Report

15/05/2023

HACETTEPE UNIVERSITY
Graduate School of Educational Sciences
To The Department of Educational Sciences

Thesis Title: Comparison of Fixed and On-The-Fly Computerized Multistage Testing

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

Time Submitted	Page Count	Character Count	Date of Thesis Defense	Similarity Index	Submission ID
08/05/2023	117	164 391	15/08/2023	%2	2087009637

Filtering options applied:
1. Bibliography excluded
2. Quotes included
3. Match size up to 5 words excluded

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

I respectfully submit this for approval.

Name Lastname: Mahmut Sami YİĞİTER

Student No.: N19146394

Department: Educational Sciences

Program: Educational Measurement and Evaluation

Status: ☐ Masters ☒ Ph.D. ☐ Integrated Ph.D.

Signature

ADVISOR APPROVAL

APPROVED

Prof. Dr. Nuri DOĞAN

Supervisor

EK-I: Yayımlama ve Fikrî Mülkiyet Hakları Beyanı

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kâğıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan **"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına ilişkin Yönerge"** kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H.Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

- Enstitü/ Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihinden itibaren 2 yıl ertelenmiştir. (1)
- Enstitü/Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren … ay ertelenmiştir. (2)
- Tezimle ilgili gizlilik kararı verilmiştir. (3)

26/05/2023

Mahmut Sami YİĞİTER

"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"

(1) Madde 6. 1. Lisansüstü tezle ilgili patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda, tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü Üzerine enstitü veya fakülte yönetim kurulu iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.

(2) Madde 6. 2. Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç; imkânı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez danışmanın önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulunun gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.

(3) Madde 7. 1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, tezin yapıldığı kurum tarafından verilir. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, ilgili kurum ve kuruluşun önerisi ile enstitü veya fakültenin uygun görüşü Üzerine üniversite yönetim kurulu tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.*

Madde 7.2. Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir

**Tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu tarafından karar verilir.*